

ISO/IEC TR 5469:2024-01 (E)

Artificial intelligence - Functional safety and AI systems

Contents		Page
Foreword		v
Introduction		vi
1	Scope	1
2	Normative references	1
3	Terms and definitions	1
4	Abbreviated terms	4
5	Overview of functional safety	4
5.1	General	4
5.2	Functional safety	5
6	Use of AI technology in E/E/PE safety-related systems	6
6.1	Problem description	6
6.2	AI technology in E/E/PE safety-related systems	6
7	AI technology elements and the three-stage realization principle	10
7.1	Technology elements for AI model creation and execution	10
7.2	The three-stage realization principle of an AI system	12
7.3	Deriving acceptance criteria for the three-stage of the realization principle	12
8	Properties and related risk factors of AI systems	13
8.1	Overview	13
8.1.1	General	13
8.1.2	Algorithms and models	13
8.2	Level of automation and control	14
8.3	Degree of transparency and explainability	15
8.4	Issues related to environments	17
8.4.1	Complexity of the environment and vague specifications	17
8.4.2	Issues related to environmental changes	17
8.4.3	Issues related to learning from environment	18
8.5	Resilience to adversarial and intentional malicious inputs	19
8.5.1	Overview	19
8.5.2	General mitigations	19
8.5.3	AI model attacks: adversarial machine learning	19
8.6	AI hardware issues	20
8.7	Maturity of the technology	21
9	Verification and validation techniques	21
9.1	Overview	21
9.2	Problems related to verification and validation	22
9.2.1	Non-existence of an a priori specification	22
9.2.2	Non-separability of particular system behaviour	22
9.2.3	Limitation of test coverage	22
9.2.4	Non-predictable nature	22
9.2.5	Drifts and long-term risk mitigations	22
9.3	Possible solutions	23
9.3.1	General	23
9.3.2	Relationship between data distributions and HARA	23
9.3.3	Data preparation and model-level validation and verification	24
9.3.4	Choice of AI metrics	25
9.3.5	System-level testing	25

9.3.6	Mitigating techniques for data-size limitation.....	26
9.3.7	Notes and additional resources	26
9.4	Virtual and physical testing.....	26
9.4.1	General.....	26
9.4.2	Considerations on virtual testing.....	26
9.4.3	Considerations on physical testing.....	28
9.4.4	Evaluation of vulnerability to hardware random failures.....	29
9.5	Monitoring and incident feedback.....	29
9.6	A note on explainable AI.....	29
10	Control and mitigation measures.....	30
10.1	Overview.....	30
10.2	AI subsystem architectural considerations.....	30
10.2.1	Overview.....	30
10.2.2	Detection mechanisms for switching.....	30
10.2.3	Use of a supervision function with constraints to control the behaviour of a system to within safe limits.....	33
10.2.4	Redundancy, ensemble concepts and diversity.....	34
10.2.5	AI system design with statistical evaluation.....	35
10.3	Increase the reliability of components containing AI technology.....	35
10.3.1	Overview of AI component methods.....	35
10.3.2	Use of robust learning.....	35
10.3.3	Optimization and compression technologies.....	36
10.3.4	Attention mechanisms.....	37
10.3.5	Protection of the data and parameters.....	37
11	Processes and methodologies.....	38
11.1	General.....	38
11.2	Relationship between AI life cycle and functional safety life cycle.....	38
11.3	AI phases.....	39
11.4	Documentation and functional safety artefacts.....	39
11.5	Methodologies.....	39
11.5.1	Overview.....	39
11.5.2	Fault models.....	39
11.5.3	PFMEA for offline training of AI technology.....	40
Annex A (informative) Applicability of IEC 61508-3 to AI technology elements.....		41
Annex B (informative) Examples of applying the three-stage realization principle.....		54
Annex C (informative) Possible process and useful technology for verification and validation.....		59
Annex D (informative) Mapping between ISO/IEC 5338 and the IEC 61508 series.....		62
Bibliography.....		65