ISO/IEC TR 29119-11:2020 (E)

Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems

Contents

Foreword

Introduction

1 Scope

2 Normative references

3 Terms, definitions and abbreviated terms

3.1 Terms and definitions
3.2 Abbreviated terms

4 Introduction to AI and testing

4.1 Overview of AI and testing
4.2 Artificial intelligence (AI)
4.2.1 Definition of ‘artificial intelligence’
4.2.2 AI use cases
4.2.3 AI usage and market
4.2.4 AI technologies
4.2.4.1 General
4.2.4.2 Robots and software agents
4.2.4.3 AI and autonomous systems
4.2.5 AI hardware
4.2.6 AI development frameworks
4.2.7 Narrow vs general AI
4.3 Testing of AI-based systems
4.3.1 The importance of testing for AI-based systems
4.3.2 Safety-related AI-based systems
4.3.3 Standardization and AI
4.3.3.1 Introduction to AI standardization
4.3.3.2 Regulatory standards for AI
4.3.3.2.1 General
4.3.3.2.2 Non-safety-related regulatory standards
4.3.3.2.3 Safety-related standards
4.3.3.3 The AI quality metamodel

5 AI system characteristics

5.1 AI-specific characteristics
5.1.1 General
5.1.2 Flexibility and adaptability
5.1.3 Autonomy
5.1.4 Evolution
5.1.5 Bias
5.1.6 Complexity
5.1.7 Transparency, interpretability and explainability
5.1.8 Non-determinism
5.2 Aligning AI-based systems with human values
5.3 Side-effects
5.4 Reward hacking
5.5 Specifying ethical requirements for AI-based systems

6 Introduction to the testing of AI-based systems

6.1 Challenges in testing AI-based systems

6.1.1 Introduction to challenges testing AI-based systems
6.1.2 System specifications
6.1.3 Test input data
6.1.4 Self-learning systems
6.1.5 Flexibility and adaptability
6.1.6 Autonomy
6.1.7 Evolution
6.1.8 Bias
6.1.9 Transparency, interpretability and explainability
6.1.10 Complexity
6.1.11 Probabilistic and non-deterministic systems
6.1.12 The test oracle problem for AI-based systems
6.2 Testing AI-based systems across the life cycle
6.2.1 General
6.2.2 Unit/component testing
6.2.3 Integration testing
6.2.4 System testing
6.2.5 System integration testing
6.2.6 Acceptance testing
6.2.7 Maintenance testing

7 Testing and QA of ML systems

7.1 Introduction to the testing and QA of ML systems
7.2 Review of ML workflow
7.3 Acceptance criteria
7.4 Framework, algorithm/model and hyperparameter selection
7.5 Training data quality
7.6 Test data quality
7.7 Model updates
7.8 Adversarial examples and testing
7.9 Benchmarks for machine learning

8 Black-box testing of AI-based systems

8.1 Combinatorial testing
8.2 Back-to-back testing
8.3 A/B testing
8.4 Metamorphic testing
8.5 Exploratory testing

9 White-box testing of neural networks

9.1 Structure of a neural network
9.2 Test coverage measures for neural networks
9.2.1 Introduction to test coverage levels
9.2.2 Neuron coverage
9.2.3 Threshold coverage
9.2.4 Sign change coverage
9.2.5 Value change coverage
9.2.6 Sign-sign coverage
9.2.7 Layer coverage
9.3 Test effectiveness of the white-box measures
9.4 White-box testing tools for neural networks

10 Test environments for AI-based systems

10.1 Test environments for AI-based systems
10.2 Test scenario derivation
10.3 Regulatory test scenarios and test environments

Annex A Machine learning

A.1 Introduction to machine learning
A.2 The machine learning workflow
A.2.1 Machine learning workflow overview
A.2.2 Understand the objectives
A.2.3 Select a framework
A.2.4 Build and compile the model

A.2.5 Source the data
A.2.6 Pre-process the data
A.2.7 Train the model
A.2.8 Evaluate the model
A.2.9 Tune the model
A.2.10 Test the model
A.2.11 Deploy the model
A.2.12 Use the model
A.2.13 Monitor and tune the model
A.3 Machine learning training and test data
A.4 Overfitting and underfitting in machine learning
A.4.1 Overfitting
A.4.2 Underfitting
A.5 Data quality
A.5.1 Data completeness
A.5.2 Data labelling
A.5.3 Distributional shift
A.6 Machine learning algorithm/model selection
A.7 Documenting ML systems
A.8 Machine learning performance metrics
A.8.1 Introduction to ML performance metrics
A.8.2 Confusion matrix
A.8.3 Accuracy
A.8.4 Precision
A.8.5 Recall
A.8.6 F1-score
A.8.7 Selection of performance metrics

Page count: 52