PD ISO/IEC TR 24029-1:2021

BSI Standards Publication

Artificial Intelligence (AI) — Assessment of the robustness of neural networks

Part 1: Overview

National foreword

This Published Document is the UK implementation of ISO/IEC TR 24029-1:2021.

The UK participation in its preparation was entrusted to Technical Committee ART/1, Artificial Intelligence.

A list of organizations represented on this committee can be obtained on request to its committee manager.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

Published by BSI Standards Limited 2021

ISBN 978 0 539 17398 7

ICS 35.020

Compliance with a British Standard cannot confer immunity from legal obligations.

This Published Document was published under the authority of the Standards Policy and Strategy Committee on 31 March 2021.

Amendments/corrigenda issued since publication

Date	Text affected

PD ISO/IEC TR 24029-1:2021

TECHNICAL REPORT

ISO/IEC TR 24029-1

First edition
2021-03

Artificial Intelligence (AI) — Assessment of the robustness of neural networks —

Part 1: Overview

Reference number
ISO/IEC TR 24029-1:2021(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Overview of the existing methods to assess the robustness of neural networks **3**
4.1 General 3
4.1.1 Robustness concept 3
4.1.2 Typical workflow to assess robustness 3
4.2 Classification of methods 6
5 Statistical methods **7**
5.1 General 7
5.2 Robustness metrics available using statistical methods 8
5.2.1 General 8
5.2.2 Examples of performance measures for interpolation 8
5.2.3 Examples of performance measures for classification 9
5.2.4 Other measures 13
5.3 Statistical methods to measure robustness of a neural network 14
5.3.1 General 14
5.3.2 Contrastive measures 14
6 Formal methods **14**
6.1 General 14
6.2 Robustness goal achievable using formal methods 15
6.2.1 General 15
6.2.2 Interpolation stability 15
6.2.3 Maximum stable space for perturbation resistance 15
6.3 Conduct the testing using formal methods 16
6.3.1 Using uncertainty analysis to prove interpolation stability 16
6.3.2 Using solver to prove a maximum stable space property 16
6.3.3 Using optimization techniques to prove a maximum stable space property 16
6.3.4 Using abstract interpretation to prove a maximum stable space property 17
7 Empirical methods **17**
7.1 General 17
7.2 Field trials 17
7.3 A posteriori testing 18
7.4 Benchmarking of neural networks 19
Annex A (informative) **Data perturbation** **20**
Annex B (informative) **Principle of abstract interpretation** **25**
Bibliography **26**