

# ISO/IEC TR 24028:2020-05 (E)

## Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence

---

<b>Contents</b>		<b>Page</b>
Foreword .....		v
Introduction .....		vi
1	Scope .....	1
2	Normative references .....	1
3	Terms and definitions .....	1
4	Overview .....	7
5	Existing frameworks applicable to trustworthiness .....	7
5.1	Background .....	7
5.2	Recognition of layers of trust .....	8
5.3	Application of software and data quality standards .....	8
5.4	Application of risk management .....	10
5.5	Hardware-assisted approaches .....	10
6	Stakeholders .....	11
6.1	General concepts .....	11
6.2	Types .....	12
6.3	Assets .....	12
6.4	Values .....	13
7	Recognition of high-level concerns .....	13
7.1	Responsibility, accountability and governance .....	13
7.2	Safety .....	14
8	Vulnerabilities, threats and challenges .....	14
8.1	General .....	14
8.2	AI specific security threats .....	15
8.2.1	General .....	15
8.2.2	Data poisoning .....	15
8.2.3	Adversarial attacks .....	15
8.2.4	Model stealing .....	16
8.2.5	Hardware-focused threats to confidentiality and integrity .....	16
8.3	AI specific privacy threats .....	16
8.3.1	General .....	16
8.3.2	Data acquisition .....	16
8.3.3	Data pre-processing and modelling .....	17
8.3.4	Model query .....	17
8.4	Bias .....	17
8.5	Unpredictability .....	17
8.6	Opacity .....	18
8.7	Challenges related to the specification of AI systems .....	18
8.8	Challenges related to the implementation of AI systems .....	19
8.8.1	Data acquisition and preparation .....	19
8.8.2	Modelling .....	19
8.8.3	Model updates .....	21
8.8.4	Software defects .....	21

8.9	Challenges related to the use of AI systems .....	21
8.9.1	Human-computer interaction (HCI) factors .....	21
8.9.2	Misapplication of AI systems that demonstrate realistic human behaviour .....	22
8.10	System hardware faults .....	22
9	Mitigation measures .....	23
9.1	General .....	23
9.2	Transparency .....	23
9.3	Explainability .....	24
9.3.1	General .....	24
9.3.2	Aims of explanation .....	24
9.3.3	Ex-ante vs ex-post explanation .....	24
9.3.4	Approaches to explainability .....	25
9.3.5	Modes of ex-post explanation .....	25
9.3.6	Levels of explainability .....	26
9.3.7	Evaluation of the explanations .....	27
9.4	Controllability .....	27
9.4.1	General .....	27
9.4.2	Human-in-the-loop control points .....	28
9.5	Strategies for reducing bias .....	28
9.6	Privacy .....	28
9.7	Reliability, resilience and robustness .....	28
9.8	Mitigating system hardware faults .....	29
9.9	Functional safety .....	29
9.10	Testing and evaluation .....	30
9.10.1	General .....	30
9.10.2	Software validation and verification methods .....	30
9.10.3	Robustness considerations .....	32
9.10.4	Privacy-related considerations .....	33
9.10.5	System predictability considerations .....	33
9.11	Use and applicability .....	34
9.11.1	Compliance .....	34
9.11.2	Managing expectations .....	34
9.11.3	Product labelling .....	34
9.11.4	Cognitive science research .....	34
10	Conclusions .....	34
Annex A (informative) Related work on societal issues .....		36
Bibliography .....		37