

ISO/IEC 24029-2:2023-08 (E)

Artificial intelligence (AI) - Assessment of the robustness of neural networks - Part 2: Methodology for the use of formal methods

Contents		Page
Foreword		iv
Introduction		v
1	Scope	1
2	Normative references	1
3	Terms and definitions	1
4	Abbreviated terms	4
5	Robustness assessment	4
5.1	General	4
5.2	Notion of domain	5
5.3	Stability	6
5.3.1	Stability property	6
5.3.2	Stability criterion	6
5.4	Sensitivity	6
5.4.1	Sensitivity property	6
5.4.2	Sensitivity criterion	7
5.5	Relevance	7
5.5.1	Relevance property	7
5.5.2	Relevance criterion	7
5.6	Reachability	8
5.6.1	Reachability property	8
5.6.2	Reachability criterion	8
6	Applicability of formal methods on neural networks	9
6.1	Types of neural network concerned	9
6.1.1	Architectures of neural networks	9
6.1.2	Neural networks input data type	10
6.2	Types of formal methods applicable	12
6.2.1	General	12
6.2.2	Solver	13
6.2.3	Abstract interpretation	13
6.2.4	Reachability analysis in deterministic environments	13
6.2.5	Reachability analysis in non-deterministic environments	14
6.2.6	Model checking	14
6.3	Summary	14
7	Robustness during the life cycle	15
7.1	General	15
7.2	During design and development	15
7.2.1	General	15
7.2.2	Identifying the recognized features	15
7.2.3	Checking separability	16
7.3	During verification and validation	16
7.3.1	General	16
7.3.2	Covering parts of the input domain	17
7.3.3	Measuring perturbation impact	17

7.4	During deployment	18
7.5	During operation and monitoring	19
7.5.1	General	19
7.5.2	Robustness on a domain of operation	19
7.5.3	Changes in robustness	20
	Bibliography	21