ISO/IEC JTC 1/SC 42/WG 3 **N 4285**

ISO/IEC JTC 1/SC 42/WG 3 "Trustworthiness"
Convenorship: **NSAI**
Convenor: **Filip David Dr**

PWI_42118_Outline for Reliability of AI systems-v5-PDF version

Document type	Related content	Document date	Expected action
Project / Other	Project: ISO/IEC PWI 42118	2024-08-26	**COMMENT/REPLY** by 2024-09-06

Outline for Form 4 – New Work Item Proposal:

Information Technology – Artificial Intelligence – Reliability assessment of AI systems

Introduction

With the wide spread roll out of AI systems in every aspect of human life, it is important to assess the reliability of AI systems before and during each real-world deployment. This is especially important for AI systems that affect various aspects of human life, such as health-care, robotic surgery, autonomous vehicles, senior citizen monitoring and care, citizen-welfare services, robotic automation etc. Reliability ~~evaluation~~ assessment is crucial because failure events of an AI system can lead to business loss, information loss, human injury and other safety issues. By measuring the reliability of AI systems for failure free operation for a specified period of time under stated conditions, one can be x% confident that the system would function well as required and not create failures and faults during its run. With large-scale deployment of an AI system, ~~evaluating~~ assessing reliability can also help system administrators to have a level of confidence in the functioning of that AI system before roll-out and during its life-cycle.

Reliability assessment focuses on ~~evaluating~~ estimating how well the AI system can perform its designed functionality without failure, for the intended period of time, under given conditions for operational profiles.

Reliability models, can give a predictive measure that the system would function at a level of performance for a period of time in a given environment. High reliability can help consumers and users be confident of the AI system against potential failures during run-time of the system. This is important for all AI systems, especially the ones that can have a direct impact on human life and safety. Reliability is estimated by analysing all failure data of the system, using statistical modelling techniques leading to building an estimate of the potential future failure prediction in various scenarios.

Reliability can be viewed as related to quality and testing of AI systems, but is a very different aspect. While testing determines whether the AI system's output matches the expected output, reliability ~~predicts~~ estimates the confidence in the system to function without failures for a specified period of time after development or deployment based on the ~~test results and~~ failure~~s~~ logs of the system. It is a time-based prediction of the performance of the AI system in real-world conditions. Quality, ~~on the other hand~~as described in ISO/IEC 25059:2023, has reliability as one of the characteristics of quality model of AI system. The Quality model, ~~measures the performance of~~evaluates the AI system based on functional and non-functional ~~specifications~~characteristics, and is a broader term that can include various aspects such as efficiency, effectiveness, functional adaptability, transparency, intervenability, societal and ethical risk mitigation etc.~~satisfaction, risk management~~ etc. . Reliability assessment does not determine whether an AI output is factually correct, fair, safe, secure, ethical or robust. In reliability assessment the focus is on estimating the failure-free operation of the AI system for a given time period in a specific context. However, because AI systems are often systems-of systems, or are embedded in other systems, issues such as malformed output, adversarially engineered data, or similar can cause system failures.

The importance of reliability assessment for AI systems is described in various literature, some of which are listed in the references of this outline.

This project describes methods and measures for ~~evaluating~~ assessing the reliability of AI systems so that it is measured and reported to the stakeholders. This can be done at any time after the ~~testing~~ development of the AI system or while it is being deployed or while it is in real-world use.

1 Scope

This document provides methods and mechanisms to ~~evaluate~~ assess the reliability of an AI system. It describes the metrics of reliability and the procedure for reliability assessment from a statistical perspective.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989 and the following apply.

a) **reliability**

property of consistent intended behaviour and results [ISO/IEC 22989:2022]

b) reliability level

measure of the reliability of an AI system for failure free operation for specified period of time under stated conditions.

c) reliability of an AI system

collective measure of the reliability level of the AI system in different stated modes of operation

4 Preconditions for the assessment of reliability of AI systems

45 Overview of statistical models for ~~Types of~~ reliability assessment ~~models~~

~~4.1~~5.1 Statistical models

~~4.2~~5.2 Nonhomogeneous Poisson process

Formatted: Indent: Left: 0.63 cm, No bullets or numbering

Formatted: Font: Bold

Formatted: Outline numbered + Level: 2 + Numbering Style: 1, 2, 3, … + Aligned at: 0 cm + Indent at: 0.63 cm

~~4.3~~5.3 **Spline models**
~~4.4~~5.4 **Musa's basic execution time model**
~~4.5~~5.5 **Musa-Okumoto model**
~~4.6~~5.6 **Gompertz model**
~~4.7~~5.7 **Weibull model**

~~5~~6 Sub-characteristics and Metrics of reliability assessment

~~6~~7 Methods to ~~measure~~ assess reliability of AI systems
7.1 **Statistical measurement**
7.2 **Analysis of failure models**
7.3 **Parameter estimation**
7.4 **Estimation of reliability**

~~7~~8 ~~Methods and m~~Mechanisms to ~~measure~~ assess reliability of AI systems with Machine Learning models

~~8~~9 ~~Methods and m~~Mechanisms to ~~measure~~ assess reliability of AI systems with Deep Learning models

~~9~~10 ~~Methods and m~~Mechanisms to ~~measure~~ assess reliability of AI systems with Reinforcement Learning models

Formatted: Indent: Left: 0.63 cm, No bullets or numbering

Appendix-A

A. Real-world examples and their results

Example 1:

In the reliability estimation and analysis conducted by Jie Min et.al. [5] on autonomous vehicles (AV), a detailed study was carried out on the AVs produced by Waymo, Cruise, Pony AI, and Zoox. The study shows strong correlation between the reliability estimates and the recurrent events data for these AVs. The study also shows that the Gompertz estimation model fits well with the events data of Waymo and Cruise, while for Pony AI the Weibull model fits, and for the Zoox the Musa-Okumoto model fits well. This variation in applicable estimation models can occur due to various reasons, such as the sample size, the driving speed when the event occurred, the environment (e.g., busy street versus highway), and vehicle event (failure) characteristics.

The operational profile is that of a test driver who marks a failure of the AV system as a disengagement event, which occurs when there is an autonomous vehicle failure indicated by a warning from the AV system, or when situations arise that require the test driver to take manual control of the vehicle to operate safely.

Formatted: Font: Not Bold

Example 2:

A business-rule engine was integrated into a banking product after extensive system-level and user-level testing, but without any measure on its reliability. When the reliability of its failure free operation was measured it was found out that the probability of failure free operation for 1 hour of continuous run of that rule engine was 10^{-8}. Later, the reliability was modelled for ensuring that the system would function failure-free for 24 hours of run with a probability of 85%, and it was discovered that there were 45 more potential failures hidden in that system that needed to be fixed. These failures were later identified with rigorous code analysis, use case analysis, user-modelling and testing. Once fixed the rule engine has been successfully running with no failures reported.

The operational profile is that of a business-rule engine user who defines and executes business rules for the banking product and failures are marked when the product does not function as expected.

B. Sample cases

Sample Case 1:

Formatted: Font: Bold

Formatted: Indent: Left: 1.27 cm

An AI agent used for proposing scheduling and booking locations for client meetings consists of a set of prompts and calls to a language model using retrieval-augmented generation from email systems of users, combined with access to their calendar for reading and proposing new events. In the system, unexpected combinations of inputs or a failure to identify the proper participants and calendar entries cause the system to propose overlapping or otherwise incorrect meetings. This is a reliability failure, as the system does not provide consistent intended behavior and results. By identifying the cases where this occurs, a reliability model may be built predicting probability of failure. Statistical methods can determine which model best fits the data, and when failure modes are identified and changed, system developers can expect measured reliability to increase. Changes to measured reliability can then be estimated by back-testing the system, and validated in production use.

Sample Case 2:

An AI system uses a commercial API for access to a large language model (LLM) as part of its operation. The domain performance of the system has increased with more recent models, and for this reason, the commercial API uses the latest version of the model. However, some past modifications of the model have introduced system failures, when a change to the LLM causes it to no longer produce conforming output, a failure which requires prompt engineering changes. Despite the fact that in the LLM this is a quality or functional adaptability issue, it is a reliability issue for the current system. Because fine tuning and model changes are relatively frequent but spaced erratically without knowledge of the model users, measuring the frequency of such breaking-changes is possible with traditional statistical measurement techniques which are applied to complex systems.

~~B.~~

Formatted: Normal, Indent: Left: 0.63 cm, No bullets or numbering

References

1 https://users.ece.cmu.edu/~koopman/des_s99/sw_reliability/

2 Anthony Corso, David Karamadian, Romeo Valentin, Mary Cooper and Mykel Kochenderfer, A Holistic Assessment of the Reliability of Machine Learning Systems, arXiv:2307.10586v2 [cs.LG] 29 Jul 2023

3 Yili Hong, Jiayi Lian, Li Xu, Jie Min, Yueyao Wang, Laura J. Freeman, and Xinwei Deng, Statistical Perspectives on Reliability of Artificial Intelligence Systems, https://arxiv.org/abs/2111.05391v1

4 Xingyu Zhao, Wei Huang, Alec Banks, Victoria Cox, David Flynn, Sven Schewe and Xiaowei Huang, Assessing the Reliability of Deep Learning Classifiers Through Robustness Evaluation and Operational Profiles, CEUR-WS.org/Vol-2916/paper_16.pdf

5 Jie Min, Yili Hong, Caleb B. King, and William Q. Meeker, Reliability Analysis of Artificial Intelligence Systems Using Recurrent Events Data from Autonomous Vehicles, Journal of the Royal Statistical Society Series C Applied Statistics, April 2022, DOI: 10.1111/rssc.12564

6 Japan Liaison Council of Three Ministries on Disaster Prevention of Petroleum Complexes, Guidelines on Assessment of AI Reliability in the Field of Plant Safety, Second edition, March 2021.

7 Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, Balaji Lakshminarayanan, Plex: Towards Reliability using Pretrained Large Model Extensions, https://arxiv.org/pdf/2207.07411

8 Saurabh Mishra, Amin Aria, Andrea Renda, OECD.AI Policy Observatory, Ensuring AI's trustworthiness: reliability engineering meets data-driven risk management, December 12, 2023, https://oecd.ai/en/wonk/reliability-engineering-data-driven-risk-management.

~~5~~9 RAND Research & Commentary Blog, Why Waiting for Perfect Autonomous Vehicles May Cost Lives Why Waiting for Perfect Autonomous Vehicles May Cost Lives, https://www.rand.org/pubs/articles/2017/why-waiting-for-perfect-autonomous-vehicles-may-cost-lives.html, 2017.