ISO/IEC TS 12791:2024-10 (E)

Information technology - Artificial intelligence - Treatment of unwanted bias in classification and regression machine learning tasks

Contents

Page

Foreword ... **iv**

Introduction ... **v**

1 Scope ... **1**

2 Normative references ... **1**

3 Terms and definitions ... **1**
3.1 General ... 1
3.2 Artificial intelligence ... 3
3.3 Bias ... 4
3.4 Testing ... 5

4 Abbreviated terms ... **6**

5 Treating unwanted bias in the AI system life cycle ... **6**
5.1 Inception ... 6
5.1.1 Stakeholder identification ... 6
5.1.2 Stakeholder needs and requirements definition ... 7
5.1.3 Procurement ... 8
5.1.4 Data sources ... 9
5.1.5 Integration with risk management ... 11
5.1.6 Acceptance criteria ... 11
5.2 Design and development ... 12
5.2.1 Feature representation ... 12
5.2.2 Metadata sufficiency ... 12
5.2.3 Data annotations ... 12
5.2.4 Adjusting data ... 13
5.2.5 Methods for managing identified risks ... 13
5.3 Verification and validation ... 13
5.3.1 General ... 13
5.3.2 Static testing of data used in development ... 14
5.3.3 Dynamic testing ... 14
5.4 Re-evaluation, continuous validation, operations and monitoring ... 15
5.4.1 General ... 15
5.4.2 External change ... 16
5.5 Disposal ... 17

6 Techniques to address unwanted bias ... **17**
6.1 General ... 17
6.2 Algorithmic and training techniques ... 17
6.2.1 General ... 17
6.2.2 Pre-trained models ... 18
6.3 Data techniques ... 19

7 Handling bias in a distributed AI system life cycle ... **19**

Annex A (informative) **Life cycle processes map** ... **21**

Annex B (informative) **Potential impacts of unwanted bias on different types of specific user** ... **22**

Bibliography ... **23**