



Assurance Case Structure for Safety Functions Enabled by Artificial Intelligence

Julieth Patricia Castellanos-Ardila^(✉) and Sasikumar Punnekkat

Mälardalen University, Västerås 721 23, Sweden
{julieth.castellanos,sasikumar.punnekkat}@mdu.se

Abstract. Artificial intelligence (AI) technologies are being adopted in high-risk systems where malfunctions may have severe consequences. However, AI's non-deterministic, adaptive, and opaque behavior challenges assumptions of predictability and control, creating uncertainty about whether AI systems will consistently avoid harm. This challenge is recognized by regulatory frameworks, especially the European Union's (EU) AI Act, which establishes mandatory assurance requirements for high-risk AI systems. Commonly, EU regulations are implemented via harmonized standards, which operationalize assurance expectations. At the same time, delays in the availability of harmonized standards under the EU AI Act, together with recent EU simplification initiatives, reinforce the need for pre-harmonization approaches relying on currently available standardization guidance and assurance practices. This paper addresses safety argumentation for safety functions enabled or supported by AI, i.e., functions necessary to achieve or maintain a safe state. Our major research contribution is twofold. First, we propose AIFAL (AI Function Assurance Level), a conceptual construct that is built upon the AI classification guidance established in ISO/IEC TR 5469. Second, we define a modular assurance case structure that embeds AIFAL within the safety argument through claims, strategies, and contextual assumptions. The structure is expressed using Goal Structuring Notation (GSN). We also illustrate the structure by applying it to the justification of an AI-enabled coordination function for System-of-Systems (SoS) operations at a construction site. Together, these contributions support standards-informed safety arguments that integrate AI-specific assurance concerns.

Keywords: Safety Functions · Artificial Intelligence (AI) · Assurance Cases · GSN · AI Standards · Functional Safety · AIFAL

1 Introduction

Artificial intelligence (AI) technologies are increasingly being integrated into high-risk systems, i.e., systems requiring mandatory assurance, as their intended

This work is supported by the Trusted Smart Systems initiative at Mälardalen University via the projects SAILS and DOSTY. It is also partially supported by Vinnova via the project SIMCON.

use may affect health, safety, and fundamental rights [8]. This integration is driven by the significant benefits AI provides despite its inherent characteristics [2], i.e., non-deterministic, adaptive, and opaque behaviors. However, AI's characteristics challenge assumptions of predictability and control typical of traditional software [16], intensifying the AI safety problem [5], i.e., the challenge of ensuring that AI-enabled systems act in ways that avoid harm. These characteristics also limit the effectiveness of testing, rendering testing-based approaches insufficient for high-risk AI systems, which demand high levels of confidence [4] or strong safety guarantees [5]. Consequently, justifying the safety of AI-enabled systems calls for structured safety argumentation, as embodied in safety assurance cases, to coherently and consistently integrate diverse sources of evidence.

Regulatory developments further reinforce the need for evidence-based safety assurance. In particular, the European Union's (EU) approach to trustworthy AI, i.e., systems that are lawful, ethical, and technically robust [13], has culminated in the adoption of the EU AI Act [8]. The Act establishes the first binding, horizontal, risk-based regulatory framework for AI systems in the EU, under which high-risk AI systems must demonstrate conformity with mandatory requirements. Article 40 introduces a presumption of conformity through harmonized standards, which provide technical support for meeting the Act's requirements. However, delays in the availability of such standards create practical challenges for operationalizing the Act's high-risk AI obligations. Recent EU regulatory initiatives, i.e., the Digital Omnibus [7], further highlight these transitional difficulties and reinforce the relevance of pre-harmonization assurance based on available standardization guidance and assurance practices. In particular, ISO/IEC TR 5469 [17], which addresses the relationship between functional safety and AI systems, can provide valuable input for pre-harmonization assurance while future harmonized standards are still under development.

This paper addresses safety argumentation for safety functions enabled or supported by AI, i.e., functions intended to reduce risk or prevent harm and whose failure could lead to unacceptable safety consequences [14]. The contribution is twofold. First, we propose AIFAL (AI Function Assurance Level), an assurance-case-oriented conceptual construct derived from the technology class and usage level categories established in ISO/IEC TR 5469. AIFAL helps assess the assurance burden and the feasibility of justifying an AI-enabled safety function realization. Second, we define a modular assurance case structure that embeds AIFAL within the safety argument through claims, strategies, and contextual assumptions. The structure is expressed using Goal Structuring Notation (GSN) [25]. We also illustrate the structure by applying it to the justification of an AI-enabled coordination function for System-of-Systems (SoS) operations at a construction site. Together, these contributions provide a standards-informed basis for constructing safety arguments for AI-enabled safety functions. The explicit integration of AI-specific assurance concerns into safety case structures may also inform future discussions on harmonized assurance practices for AI-enabled safety-critical systems. The remainder of the paper is structured as follows. In Sect. 2, we present essential background. In Sect. 3, we describe our

assurance case structure. In Sect. 4, we illustrate the structure with a case study. In Sect. 5, we discuss our findings. In Sect. 6, we discuss related work. Finally, in Sect. 7, we conclude our work and present future directions.

2 Background

This section provides the essential background necessary for the rest of the paper.

2.1 Assurance Cases

Assurance provides a justified basis for confidence in the validity of a claim [18], where a claim states the system property to be justified. Claim validity depends on reasonable arguments about the operational context, as well as on stated assumptions and justifications regarding the system and its use, and must align with the level of risk reduction required. In practice, assurance information is captured in assurance cases, which present structured arguments supported by evidence and associated confidence levels to justify the stated claims.

We use a graphical notation called GSN [25]. Arguments in GSN start with a top-level goal, supported by strategies linking it to subgoals and solutions. Context, assumptions, and justifications are also provided to support claim acceptability. Those elements (see Fig. 1a) are connected with relationships, i.e., SupportedBy (link goals with strategies/solutions) and InContextOf (link goals/strategies with contextual information). GSN provides decorators (see Fig. 1b). For example, the hollow diamond, added to a goal, represents an undeveloped goal, i.e., a goal to which the line of argument still needs to be developed. GSN structures can also be partitioned into separate packages (see Fig. 1c), e.g., an away goal represents a claim presented in another module.

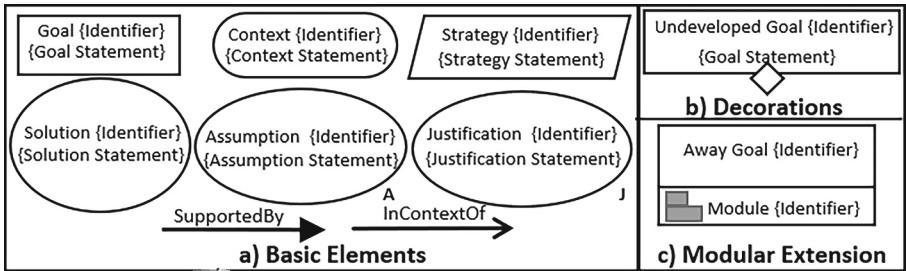


Fig. 1. GSN (Goal Structuring Notation) Elements [25].

2.2 Functional Safety

Functional safety (FuSa) addresses the portion of overall safety that depends on the correct operation of functions intended to prevent or mitigate hazards in safety-critical systems (i.e., systems whose failure or malfunction can lead to unacceptable harm). FuSa guidance, provided in IEC 61508:2010 [14], establishes the need for **safety functions**, which are implemented by safety-related systems to maintain a safe state under defined conditions. To ensure adequate risk reduction, IEC 61508 uses **Safety Integrity Levels (SILs)**. SILs are discrete levels (1–4) specifying increasing safety integrity and the rigor of development, verification, validation, and assurance activities. Compliance with the standard is demonstrated through evidence-based safety justification, in which supporting evidence is used to show that safety requirements are met and that safety functions operate correctly under stated assumptions and conditions.

2.3 FuSa and AI Systems

ISO/IEC TR 5469:2024 [17] complements established FuSa practices (see Sect. 2.2) with AI-specific considerations. Key aspects are as follows:

- a. **AI Technology Class (TC).** Classification reflecting increasing challenges in assurance feasibility. *TC I* exhibits deterministic behavior, enabling conventional software assurance. *TC II* shows limited data dependence remaining sufficiently constrained to support safety justification with additional measures. *TC III* typically adapts at runtime, posing significant challenges for safety justification not yet solved in the state of the art.
- b. **AI Usage Levels (UL).** Classification of AI involvement in safety. *UL A* refers to AI used within safety-related functions. *UL B* refers to AI used during the development of safety-related functions. Further, *UL C* referring to AI in adjacent functions and *UL D* referring to AI included in architecturally segregated functions, correspond to safety-relevant, but indirect functions. In both cases, the AI system may still impact overall system safety. *UL A* and *UL B* are further subdivided into A1/B1 for systems with automated decision-making and A2/B2 for systems without automated decision-making.
- c. **Levels of autonomy.** The degree to which an AI system operates independently of human supervision, directly affecting system controllability. Automation levels are *no automation*, *assistance*, *partial automation*, *conditional automation*, *high automation*, *full automation* and *autonomous systems*. Higher levels of automation increase assurance challenges, as unexpected or adaptive behavior becomes harder to detect, control, and validate.
- d. **AI Realization Dimensions.** In this work, we collectively refer to complementary aspects through which an AI-enabled capability is constructed and realized as AI realization dimensions, including:
 - **AI System Technology Layers.** Architectural layers, spanning algorithmic decision logic, software realization, and underlying physical systems, through which an AI capability is realized.

- **AI realization stages.** They distinguish between stages such as data preparation, knowledge induction, and operational use, capturing how the AI-enabled capability evolves across the lifecycle.
- **AI capability components.** They refer to the technical basis underlying AI behavior, combining the model structure and learned parameters that together determine the capability realized by the AI system

The AI realization dimensions are not used to classify the AI system itself. Instead, they structure the evidence needed to justify the AI-enabled realization within the AIFAL-based assurance argument (See Sect. 3.1). ISO/IEC TR 24028 [15] further identifies **AI trustworthiness** attributes that influence confidence in AI behavior under data variability and uncertainty.

3 Assurance Cases for Safety Functions Enabled by AI

This section presents the assurance case structure introduced in this paper.

3.1 AIFAL

Safety functions (see Sect. 2.2) require evidence-based justification of their ability to prevent or mitigate hazardous situations. When such functions are enabled or supported by AI, this justification must also address how the characteristics of the AI technology affect assurance. To support such justification for safety functions enabled or supported by AI, we propose AIFAL (AI Function Assurance Level). AIFAL is a conceptual construct derived from the AI classification guidance in ISO/IEC TR 5469. First, the AI's assurance potential, analogous to TC (see Sect. 2.3.a), reflects how compatible the AI technology is with established FuSa assurance methods. Second, the system's safety exposure, analogous to the UL (see Sect. 2.3.b), reflects the criticality of the AI function's role in FuSa. Together, these parameters indicate the assurance burden associated with justifying the AI-enabled safety function.

Figure 2 presents AIFAL as a two-dimensional positioning scheme, where safety exposure is represented along the horizontal axis and assurance potential along the vertical axis. The colors indicate increasing assurance burden across this space. For each region, the AI realization dimensions (see Sect. 2.3.d) are used to guide the identification of relevant assurance evidence, as follows:

- **Green.** Indicates low assurance burden, typically due to limited safety exposure. In this case, the AI realization dimensions are considered mainly to confirm limited safety relevance, sufficient bounds, segregation, or compatibility with conventional FuSa evidence.
- **Yellow.** Indicates moderate assurance burden. In this case, the AI has a safety-relevant role, but its assurance potential remains sufficiently high for existing FuSa methods to provide the main basis for justification. The AI realization dimensions should therefore be considered to identify where AI-specific evidence is needed to complement conventional FuSa evidence.

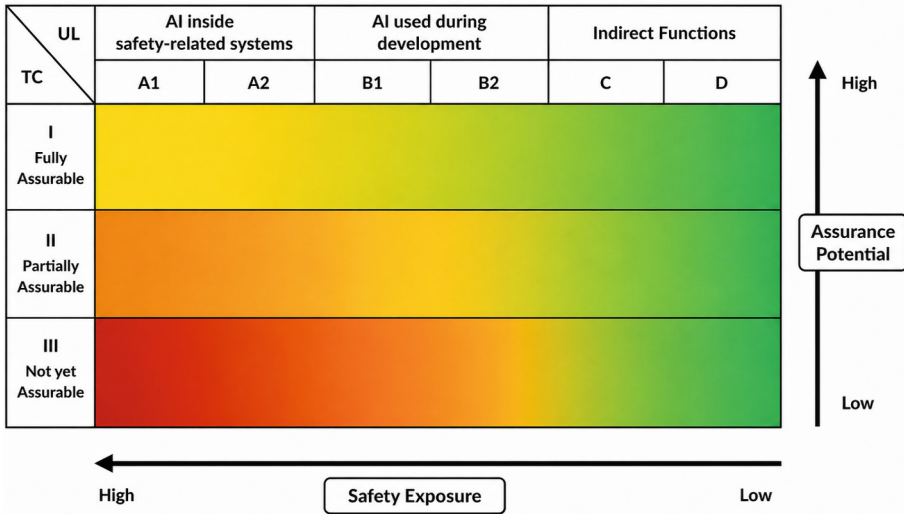


Fig. 2. AI Function Assurance Level (AIFAL).

- **Orange.** Indicates high but feasible assurance burden. In this region, the AI may have safety involvement comparable to yellow cases, but its assurance potential is lower. Existing FuSa methods do not provide the full basis for justification. Therefore, additional AI-specific evidence is needed across the relevant AI realization dimensions to address the residual assurance gap.
- **Red:** Represents critical scenarios in which low assurance potential prevents the formulation of convincing safety arguments using current methods. In such cases, the AI-enabled design should be reconsidered or properly constrained before a convincing justification can be formulated.

The shaded colors transitions reflect the continuous and context-dependent nature of assurance burden rather than sharp categorical boundaries. Thus, AIFAL should not be interpreted as a fixed compliance classification like SIL. Distinctions such as A1/A2 and B1/B2 remain relevant because they indicate whether AI is involved in automated decision-making within its usage context. This aspect is related to the levels of autonomy introduced in Sect. 2.3.c., which describe the degree of independence from human supervision or intervention. Thus, the same AIFAL region may require different assurance arguments depending on the autonomy level of the AI-enabled function.

3.2 Top-Level Goal and Supporting Arguments

AI-enabled system behavior introduces uncertainty that cannot be fully justified by conventional deterministic assurance reasoning alone, since AI-enabled systems are typically probabilistic and data-driven. Therefore, the safety of an

AI-enabled safety function must be argued through structured evidence that supports sufficient confidence in safe operation. In this work, the top-level goal, $G1$, captures this primary claim, i.e., the AI-enabled safety function (SF) operates safely with demonstrable confidence. The SF is assumed to be part of a safety architecture with an assigned SIL, derived from prior hazard and risk analysis performed outside this assurance case.

Accordingly, the claim is characterized by four elements. A SIL, in $C1$, represents the required integrity and risk reduction. An AIFAL, in $C2$, represents the assurance effort required to justify the AI-enabled realization of the function. An autonomy level (AL), in $C3$, clarifies the allocation of responsibilities between humans and the AI system. Finally, a specified set of environmental conditions (EC), in $C4$, defines the operational scope in which the SF is expected to operate. Together, these elements delimit the conditions under which the claim must hold. This claim, captured in the top-level goal in Fig. 3 using GSN elements (see Fig. 1.a), anchors the subsequent argumentation.

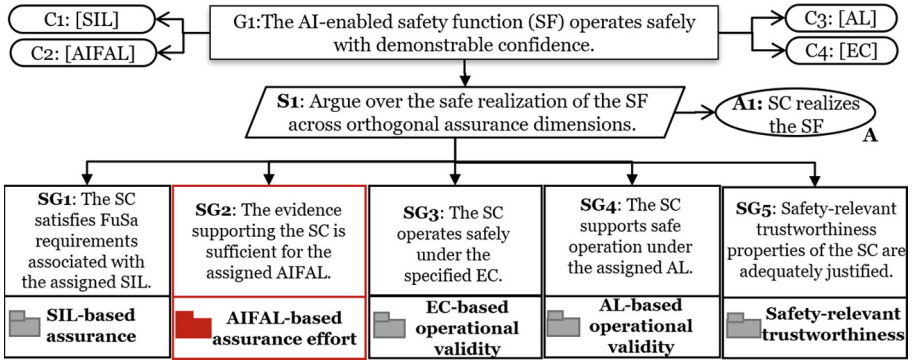


Fig. 3. Top-level Goal and Supporting Arguments.

The strategy $S1$ proceeds by arguing for the safe realization of the SF. For this, the AI safety capability (SC) is assumed to be the element realizing the SF, and the supporting arguments are constructed around its justification. The arguments are organized into orthogonal assurance modules, where each module captures a distinct assurance concern associated with the contextual characterization of the top-level goal. Accordingly, the modules address SIL-based functional safety assurance, $SG1$, AIFAL-based assurance effort, $SG2$, EC-based operational validity, $SG3$, and AL-based operational validity, $SG4$. In addition, a separate module addresses safety-relevant trustworthiness properties of the SC, $SG5$, capturing AI-specific quality concerns that may influence confidence in the safe realization of the SF beyond the contextual characterization of the top-level goal. Among these modules, this work focuses specifically on the AIFAL-based assurance effort module (highlighted in red in Fig. 3).

3.3 AIFAL-Based Assurance Effort

This module (see Fig. 4) addresses evidential sufficiency for the assigned AIFAL, as formalized by the top-level claim $G2.1$. The strategy node, $S2.1$, argues that such sufficiency can be established across AI realization dimensions, which, as justified by $J2.1$, are interpreted according to ISO/IEC TR 5469 (see Sect. 2.3.d). These dimensions capture distinct assurance concerns and therefore provide a structured basis for establishing evidential sufficiency. In particular, $AG2.1$ addresses AI capability components, characterizing what constitutes the AI-enabled realization of the SC (e.g., AI models and learned parameters). $AG2.2$ addresses AI realization stages, characterizing how assurance concerns emerge across the lifecycle (e.g., data acquisition, knowledge induction, and generation of outputs). $AG2.3$ addresses AI technology layers (e.g., software, hardware, libraries), characterizing where the realization is implemented within the system architecture. Finally, $AG2.4$ addresses the coherent integration of these dimensions. Together, these four sub-goals provide structured support for evidential sufficiency under the assigned AIFAL.

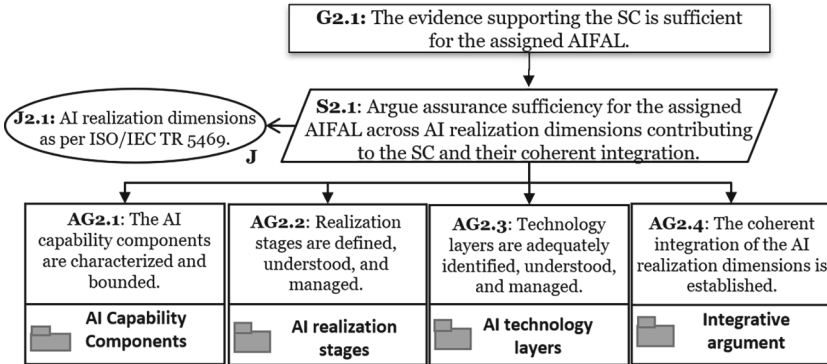


Fig. 4. AI Safety Capability Assurance Argument.

4 Case Study

The case study considers mass removal operations at a construction site. The solution [1] adopts a System of Systems (SoS) perspective, where constituent systems (CS), i.e., autonomous diggers, loaders, and transporters, collaborate while retaining operational autonomy. As shown in Fig. 5, these CS operate in predefined zones and are coordinated by an AI-based orchestrator responsible for task allocation and synchronization. At the SoS level, safety is enforced through coordination contracts. We focus on contract SoS-C2, a safety function that relies on an AI-enabled coordination capability to enforce exclusive access to the shared material pile using a learned orchestration strategy.

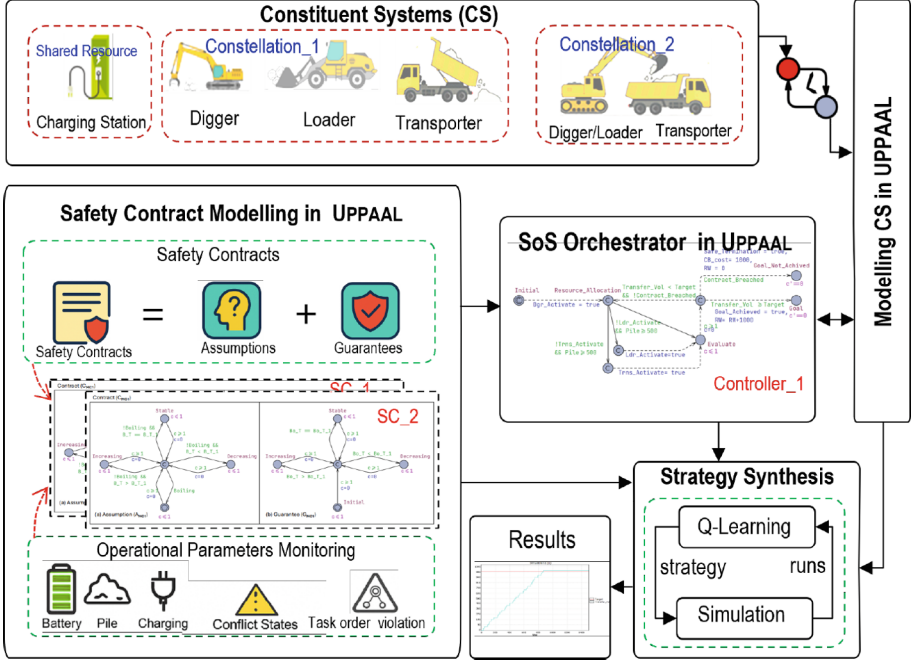


Fig. 5. Mass Removal in a Construction Site (taken from [1]).

4.1 Top-Level Goal and Supporting Arguments

In this case study, the selected safety function (SF) prevents unsafe simultaneous access of autonomous machines to the shared pile during mass removal operations ($G1$), addressing hazards related to proximity and potential collisions. Based on the risk reduction required for the selected safety function, the SF is assigned SIL 2 ($C1$). The SF is realized through an AI-enabled safety capability (SC) providing SoS-level coordination of access, implemented via offline Q-learning in UPPAAL Stratego. This SC produces a fixed coordination policy, verified at design time and deployed at runtime for automated safety-relevant decisions (UL A1). Since learning is performed offline and the deployed policy does not adapt at runtime, the SC is classified as TC II rather than TC III. This results in a high but feasible assurance burden corresponding to the AIFAL orange region ($C2$) (see Fig. 2). Runtime operation will require human supervision, corresponding to conditional automation ($C3$), while the operational scope is restricted to a controlled construction site with predefined zones and known terrain constraints ($C4$). Figure 6 shows the instantiation of the top-level safety goal and its supporting arguments. The subsequent analysis focuses specifically on the AIFAL-based assurance effort module ($SG2$).

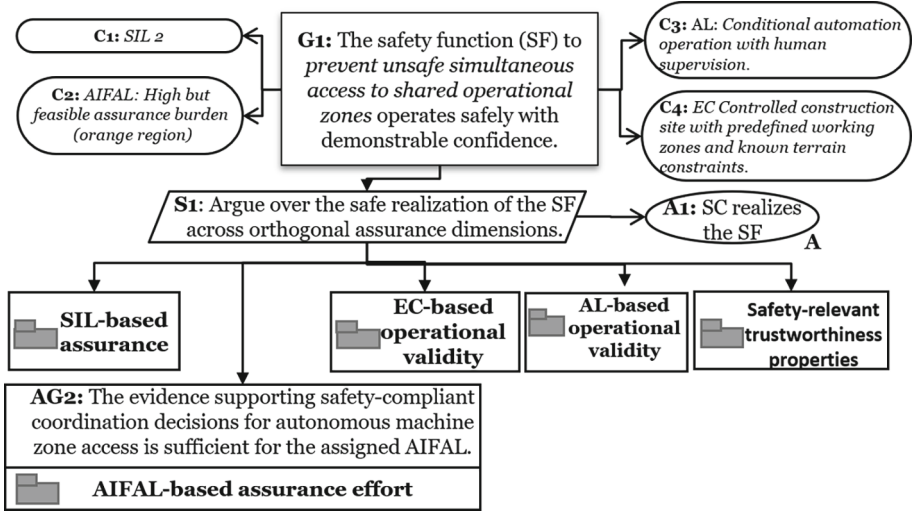


Fig. 6. Top-Level Safety Goal and Supporting Assurance Arguments.

4.2 AIFAL-Based Assurance Effort

The proposed argument structure introduces a set of sub-goals that collectively support the justification of the assigned AIFAL across relevant realization dimensions, as illustrated in Fig. 4. The detailed instantiation of all these sub-goals is inherently case-dependent, since different application domains, system architectures, and AI technologies impose distinct assumptions, constraints, and assurance concerns. For this reason, and due to space limitations, this paper focuses on the development of one representative sub-goal, AG2.3. This sub-goal addresses assurance of the AI technology layers, which, according to ISO/IEC TR 5469 (see Sect. 2.3.d), characterize how AI-enabled functionality is realized across complementary architectural layers that constrain AI behavior.

In the present case study, the AI-enabled coordination capability is realized across three AI technology layers. The *decision layer* comprises the Q-learning-based coordination logic synthesized in UPPAAL Stratego and executed by the SoS orchestrator to regulate access to the shared operational zone. The *software/model layer* consists of the UPPAAL models and safety contracts that define and verify the assumptions and guarantees governing coordination behavior. The *physical layer* includes the CS, predefined operational zones, shared resources, and monitored parameters that form the execution environment.

The goal *G2.3.1* (see Fig. 7) is supported by corresponding assurance claims in accordance with the layered realization previously presented. At the *decision layer*, the claim is that the AI supports safe coordination decisions. This is supported by *E.2.3.1.1* corresponding to the synthesized coordination policy and simulation results. At the *software/model layer*, the claim is that this layer preserves safety assumptions and guarantees, supported by *E.2.3.1.2* corresponding

to formal models and safety contracts. At the *physical layer*, the claim is that this layer enforces safe coordination decisions, supported by *E.2.3.1.3* corresponding to system models and representative execution scenarios.

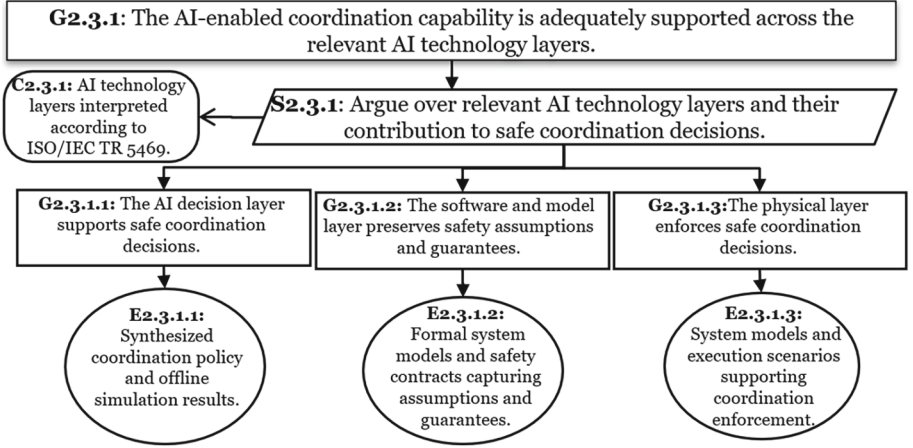


Fig. 7. Goal Supporting the AI Safety Capability Assurance.

4.3 Case Study Implications

The case study is not intended to demonstrate general applicability across all AI-enabled safety functions. It was selected because it provides a concrete safety-relevant AI coordination function in a bounded operational context, allowing the proposed assurance structure to be instantiated without excessive domain-specific complexity. However, broader validation will require application to more complex settings, such as systems involving online learning, richer human–AI interaction, and less structured operational environments.

The case study also illustrates how the assigned AIFAL region can guide expected evidence without prescribing a fixed checklist. Since the selected safety function is positioned in the orange region, conventional FuSa evidence needs to be complemented by AI-specific evidence. To illustrate this, the paper develops one representative branch of the AIFAL-based assurance argument, focused on AI technology layers. Thus, the case study should be read as a partial instantiation rather than as a complete safety case. In this way, AIFAL helps orient evidence generation while keeping concrete evidence requirements case-dependent.

5 Discussion

This section discusses key aspects of the proposed structure. First, AIFAL (Sect. 3.1) is introduced as a qualitative positioning tool derived from the UL–TC combination in ISO/IEC TR 5469. As a complementary construct, AIFAL

captures differences in assurance effort beyond what SIL alone represents. For example, functions in the green region may be justified using SIL-dependent FuSa measures alone, whereas those in the yellow or orange regions require progressively stronger AI-specific assurance activities. In this way, AIFAL extends the descriptive UL–TC classification into a parameter that supports safety case construction. It can also be seen as a dynamic filter, where functions may shift toward lower assurance burden as AI techniques mature, i.e., moving from red to orange as decision-making becomes more transparent or traceable. These regions may also provide a basis for future graded assurance approaches. However, their formalization requires further refinement and consensus, as AIFAL remains a continuous, context-dependent construct rather than a system of fixed levels.

Second, the formulation of the top-level goal (see Sect. 3.2) does not replace established FuSa principles, but extends them by incorporating AI-specific elements, i.e., safety capability (SC), autonomy level (AL), and AIFAL. This extension makes explicit the dependencies on AI and the conditions under which it supports the safety function. Scalability is supported by the modular structure of the safety case, which allows distinct assurance concerns to be developed, assessed, reused, and refined separately. However, such scalability depends on making assumptions, context, and evidence dependencies explicit, so that argument branches can be safely maintained and reused. The proposed structure further enables AIFAL to be operationalized through explicit reasoning across AI realization dimensions (see Sect. 3.3), transforming assurance effort into decomposable argument structures. In this way, not all dimensions need to be developed at once. Instead, branches can remain undeveloped until required by the system context, domain needs, available evidence, or emerging standards.

The case study (Sect. 4) also illustrates how AIFAL can inform early design decisions by making the assurance consequences of AI technology choices explicit. For example, extending the AI capability from offline to online learning would shift its positioning toward the red region, where convincing safety justifications are not yet feasible with current methods. In this sense, AIFAL can help identify when an AI-enabled design remains within a feasible assurance space and when additional constraints, safeguards, or design alternatives are needed. However, AIFAL should be understood as an orientation mechanism rather than a predefined evidence assessment scheme. In particular, it indicates the expected assurance burden, but does not yet define region-specific sufficiency criteria. Refining such criteria remains an important direction for future work.

6 Related Work

Existing literature proposes capability- and risk-based classifications for AI systems. For example, Anthropic’s AI Safety Levels (ASL) [3] organize governance responses by increasing model capability and catastrophic risk, from ASL-1 (low-risk systems) to ASL-5+ levels associated with superhuman intelligence and large-scale harm. Similarly, the Future of Life Institute (FLI) [9] classifies AI systems by escalating capability (ASL-1 to -4) and societal risk within a binding, regulator-led governance framework, coupling each level to enforceable safety

requirements and oversight. Dalrymple et al. [5] use capability-based safety levels (e.g., ASL) to determine assurance strength and motivate architectures delivering quantitative safety guarantees. Complementary to these classifications, Khlaaf et al. [20] propose a hazard and risk assessment framework for code-synthesis large language models, introducing a Hazard Risk Index (HRI) that combines severity and likelihood to prioritize concrete deployment risks. Diemert et al. [6] propose AI-SIL, which combines an allocated SIL with AI task complexity, expressed through input entropy and output non-determinism, to guide level-of-rigor activities for AI-based components. Porter et al. [24] propose INSYTE, a multidimensional classification framework for traditional to agentic AI systems, characterizing system design, functionality, operating environment, and independence from human operational control. Unlike these capability-, risk-, or level-of-rigor approaches, the proposed AIFAL regions characterizes the assurance burden and feasibility of justifying safety functions enabled or supported by AI within a safety case structure, consistent with ISO/IEC TR 5469.

Beyond classification frameworks, AI safety argumentation builds on modularity introduced by Kelly and McDermid [19]. In particular, Habli et al.'s BIG argument [10] presents a multi-layer GSN framework integrating ethical, system-level, and AI-specific assurance reasoning. Bloomfield and Rushby [4] analyze AI assurance from a classical dependability perspective, advocating guarded architectures and rigorous deductive arguments rather than structured GSN extensions. Paterson et al. [23] propose AMLAS, a lifecycle-based GSN methodology ensuring traceability from hazards to ML artefacts. Lee et al. [21] propose a reusable CAE-based framework for AI safety cases, introducing taxonomies of AI-specific claim types, reasoning logics, and composable patterns. Our proposed structure also employs a modular GSN framework but is scoped to safety functions enabled or supported by AI within established functional safety contexts.

In parallel, literature also addresses conformity assessment and emerging standardization for AI-enabled systems. Thelisson and Verma [26] analyze the conformity assessment regime of the EU AI Act within the EU New Legislative Framework, discussing internal and external assessment pathways and the structure of ex-ante compliance mechanisms. Heidemann et al. [12] present a standards-oriented compliance methodology that translates the EU AI Act into extended quality attributes and proposes a contract-based, GSN-supported verification framework to operationalize conformity across the AI value chain. Complementing these perspectives, Hayama et al. [11] provide a GSN-based requirement analysis that maps high-risk classification and lifecycle obligations of the AI Act into a structured regulatory argument, improving traceability of legal provisions. Locherer et al. [22] critically examine ISO/IEC TR 5469 and highlight that, while it provides useful guidance on AI and functional safety, it remains informative and does not by itself define a complete basis for rigorous assurance argumentation. Against this background, our proposed approach moves from governance-level compliance and standardization analysis to the technical justification of safety functions, using selected concepts from ISO/IEC TR 5469 to structure AI-specific assurance reasoning within a modular safety case.

7 Conclusions and Future Work

This paper presented an assurance case structure for safety functions enabled or supported by AI, informed by available standardization guidance on the relationship between functional safety and AI systems. The proposed AIFAL construct supports the explicit positioning of an AI-enabled safety function in terms of assurance burden, helping to clarify when conventional FuSa evidence may be sufficient and when AI-specific assurance is needed. The modular GSN structure embeds this reasoning within the safety argument by relating the top-level safety claim to assumptions, contextual conditions, and supporting assurance modules. The construction-site case study illustrated how the structure can be instantiated for an AI-enabled coordination function in System-of-Systems operations. Overall, the paper shows how AI-specific assurance concerns can be incorporated into safety case structures in a systematic and traceable way, offering a basis for further work on assurance practices for AI-enabled safety-critical systems.

Future work will refine AIFAL by developing more explicit guidance for interpreting and assigning its regions. This includes defining criteria or decision questions related to assurance potential and safety exposure, as well as representative assurance measures that can guide the development of region-specific arguments branches. The proposed structure should also be applied across diverse industrial case studies to assess its usability, scalability, and applicability to different AI technologies, autonomy levels, and operational contexts. Finally, tool support is needed to enable more efficient, traceable, and maintainable safety justification for AI-enabled systems, particularly when assurance arguments evolve together with system designs, assumptions, and supporting evidence.

References

1. Ali, N., Naeem, M., Castellanos-Ardila, J.P., Punnekkat, S.: Safety-aware strategy synthesis for autonomous system of systems with UPPAAL. In: Törngren, M., Galina, B., Schoitsch, E., Troubitsyna, E., Bitsch, F. (eds.) *International Conference on Computer Safety, Reliability, and Security*, pp. 73–87. Springer, Cham (2025). https://doi.org/10.1007/978-3-032-02018-5_6
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in AI safety. arXiv preprint [arXiv:1606.06565](https://arxiv.org/abs/1606.06565) (2016)
3. Anthropic: Anthropic’s responsible scaling policy (2023). <https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf>. Accessed 13 Jan 2026
4. Bloomfield, R., Rushby, J.: Assurance of AI systems from a dependability perspective. arXiv preprint [arXiv:2407.13948v3](https://arxiv.org/abs/2407.13948v3) (2025)
5. Dalrymple, D., et al.: Towards guaranteed safe AI: a framework for ensuring robust and reliable AI systems. arXiv preprint [arXiv:2405.06624](https://arxiv.org/abs/2405.06624) (2024)
6. Diemert, S., Millet, L., Groves, J., Joyce, J.: Safety integrity levels for artificial intelligence. In: Guiochet, J., Tonetta, S., Schoitsch, E., Roy, M., Bitsch, F. (eds.) *International Conference on Computer Safety, Reliability, and Security*, pp. 397–409. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-40953-0_34

7. European Commission: Digital omnibus on AI regulation proposal (COM(2025) 836 final) (2025). <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52025PC0836>
8. European Union: Regulation (EU) 2024/1689 (Artificial Intelligence Act) (2024). <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L.202401689>, official Journal of the European Union
9. Future of Life Institute: AI governance scorecard and safety standards policy (2023). https://futureoflife.org/wp-content/uploads/2023/11/FLI_Governance_Scorecard_and_Framework.pdf. Accessed 13 Jan 2026
10. Habli, I., et al.: The BIG argument for AI safety cases. arXiv preprint [arXiv:2503.11705](https://arxiv.org/abs/2503.11705) (2025)
11. Hayama, N., Yamagata, Y., Nishihara, H., Matsuno, Y.: A GSN-based requirement analysis of the EU AI regulation. In: Törnngren, M., Gallina, B., Schoitsch, E., Troubitsyna, E., Bitsch, F. (eds) International Conference on Computer Safety, Reliability, and Security, pp. 183–196. Springer, Cham (2025). https://doi.org/10.1007/978-3-032-02018-5_14
12. Heidemann, L., et al.: The European Artificial Intelligence Act Overview and Recommendations for Compliance. Whitepaper-Fraunhofer (2024)
13. High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI (2019). <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>. Accessed Jan 2026
14. International Electrotechnical Commission: IEC 61508:2010. Functional safety of electrical/electronic/programmable electronic safety-related systems (2010). <https://webstore.iec.ch/en/publication/5515>
15. ISO/IEC JTC 1/SC 42: ISO/IEC TR 24028:2020. Information technology — Artificial intelligence - Overview of trustworthiness in artificial intelligence (2020). <https://www.iso.org/standard/77608.html>
16. ISO/IEC JTC 1/SC 42: ISO/IEC 23894:2023. Information technology — Artificial intelligence - Guidance on risk management (2023). <https://www.iso.org/standard/77304.html>
17. ISO/IEC JTC 1/SC 42: ISO/IEC TR 5469:2024. Artificial intelligence - Functional safety and AI systems (2024). <https://www.iso.org/standard/81283.html>
18. ISO/IEC JTC 1/SC 7: ISO/IEC/IEEE 15026:2019. Systems and software engineering — Systems and software assurance - Part 1:vocabulary and concepts (2025). <https://www.iso.org/standard/89808.html>
19. Kelly, T.P., McDerimid, J.A.: Safety case construction and reuse using patterns. In: Daniel, P. (ed.) 16th International Conference on Computer Safety, Reliability and Security. pp. 55–69. Springer, London (1997). https://doi.org/10.1007/978-1-4471-0997-6_5
20. Khlaaf, H., Mishkin, P., Achiam, J., Krueger, G., Brundage, M.: A hazard analysis framework for code synthesis large language models. arXiv preprint [arXiv:2207.14157](https://arxiv.org/abs/2207.14157) (2022)
21. Lee, S.U., Zhu, L., Shamsujjoha, M., Dong, L., Lu, Q., Chen, J.: Constructing safety cases for AI systems: a reusable template framework. arXiv preprint [arXiv:2601.22773](https://arxiv.org/abs/2601.22773) (2026)
22. Locherer, M., Kordovan, M., Buxbaum, B., Kever, T.: No safe AI standardization in sight at present: an investigation into ISO/IEC TR 5469. In: 9th International Conference on System Reliability and Safety (ICSRS), pp. 87–94. IEEE (2025)
23. Paterson, C., Hawkins, R., Picardi, C., Jia, Y., Calinescu, R., Habli, I.: Safety assurance of machine learning for autonomous systems. Reliab. Eng. Syst. Saf. **264**, 111311 (2025)

24. Porter, Z., et al.: INSYTE: a classification framework for traditional to agentic AI systems. *ACM Trans. Auton. Adapt. Syst.* **20**(3), 1–39 (2025)
25. The Assurance Case Working Group (ACWG): GSN Community Standard. Version, vol. 3 (2021)
26. Thelisson, E., Verma, H.: Conformity assessment under the EU AI act general approach. *AI Ethics* **4**(1), 113–121 (2024)