ATARC nnnn:2020 (E)

ATARC Working Group n

Secretariat: ATARC

Information Technology — Artificial Intelligence — Machine Learning (ML) model transparency

WD stage

Warning for WDs and CDs

This document is not a Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as a Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Copyright notice

[need to add copyright notice]

Contents Page

[1 Scope 1](#_Toc50708498)

[2 Normative references 1](#_Toc50708499)

[3 Terms and definitions 1](#_Toc50708500)

[3.1 General 1](#_Toc50708501)

[3.2 Terms relating to artificial intelligence 1](#_Toc50708502)

[3.3 Terms related to machine learning 3](#_Toc50708503)

[3.4 Terms related to neural networks 6](#_Toc50708504)

[3.5 Terms related to trustworthiness 7](#_Toc50708505)

[4 Conformance 7](#_Toc50708506)

[5 Transparency assessment overview 8](#_Toc50708507)

[6 Transparency assessment criteria 10](#_Toc50708508)

[6.1 Algorithm explainability 10](#_Toc50708509)

[6.2 Identification of data sources used for training 10](#_Toc50708510)

[6.3 Methods used for data selection 10](#_Toc50708511)

[6.4 Identification of data set bias and methods used for reduction 11](#_Toc50708512)

[6.5 Method and means by which model will be versioned 11](#_Toc50708513)

[7 Measurement and assessment techniques 11](#_Toc50708514)

[Annex A (informative) Sources and Contributors 12](#_Toc50708515)

[A.1 Sources 12](#_Toc50708516)

[A.2 Contributors and acknowledgments 12](#_Toc50708517)

[Annex B (informative) Rationale 13](#_Toc50708518)

[B.1 Overview 13](#_Toc50708519)

[B.2 A different take on transparency 13](#_Toc50708520)

Foreword

Introduction

This document is meant to illustrate a framework and assessment system by which machine learning models and artificial intelligence algorithms can be evaluated for transparency on a variety of factors. The intent is that this transparency assessment will give model consumers the ability to determine whether or not a given model will be appropriate and applicable to their situation with respect to visibility into a number of factors. These factors may be important to determine whether or not the model can be used for a particular application. They also aid adherence to principles of ethical artificial intelligence, such as responsible, equitable, traceable, reliable, and governable.

The transparency assessment is not meant to be a quantitative ranking or a qualititative rating. Rather, it’s meant to be a point-in-time evaluation of that particular model iteration’s transparency based on certain self-documented and self-declared factors.

This assessment is not meant to be applied by a third party, and this is not a certification. It is meant to be used by model creators to communicate, document, and illustrate visibility into various factors that will give further understanding into the origins and methods by which the model was created.

Information Technology — Artificial Intelligence — Machine Learning (ML) model transparency

# Scope

This document establishes assessment criteria for measuring and assessing machine learning (ML) models..

# Normative references

The following referenced documents are indispensable for the application 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 - Information Technology — Artificial Intelligence — Artificial Intelligence Concepts and Terminology

# Terms and definitions

## General

For the purposes of this document, the terms and definitions in the following apply.

ISO and IEC maintain terminological databases for use in standardisation at the following addresses:

* IEC Electropedia: available at http://www.electropedia.org/
* ISO Online browsing platform: available at <http://www.iso.org/obp>

## Terms relating to artificial intelligence

###

**artificial intelligence**

<system>capability of an engineered system to acquire, process and apply knowledge and skills

Note 1 to entry: knowledge are facts, information, and skills acquired through experience or education.

**artificial intelligence**

<engineering discipline>discipline which studies the engineering of systems with the capability to acquire, process and apply knowledge and skills

Note 1 to entry: knowledge are facts, information, and skills acquired through experience or education.

**AI system**

system using AI (3.2.2)

**automation
automated**

characteristic of a system where work is performed that might previously have been done by a living being and that is governed by rules determined outside of the system

Note 1 to entry: Such systems are subject to external control and oversight.

Note 2 to entry: Automation implies the (revocable) delegation to a machine of a specific and defined set of “skills”, operations, processes, or procedures.

**bayesian network**

probabilistic model that represents a set of variables and their conditional dependencies via a directed acyclic graph

**continuous learning
continual learning
lifelong learning**

Incremental training of an AI system that takes place on an ongoing basis while the system is running in production

explainability

property of an AI system that important factors influencing the prediction decision can be expressed in a way that humans would understand

**genetic algorithm**

algorithm simulating natural selection by creating and evolving a population of individuals (solutions) for optimization problems

**lifecycle**

evolution of a [system,](https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:29110:-1:ed-2:v1:en:term:3.62) product, [service,](https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:29110:-1:ed-2:v1:en:term:3.53) [project](https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:29110:-1:ed-2:v1:en:term:3.47)or other human-made [entity,](https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:29110:-1:ed-2:v1:en:term:3.25) from conception through retirement

[SOURCE: ISO/IEC/IEEE 15288]

transparency

open, comprehensive, accessible, clear and understandable presentation of information

[SOURCE: ISO 20294:2018, 3.3.11]

-OR-

openness about activities and decisions that affect stakeholders and willingness to communicate about these in an open, comprehensive, accessible, clear and understandable manner

verification

confirmation, through the provision of objective evidence, that specified requirements have been fulfilled

Note 1 to entry: Verification only provides assurance that a product conforms to its specification.

[SOURCE: ISO/IEC 27042:2015, 3.21]

validation

confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled

[SOURCE: ISO/IEC 27043:2015, 3.16]

## Terms related to machine learning

decision trees

supervised-learning model for which inference can be represented by traversing one or more tree-like structures

[SOURCE: ISO/IEC 23053, 3.12]

human-machine teaming

efficient and effective integration of human interaction with machine intelligence capabilities

Note 1 to entry: In contrast to automation, where a machine substitutes for human work, in some cases a machine will complement human work. This may happen as a side-effect of AI development, or a system might be developed specifically with the goal of creating a human-machine team. Systems that aim to complement human cognitive capabilities are sometimes referred to as intelligence augmentation.

long short-term memory networks

artificial recurrent neural network that addresses the problem of the vanishing gradient

[SOURCE: ISO/IEC 23053, 3.41]

machine learning

process using computational techniques to enable systems to learn from data or experience

[SOURCE: ISO/IEC 23053, 3.16]

reinforcement learning

task of building an ML model using a process of trial and reward to achieve an objective

Note 1 to entry: A reinforcement learning task can include the training of a machine learning model in a way similar to supervised learning plus training on unlabelled inputs gathered during the operation phase of the AI system. Each time the model makes a prediction, a reward is calculated, and further trials are run to optimize the reward.

Note 2 to entry: In reinforcement learning, the objective, or definition of success, can be defined by the system designer.

Note 3 to entry: In reinforcement learning, the reward can be a calculated number that represents how close the AI system is to achieving the objective for a given trial.

[SOURCE: ISO/IEC 23053, 3.18]

**retraining**

generation of new model parameters by applying different training data to a trained model

[SOURCE: ISO/IEC 23053 3.50]

**semi-supervised machine learning**

task of learning a function that makes use of both labelled and unlabelled data during training

Note 1 to entry: The training data for a semi-supervised machine learning task can include a majority of unlabelled inputs.

[SOURCE: ISO/IEC 23053, 3.19]

supervised machine learning

task of learning a function that maps an input to an output based on labeled example input-output pairs

 [SOURCE: ISO/IEC 23053, 3.20]

support vector machines

maximum-distance classification algorithm

[SOURCE: ISO/IEC 23053, 3.21]

trained model

result of model training

[SOURCE: ISO/IEC 23053 3.49]

training

process to establish or to improve the parameters of a machine learning model, based on a machine learning algorithm, by using training data

Note 1 to entry: for supervised learning, the machine learning model can be trained (learn from) data that is similar to input data.

Note 2 to entry: for transfer learning, the input data is not necessarily similar to the training data.

Note 3 to entry: for unsupervised learning, the machine learning model is trained (learns from) and makes inferences, or predictions, based on the same data.

[SOURCE: ISO/IEC 23053, 3.9]

training data

samples used to fit a machine learning model

[SOURCE: ISO/IEC 23053, 3.8]

unsupervised machine learning

task of learning a function that maps unlabelled input data to an output

[SOURCE: ISO/IEC 23053, 3.22]

## Terms related to neural networks

###

deep learning

approach to creating rich hierarchical representations through the training of neural networks with many hidden layers

Note 1 to entry: Deep learning uses multi-layered networks of simple computing units (or “neurons”). In these neural networks each unit combines a set of input values to produce an output value, which in turn is passed on to other neurons downstream.

[SOURCE: ISO/IEC 23053, 3.13]

convolutional / deep convolutional neural networks

feed forward neural networks which are using convolution instead of matrix multiplication in at least one of their layers

Note 1 to entry: Convolutional networks assume that their input have a grid-like topology, like images (2D grid) or time series (1D grid).

Note 2 to entry: Convolutional networks layers most often are made of a succession of Convolution, Pooling, and Rectified Linear Unit ReLU layers ended by a fully connected layer.

[SOURCE: ISO/IEC 23053, 3.14]

**feed forward neural networks**

artificial neural network where information is feed from the input to the output in one direction only

[SOURCE: ISO/IEC 23053, 3.15]

**neural network
neural net
artificial neural network**

network of primitive processing elements connected by weighted links with adjustable weights, in which each element produces a value by applying a nonlinear function to its input values, and transmits it to other elements or presents it as an output value

Note 1 to entry: Whereas some neural networks are intended to simulate the functioning of neurons in the nervous system, most neural networks are used in artificial intelligence as realizations of the connectionist model.

Note 2 to entry: Examples of nonlinear functions are a threshold function, a sigmoid function, and a polynomial function.

Note 3 to entry: This entry is an improved version of the entry 28.01.22 in ISO/IEC 2382-28:1995.

Note 4 to entry: neural network; neural net; NN; artificial neural network; ANN: terms, abbreviations and definition standardized by ISO/IEC [ISO/IEC 2382-34:1999].

Note 5 to entry: 34.01.06 (2382)

[SOURCE: ISO/IEC 2382:2015]

###

**recurrent neural network**

neural network where neurons are fed information not just from the previous layer, but also from themselves from the previous pass

Note 1 to entry: RNN are well suited to process sequential input data of variable length and to output sequential data of variable length.

Note 2 to entry: there are two common kinds of RNN which are the Long Short Term Memory networks (LSTM) and the Gated Recurrent Unit (GRU) networks. Each of the LSTM cells have both an internal memory and a hidden state. LSTM have been introduced to solve the vanishing gradient problem in RNN. GRU is a simpler variant of LSTM.

[SOURCE: ISO/IEC 23053, 3.17]

## Terms related to trustworthiness

**robustness**

ability of a system to maintain its level of performance under any circumstances

# Conformance

The following table identifies the conformance paradigms. \*\*\* NEED BETTER EXPLANATION \*\*\*

ML Model Transparency Metric Options

|  |  |  |  |
| --- | --- | --- | --- |
| Possible Components of Transparency | Description | Quantitative, Qualitative or Both? | Relevant to Individual Model, AI System, or Both? |
| Algorithm Explainability | In this factor, the Transparency Assessment measures the extent to which the algorithm used has explainable features for which documentation is provided with the algorithm.  | Both | Both |
| Identification of Data sources used for Training | In this factor, the Transparency Assessment measures the extent to which the sources of data used for training the model have been adequately or fully disclosed. | Qualitative | Individual Model |
| Methods used for Data Selection | In this factor, the Transparency Assessment measures the extent to which the sources of data used for training the model have been adequately or fully disclosed | Qualitative | Individual Model |
| Identification of Data Set Bias and Methods used for Reduction | In this factor, the Transparency Assessment measures the extent to which bias in a data source has been identified and the extent to which it has been mitigated. | Qualitative | Individual Model |
| Method and means by which model will be versioned | In this factor, the Transparency Assessment evaluates the method by which the model will be versioned and how model consumers will be notified of new model versions | Qualitative | AI System |
| Documentation of model purpose and impact assessment | What is the model’s purpose? Was there an analysis done on the impact this model might have on the people using it or being analyzed by it? | Qualitative | Individual Model |
| Origin of data and model | Where did the data come from? How was it trained and validated? Who created this model? | Qualitative | Individual Model |
| Technical performance | What are this model's performance metrics? How does it compare to models that are similar in function? | Quantitative | Both |
| Real-world inference | How does this model perform in production? (computational cost, throughput, latency, etc.) | Both | Both |
| Auditing | What predictions has this model made in the past? (auditing). Who is using this model and on what data/systems? (governance and insider threat question) | Both | AI System |

# Transparency assessment overview

\*\*\* TO BE SUPPLIED \*\*\*



**Figure 1 - Transparency Assessment**

In the above Transparency Assessment system, there are five main factors:

1. Algorithm Purpose
2. Algorithm Explainability
3. Identification of Data Sources used for Training
4. Methods used for Data Selection
5. Identification of Data Set Bias and Methods used for Reduction
6. Method and means by which model will be versioned

The Transparency assessment provides a quantitiatve rating from 1-5 on each of the above factors that determines the extent to which each qualitiative factor has been fully documented or the level to which such transparency is provided. An alternate visualization of the same model transparency is shown below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **METRIC** |  **0** |  **1** |  **2** |  **3** |  **4** |  **5** |
| Algorithm Explainability |  |  |  |  3 |  |  |
| Identification of Data Sources |  |  |  |  3 |  |  |
| Methods of Data Selection |  |  |  |  |  |  5 |
| Reduction of Data Set Bias |  |  |  2 |  |  |  |
| Model Versioning Method |  |  |  |  |  4 |  |

The specifics of each factor and means for ranking are further outlined below.

# Transparency assessment criteria

\*\*\* TO BE SUPPLIED \*\*\*

## Algorithm explainability

In this factor, the Transparency Assessment measures the extent to which the algorithm used has explainable features for which documentation is provided with the algorithm. For example, it is not sufficient to merely assert that an algorithm is explainable – rather, the documentation behind the assertion of explainability is provided with the algorithm whenever the algorithm is distributed.

1. No documented explainability in any form - complete black box
2. Algorithm used is identified, but algorithm provides no explainability
3. Algorithm has partial explainability, and some factors for explainability are documented
4. Fully explainable algorithm is used, but limited or no visibility into hyperparameters
5. Fully explainable algorithm is used and all hyperparameter configurations and settings have been documented

## Identification of data sources used for training

In this factor, the Transparency Assessment measures the extent to which the sources of data used for training the model have been adequately or fully disclosed.

1. No identification at all of training data set used
2. Training set data is identified but limited visibility into size and scope of data
3. Training set data is identified along with measures of its size and scope, but limited visibility into examples of training set data
4. Training set data is identified with full details on size and scope of training data and examples provided
5. Training set data is identified, examples provided, and a link to the full training set data is provided or other means to access the training set data is provided

## Methods used for data selection

In this factor, the Transparency Assessment measures the extent to which the sources of data used for training the model have been adequately or fully disclosed

1. No visibility provided into the methods or means that data is selected or not from the training data set
2. Limited visibility into the methods and means by which data has been selected for the training data set
3. The specific filtering approach, methods, or criteria for data inclusion or exclusion is defined and communicated, but no access is provided to filtering method
4. Full definition of data exclusion and inclusion provided, the specific filtering method specified and defined, but no access to the included and excluded data is provided
5. Full definition of data exclusion and inclusion provided, the specific filtering method specified and defined, and access to the included and excluded data is provided

## Identification of data set bias and methods used for reduction

In this factor, the Transparency Assessment measures the extent to which bias in a data source has been identified and the extent to which it has been mitigated.

1. No visibility or understanding provided into bias in training data sets
2. Some explanation of possible areas of data set bias, but no tests for actual bias
3. Explanation of areas of data set bias, and tests for actual potential bias in the data sets, but no mitigation for that bias identified
4. Explanation of data set bias, tests illustrated for bias, methods by which the data bias could be mitigated are explained, but not actually mitigated
5. Explanation of data set bias, tests illustrated for bias, methods by which the data bias could be mitigated are explained, and data bias mitigated according to the method(s) identified

## Method and means by which model will be versioned

In this factor, the Transparency Assessment evaluates the method by which the model will be versioned and how model consumers will be notified of new model versions

1. No method or visibility into the frequency of model versioning, nor is version history shared
2. Model version history is shared, but no visibility into the method or visibility into model versioning.
3. Model version history shared, and the frequency of which the model versioning is disclosed, but no method by which model versioning will be communicated to consumers is disclosed
4. Model version history shared, and the frequency of which the model versioning is disclosed, method by which model versioning is communicated to consumers
5. Full visibility and participation in model versioning continuous integration loop

# Measurement and assessment techniques

\*\*\* TO BE SUPPLIED \*\*\*

1. (informative)

Sources and Contributors
	1. Sources
* High-Level Expert Group on Artificial Intelligence. (2019). Ethics Guidelines for Trustworthy AI.
	1. Contributors and acknowledgments
* Ronald Schmelzer, Managing Partner and Principal Analyst, Cognilytica
* Kathleen Walch, Managing Partner and Principal Analyst, Cognilytica
* Anil Chaudhry, Chair, ATARC, Ethics and Responsible AI Working Group
* Michael Hauck, Data Scientist and Artificial Intelligence Consultant
* Seth Clark, Booz Allen Hamilton (and Modzy)
* Maëva Ghonda, President’s Entrepreneurial Fellow, UM Ventures
1. (informative)

Rationale
	1. Overview

\*\*\* TO BE SUPPLIED

* 1. A different take on transparency

As an alternative option to consider, below is another taxonomy for how we might think about AI Transparency. At its core, transparency is about ensuring that anyone who is looking for information has the ability to uncover that information, at will. A good example of this is the Sunlight Foundation’s mission statement: “The Sunlight Foundation is a national, nonpartisan, nonprofit organization that uses civic technologies, open data, policy analysis[[1]](#footnote-1) and journalism[[2]](#footnote-2) to make our government and politics more accountable and transparent to all.”

With that perspective in mind, AI Transparency may be more about clear communication and exceptional documentation than it is about technical innovation. For instance, one could create a very poor AI model, but it could still receive a high transparency assessment if the model and all of it’s metadata were made available to the public.

Transparency in...

* Purpose
	+ What is the model’s purpose?
	+ Was there an analysis done on the impact this model might have on the people using it or being analyzed by it?
* Origin
	+ Where did the data come from?
	+ How was it trained and validated?
	+ Who created this model?
* Technical performance
	+ What are this model's performance metrics?
	+ How does it compare to models that are similar in function?
* Real-world inference
	+ How does this model perform in production? (computational cost, throughput, latency, etc.)
	+ How does the model work on a range of different data sets? (more transferrable models are better)
	+ What systems and data is this model being used on?
	+ Have impacted parties been informed that AI is being used?
	+ How are the model’s performance metrics changing over time?
* Auditing
	+ What predictions has this model made in the past? (auditing)
	+ Who is using this model and on what data/systems? (governance and insider threat question)
	+ How did the model arrive at the output it generated (prediction, classification, etc.)? (Not just explainability, necessarily, decision trees as well as publishing prediction distributions can help answer this question)
1. <http://sunlightfoundation.com/policy/> [↑](#footnote-ref-1)
2. <http://sunlightfoundation.com/?s=investigations> [↑](#footnote-ref-2)