

ISO/IEC TS 6254:2025-09 (E)

Information technology - Artificial intelligence - Objectives and approaches for explainability and interpretability of machine learning (ML) models and artificial intelligence (AI) systems

Contents		Page
Foreword		v
Introduction		vi
1	Scope	1
2	Normative references	1
3	Terms and definitions	1
4	Symbols and abbreviated terms	5
5	Overview	6
6	Stakeholders' objectives	6
6.1	General	6
6.2	AI user	7
6.3	AI developer	7
6.4	AI product or service provider	7
6.5	AI platform provider	8
6.6	AI system integrator	8
6.7	Data provider	8
6.8	AI evaluator	8
6.9	AI auditor	8
6.10	AI subject	8
6.11	Relevant authorities	8
6.11.1	Policy makers	8
6.11.2	Regulators	8
6.11.3	Other authorities	9
7	Explainability considerations throughout the AI system life cycle	9
7.1	General	9
7.2	Inception	10
7.3	Design and development	10
7.3.1	General	10
7.3.2	Development of the explainability component	10
7.3.3	Explainability's contribution to development	11
7.4	Verification and validation	11
7.4.1	General	11
7.4.2	Evaluation of the explainability component	11
7.4.3	Explainability's contribution to evaluation	13
7.5	Deployment	14
7.5.1	General	14
7.5.2	Deployment of the explainability component	14
7.5.3	Explainability's contribution to deployment	14
7.6	Operation and monitoring	14
7.7	Continuous validation	14
7.8	Re-evaluation	14
7.9	Retirement	15
8	Property taxonomy of explainability methods and approaches	15
8.1	General	15
8.2	Properties of explanation needs	16

8.2.1	General	16
8.2.2	Expertise profile of the targeted audience	16
8.2.3	Frame activity of interpretation or explanation	17
8.2.4	Scope of information	17
8.2.5	Completeness	17
8.2.6	Depth	18
8.2.7	Reasoning path	18
8.2.8	Implicit and explicit explanations	19
8.3	Forms of explanation	19
8.3.1	General	19
8.3.2	Numeric	19
8.3.3	Visual	19
8.3.4	Textual	20
8.3.5	Structured	20
8.3.6	Example-based	20
8.3.7	Interactive exploration tools	20
8.4	Technical approaches towards explainability	20
8.4.1	General	20
8.4.2	Empirical analysis	21
8.4.3	Post hoc interpretation	21
8.4.4	Inherently interpretable components	21
8.4.5	Architecture- and task-driven explainability	22
8.5	Technical constraints of the explainability method	22
8.5.1	General	22
8.5.2	Genericity of the method	22
8.5.3	Transparency requirements	23
8.5.4	Display requirements	23
9	Approaches and methods to explainability	23
9.1	General	23
9.2	Empirical analysis methods	24
9.2.1	General	24
9.2.2	Fine-grained evaluation	25
9.2.3	Error analysis	25
9.2.4	Analysis-oriented datasets	25
9.2.5	Ablation	26
9.2.6	Known trends	26
9.3	Post hoc methods	27
9.3.1	Local	27
9.3.2	Global	32
9.4	Inherently interpretable components	36
9.4.1	General	36
9.4.2	Legible models	37
9.4.3	Meaningful models	39
9.4.4	Models with explicit knowledge	41
9.5	Architecture- and task-driven methods	43
9.5.1	General	43
9.5.2	Informative features	43
9.5.3	Rich and auxiliary inputs	44
9.5.4	Multi-step processing	44
9.5.5	Rich outputs	45
9.5.6	Rationale-based processing	46
9.5.7	Rationale generation as auxiliary output	46
9.6	Data explanation	47
Annex A (informative)	Extent of explainability and interaction with related concepts	48
Annex B (informative)	Illustration of methods' properties	51
Annex C (informative)	Concerns and limitations	61
Bibliography	65