Functional Safety and AI technologies: background, standardization landscape and overview of ISO/TR 5469

Takashi Egawa, Riccardo Mariani

2nd ISO/IEC AI workshop, November 30th, 2022
IEC 61508-4 defines functional safety as that “part of the overall safety relating to the EUC (Equipment Under Control) and the EUC control system that depends on the correct functioning of the E/E/PE (Electrical/Electronic/Programmable Electronic) safety-related systems and other risk reduction measures.”

The E/E/PE safety-related system is delivering a “safety function”, which is defined in IEC 61508-4 as a “function to be implemented by an E/E/PE safety-related system or other risk reduction measures, that is intended to achieve or maintain a safe state for the EUC, in respect of a specific hazardous event.”
Example of safety functions

Emergency Stop

Protective Stop

- $s_1$
- $s_0$

Protective Stop Zone

Safeguarded Zone

Robot
Functional Safety Basics

- **Fault:** abnormal condition that may cause a reduction in, or loss of, the capability of a functional unit to perform a required function
- **Error:** discrepancy between a computed, observed or measured value or condition and the true, specified or theoretically correct value or condition
- **Failure:** termination of the ability of a functional unit to provide a required function or operation of a functional unit in any way other than as required

**Systematic Fault (HW and SW)**
- Human-induced, such as software bugs or hardware design marginalities
- Deterministic occurrence
- Focus is mainly on fault avoidance
- Mitigations primarily focus on process-related improvements and diversity

**Random HW Fault**
- Not human-induced, such as electromigration, radiation effects, interferences
- Non-deterministic: following a probability distribution
- Focus is mainly on fault detection
- Mitigations primarily focus on safety mechanisms and redundancy concepts
IEC 61508

• Overview:
  • It is a basic functional safety standard applicable to all kinds of industry
  • It addresses system/hardware and software
  • It introduces a safety lifecycle
  • It uses a probabilistic failure approach
Functional Safety and AI

• The use of artificial intelligence (AI) technology in industry has increased significantly in recent years and AI has been demonstrated to deliver benefit in several applications.

• However, there is limited guidance on specification, design and verification of functionally safe AI systems or on how to apply AI technology for functions that have safety-related effects.
Safety AI landscape (1/3)

- ISO/IEC JTC 1/SC 42
  - TR 5469 – “Artificial intelligence — Functional safety and AI systems”

- ISO/TC 22/SC 32
  - ISO/PAS 8800 - “Road Vehicles – Safety and Artificial Intelligence”
Safety AI landscape (2/3)

- IEEE
  - IEEE CS FSSC – Functional Safety Standards Committee
  - IEEE P2851 - Standard for functional safety data format for interoperability within the dependability lifecycle
Safety AI landscape (3/3)

• **Other**
  - VDE-AR-E 2842-61 - Development and trustworthiness of autonomous/cognitive systems
  - AS6983 - Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI
  - EUROCAE WG114 - to prepare technical standards, guides and any other material required to support the development of systems and the certification of aeronautical systems implementing AI-technologies
TR 5469

- Currently in DTR stage
- Developed in liaison with MT 61508 (AIFS task force of MT 61508-3)
- TR expected to be published in Summer 2023
TR 5469 contents

Contents

27 Foreword ........................................ vi
28 Introduction ................................... vii
30 1 Scope ......................................... 1
31 2 Normative references ...................... 1
32 3 Terms and definitions ...................... 1
33 4 Abbreviations ................................ 4
34 5 Overview of functional safety .......... 4
35 5.1 Introduction to functional safety ....... 4
36 5.2 Functional safety .......................... 5
37 6 Use of AI technology in safety-related E/E/PE systems .......................... 6
38 6.1 Problem description ....................... 6
39 6.2 AI technology in safety-related E/E/PE systems .......................... 6
40 7 AI technology elements and the three-stage realization principle .... 9
41 7.1 Technology elements for AI model creation and execution .... 9
42 7.2 The three-stage realization principle of an AI system .... 11
43 7.3 Deriving acceptance criteria for the three-stage of the realization principle .... 12
44 8 Properties and related risk factors of AI systems .... 13
45 8.1 Introduction ................................ 13
46 8.1.1 General .................................. 13
47 8.1.2 Algorithms and models .............. 13
48 8.2 The level of automation and control ........ 14
49 8.3 The degree of transparency and explainability .... 15
50 8.4 The complexity of the environment and vague specifications .......... 16
51 8.4.1 Overview ................................ 16
52 8.4.2 Data drift ............................... 17
53 8.4.3 Concept drift ............................ 18
54 8.4.4 Reward hacking algorithms .......... 18
55 8.4.5 Safe exploration ................. 19
56 8.5 Resilience to adversarial and intentional inputs .................. 19
57 8.5.1 Introduction ............................. 19
58 8.5.2 General mitigations ................. 19
59 8.5.3 AI model attack: adversarial machine learning .......... 20
60 8.6 AI hardware issues ................. 21
61 8.7 The readiness of the technology .......... 21
62 9 Verification and validation techniques .............. 21
63 9.1 Introduction ................................ 21
64 9.2 Problems related to verification and validation .......... 22
65 9.2.1 Existence of a priori specification ........ 22
66 9.2.2 Non-understandability of particular system behaviour .......... 22
67 9.2.3 Limitation of test coverage .......... 22
68 9.2.4 Non-predictable nature .............. 23
69 9.2.5 Long-term stability of risk mitigations .......... 23
70 9.3 Possible solutions ................. 23
71 9.3.1 General ................................ 23
72 9.3.2 Relation between data distributions and HARA .......... 24
73 9.3.3 Data preparation and model-level validation and verification .......... 24
74 9.3.4 Choice of AI metrics .............. 24
75 9.3.5 System-level testing ............... 24
76 9.3.6 Mitigating techniques for data-size limitation .......... 27
77 9.3.7 Notes and additional resources .......... 27
78 9.4 Virtual and physical testing .......... 27
79 9.4.1 General ................................ 27
80 9.4.2 Considerations on virtual testing .......... 27
81 9.4.3 Considerations on physical testing .......... 29
82 9.4.4 Evaluation of vulnerability to hardware random failures .......... 29
83 9.5 Monitoring and incident feedback .......... 30
84 9.6 A note on explainable AI .......... 30
85 10 Control and mitigation measures .......... 31
86 10.1 Introduction ................................ 31
87 10.2 AI subsystem architectural considerations .......... 31
88 10.2.1 Introduction ................................ 31
89 10.2.2 Detection mechanisms for switching .......... 32
90 10.2.3 Use of a supervision function with constraints to control the behaviour of a system within safe limits .......... 32
91 10.2.4 Recessiveness concepts and diversity .......... 33
92 10.2.5 AI system design with statistical evaluation .......... 35
93 10.3 Increase of the reliability of components containing AI technology .......... 36
94 10.3.1 Introduction to AI component methods .......... 36
95 10.3.2 Use of robust learning ............... 36
96 10.3.3 Optimisation and compression technologies .......... 37
97 10.3.4 Attention mechanisms .......... 37
98 10.3.5 Protection of the data and parameters .......... 38
99 11 Preccesses and methodologies .......... 39
100 11.1 General .................................. 39
101 11.2 Relationship between AI lifecycle and functional safety lifecycle .......... 39
102 11.3 AI phases ................................ 40
103 11.4 Documentation and functional safety artefacts .......... 40
104 11.5 Methodologies ............................ 41
105 11.5.1 Introduction ............................. 41
106 11.5.2 Fault models .......................... 41
107 11.5.3 PFMEA of offshore training of AI technology .......... 41
108 Annex A (informative) Applicability of IEC 61508-3 to AI technology elements .......... 42
109 A1. Introduction ................................ 42
110 A1.1 Relationship between AI and IEC 61508-3 elements .......... 42
111 A2. Analysis of applicability of techniques and measures in IEC 61508-3:2010 Annexes A and B to AI technology elements .......... 42
112 Annex B (informative) Examples of applying the three-stage realization principle .......... 55
113 B1. Introduction ................................ 55
114 B1.2 Example for an automotive use case .......... 55
115 B1.3 Example for a robotics use case .......... 57
116 Annex C (informative) Possible process and useful technology for verification and validation .......... 61
117 C1 General .................................. 61
118 C2 Data distribution and HARA .......... 61
119 C3 Coverage of data for identified risks .......... 62
120 C4 Data diversity for identified risks .......... 62
121 C5 Reliability and robustness .......... 63
122 Annex D (informative) Mapping between ISO IEC 5338 and IEC 61508 series .......... 64
123 Bibliography .............................. 66
TR 5649 scope

- The DTR 5469 standard describes the properties, related risk factors, available methods and processes relating to:

  01 Use of AI inside a safety related function to realize the functionality
  02 Use of non-AI safety related functions to ensure safety for an AI controlled equipment
  03 Use of AI systems to design and develop safety related functions
Desirable Properties

- Specifiability
- Domain shift
- Verifiability
- Robustness
- Interpretability
- Explainability

Topics

- e.g. for verifiability
- How is the neural network performance assessed?
- How to determine when verification is complete?

Methods & Techniques

- e.g. for performance
- Definition of what constitutes a “correct” output by the network.
- Definition of what range of inputs is evaluated

Acceptance Criteria / KPI

- e.g. for performance
- Pixel level KPIs
- Image level KPIs
- Sequence level KPIs
- Dataset level KPIs

TR 5469 three stages realization principle
AI lifecycle vs functional safety lifecycle

- ISO/IEC 22989 describes a high-level lifecycle model of AI systems
- ISO/IEC 5338 defines the lifecycle processes of AI systems
- IEC 61508 describes the functional safety lifecycle

An example of mapping between ISO/IEC 5338 and the IEC 61508 series is provided in Annex D.
TR 5469 hierarchy of technology elements

- Application graph, machine learning framework
- Machine learning graph compiler, machine learning model
- Libraries of calculation operands, set of calculations
- Executable machine code, compiler
- Computational HW
TR 5469 AI classification scheme

DTR stage, November 2022

<table>
<thead>
<tr>
<th>AI Technology Class =&gt; AI application and usage level</th>
<th>AI technology Class I</th>
<th>AI technology Class II</th>
<th>AI technology Class III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage Level A1 (1)</td>
<td>Application of risk reduction concepts of existing functional safety International Standards possible</td>
<td>Appropriate set of requirements (5)</td>
<td>Not recommended</td>
</tr>
<tr>
<td>Usage Level A2 (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usage Level B1 (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usage Level B2 (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usage Level C (1,3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usage Level D (2)</td>
<td>No specific functional safety requirements for AI technology, but application of risk reduction concepts of existing functional safety International Standards (4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Static (offline) (during development) teaching or learning only
2. Dynamic (online) teaching or learning possible
3. AI techniques clearly providing additional risk reduction and whose failure is not critical to the level of acceptable risk.
4. Additionally, other safety aspects (not being addressed with functional safety methods) can possibly be adversely affected by AI usage.
5. The appropriate set of requirements for each usage level can be established in consideration of Clauses 8, 9, 10 and 11. Examples are provided in Annex B.
### Redundancy
- ML redundancy with output voters or aggregators

### Supervision
- A safe subset of the action space can be determined using a supervisor function with constraints or limits

### Back-up
- Safe (suboptimal) back-up function to the ML component can be designed with “non-AI” techniques.
- The back-up action allows the use of detection methods to switch the output when unsafe conditions are detected.
Redundancy is combined with diversity to reduce the likelihood of systematic failures during development.

This is related to multiple AI technologies exhibiting the same behaviour, but implemented:

• by different teams;
• using separate labelling rules;
• using different problem formulations;
• using different training data;
• executing on diverse hardware (also valid for non-AI technology specific failure modes);
• with diversity of sensing;
• with diversity of self-check or self-validation methods;
• with diversity of AI technology itself.
• An effective and objective way to demonstrate a system’s performance is via **virtual testing or simulation**, where a curated set of well-chosen stress-test scenarios can be exercised during the qualification and certification activities.

• **Physical tests** also have their place as a tangible way to correlate simulation results, validate KPIs and uncover unknown unknowns.
  
  • Physical tests are far more limited than simulation in their ability to probe the domain space due to cost and time limitations but do test some aspects that are difficult to emulate in a simulation, for example, the effect of hardware delays on feedback loops and cascade effects.
Thank you!