EC3.21

# Methodological Guideline for Model ODD Characterization

**L3.5.2.3**

contact@confiance-ai.fr | www.confiance.ai

**Document reference: 321AA**

## Contributors

| | Name | Organisation | Role |
|---|---|---|---|
| Responsible for the deliverable | Paul-Marie RAFFI | IRT SystemX | Research Engineer |
| Scientific responsible | Fateh Kaakai | Thales | Research Director |
| Co-authors | Paul-Marie RAFFI | IRT SystemX | Research Engineer |
| | Thibault ROYET | IRT SystemX | Research Engineer |

## Document Control

| Revision | Date | Commentary | Author |
|---|---|---|---|
| v1.0 | 13/11/2023 | Initialization | Paul-Marie RAFFI |
| v1.1 | 18/12/2023 | Delivery | Paul-Marie RAFFI |
| | | | |

# Contents

# A. Introduction and abstract

## A.1. General introduction to trustworthy AI challenges

Trustworthiness in AI within critical systems (systems that can directly or indirectly affect human life and moral entities) is essential for its widespread adoption (by the industry, the decision makers, the general public, etc.) and poses the following significant challenges.

- First, how to design AI models, so that, by construction, they satisfy trustworthy properties (accuracy, robustness. . . ).
- Secondly, how to characterize these AI models, for example to understand and explain their behavior and their adequacy to the operational domain.
- Then, how to implement and embed those AI models on hardware, by making them fit for the target without losing their trustworthy properties.
- Another question is, what methods of data engineering to apply in order to, among other topics, manage important volumes of data and adapt to the evolution of the operational domain.
- At system level, what verification and certification processes to consider specifically for AI-based systems.
- Finally, a federation of all these matters is necessary to build an end-to-end methodological approach, supported by a consistent engineering environment compatible with industrial practices.

These are the challenges, among others, that the Confiance.ai program addresses.

## A.2. Introduction to MOODD

### A.2.1 Rationale for this methodology

Rule-based Model ODD Characterization is part of the methodologies that can be used to study the model performance for any ODD parameter that can be synthetically added in a range of intensities to augment samples of the training dataset in a controlled way. This allows to compute performance metrics that can be used to compare models for non regression purposes, or to train several models by varying parameters and then pick the model with the best performance, or to define ODD parameter ranges where the model performs well and can operate safely.

### A.2.2 Scientific Challenges

This document addresses the scientific challenge How to define the monitoring requirements on a model

### A.2.3 Trustworthiness Attributes

Robustness

### A.2.4 Target audience and disclaimer

This document has been written in a focus to be easy to comprehend for any reader. It is intended to give rule-based methods for:

- Data Scientists and Data Managers who want to characterize the robustness of their model to anomalies.
- AI Algo Engineers and Safety Engineers who want to control the inputs and outputs of AI models in operation. This method is indeed a prerequisite to rule-based Monitoring.

### A.2.5 Glossary and terminology

## A.3. Operational Design Domain (ODD) Definition

According to the taxonomy document developed by EC2 [L1.1.1 - 1st version of Confiance.ai Taxonomy], the Operational Design Domain (ODD) is defined as follows:

#### A.3.0.0.1 ODD definition in EC2 Taxonomy

*Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristic* SAE J3016 (2018).

This definition in EC2 taxonomy document is coming from the automotive domain (Road Motor Vehicle Automated Driving Systems) has been considered by project EC6 to develop a top-down approach to specify ODD from system considerations.

#### A.3.0.0.2 Concept of AI Model ODD

Now, if we assume that this "ODD as specified" exists and has driven the development of an AI model as depicted in in Figure A.1. Let also assume that the AI Model has been verified against its allocated requirements and is ready for deployment in the field. Therefore, at the end of the development process, we obtain what we call an "ODD of the AI Model as designed, implemented and verified)" or simply and shortly "AI Model ODD". To characterize this AI Model ODD, we need to measure the performance of the AI Model in operational conditions as close as possible to real one (i.e., nominal and abnormal conditions). By using inference results of the AI Model to input data containing nominal values and anomalies, we characterize the AI Model ODD through the measured ranges of parameters where the AI Model provides a correct output within an acceptable margin error. This AI Model ODD will be the baseline for the online monitoring capability since it corresponds really to the domain where the AI Model guarantee the expected behavior and the expected level of performance (since it has been verified against its allocated requirements during the development process). At this point, many questions arise like: *does the AI Model ODD perfectly fit with the specified ODD? Is the AI Model ODD a superset or a subset of the specified ODD?* All these questions are legitimate and should be managed by relevant process interface between the system engineering processes and AI Model development processes (EC2, EC4 and EC6 scopes). In this report, we develop more deeply this concept of AI Model ODD since it is a key concept for the online monitoring capability.

#### A.3.0.0.3 Principles of the AI Model ODD

There are various potential classes of anomalies depending on the data type (images, time series,

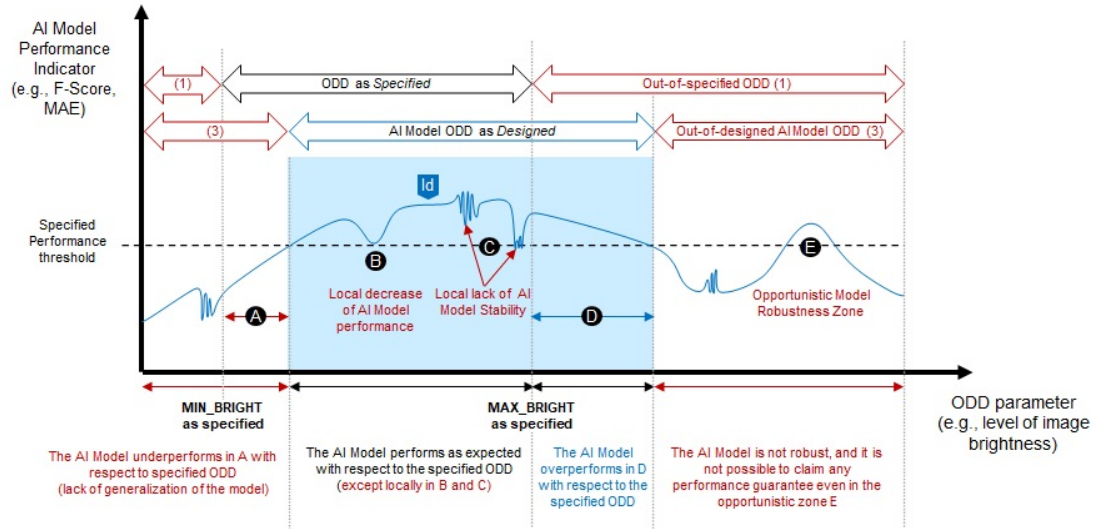


Figure A.1: Specified ODD at system level and AI Model ODD as designed

natural language, ...) and on the type of problem the industrial is trying to address (classification, object detection, semantic segmentation, regression task, ...). For example, the Renault welding Use Case is a problem of image classification into 3 classes: "Compliant welding", "non-compliant welding", or "Unknown". Based on the system ODD as specified with the use case owner, we identified potential anomalies of interest for this type of problem (images with too much color, blur, translation, rotation, brightness, darkness, ...) and defined ranges and increment steps (sampling) for these anomalies. Some anomalies such as Rotation can be selected in their whole range, from 0 degree to 360 degrees, and only the increment step is chosen. Other anomalies such as Blur or Brightness cannot be estimated easily until tested against the AI Model during a calibration phase (few cycles of tests before finding a range of interest). There are also hardware limitations that can reduce the number of steps. A trade-off has been found for each class of anomalies between storage, time of computation and performance of the Model ODD.

### A.3.1 Overall vision, context, motivation, objective, added value of the proposed method

This method is generic and the only constraint is that a single metric should be chosen to represent the inference score. This is easy for image classification, but not for object detection where metrics are often given per class. The goal of this method is to characterize macro and micro trends of the model performance for variations of an ODD parameter. This method has been originally designed to automatize the calibration of a rule-based monitor. But it can also be used to highlight robustness or stability weaknesses in a model, to guide model robustness improvements, or to measure if existing efforts to improve the model robustness meet a certain level, or to compare models for non regression of robustness and stability.

# B. Description of the method

## B.1. Position in the End-to-End approach

C.9 “Evaluation of ML Model"
C.9.2.3 Test of ML Model robustness by sampling and perturbation

## B.2. Prerequisites

- The ODD parameters of interest should be chosen.
- A subset of the training dataset should be selected.
- For each ODD parameter, a method to perturb the subset should be chosen.
- For each ODD parameter a range of intensity levels of perturbation should be defined. The intensity level should correspond to an input parameter of the perturbation method. This range should be wide enough to contain all the possible values that can be found in operation.
- Choose a metric for inference evaluation. This metric can depend on the type of data and the type of problem. For image classification, F1 score or inference score of the predicted class can be used. For object detection, mean average precision can be more appropriate.
- Choose a tolerance threshold: a percentage that this metric should maintain for safe operation of the system.

## B.3. Level of maturity

This method has been implemented and tested on Confiance use cases of image classification, object detection, time series regression and time series classification.

## B.4. MOODD Guidelines

There are 4 steps once an ODD parameter is defined with its range as detailed in the prerequisites:

- Generate N*R samples with the range R of perturbation from the subset of N samples
- Run inference on those N*R perturbed samples
- Compute the metric of inference evaluation for each perturbed sample. For each range step k from 1 to R, compute the mean or median value of the metric on the N samples of the subset. This will give a curve C containing the reference metric on the range R.
- Analyze the curve C for trends of stability, instability and zones where the tolerance threshold is maintained.

Steps 1, 2 and 3 are presented in Figure B.1 and give as output a Curve C representing the evolution of the reference metric per perturbation intensity.
This curve C is then analyzed by a set of rules to identify high level and low level zones.
High level zones:

- Model ODD Zone: where the model performance is mostly above the tolerance threshold.

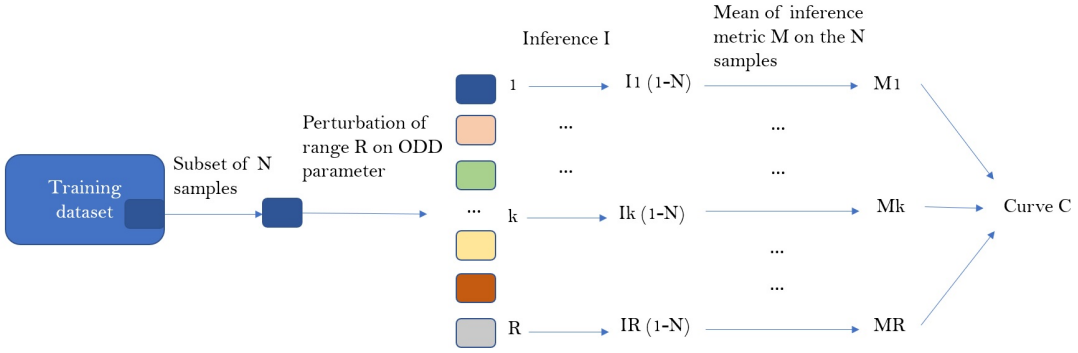


Figure B.1: Computing inference metric curve

- Out of Model ODD Zone: where the model performance is mostly below the tolerance threshold.

Low level zones:

- Lack of stability: instability zone due to oscillations or sudden change of regime.
- Local lack of robustness: zone where the model performance is below the tolerance threshold
- Model stability zone: zone where the model performance is above the tolerance threshold and relatively close to the base value, without instability or sudden change of regime
- Opportunistic ODD zone: zone where the model performance is above the tolerance threshold but this zone is separated from the stability zone(s) close to the base by one or several local lack of robustness zone(s).

Figure B.2 shows those high level and low level zones in a fictive curve C. When possible, the Business-driven ODD should be defined on the same range of perturbation intensities of the ODD parameter. This would allow to compare the Business-driven ODD with the high level zones. Some comparison zones could then be defined where:

- The Model performs as expected (B): When a portion of the Business-driven ODD zone fits with a portion of the Model ODD zone. The Model performs as expected except locally in (E) and (F) due to small lacks of stability.
- The Model under performs (A): When a portion of the Business-driven ODD zone fits with a portion of the Out of Model ODD zone.
- The Model over performs (C): When a portion of the Business-driven Out of ODD zone fits with a portion of the Model ODD zone.
- The Model is not robust (D): When a portion of the Business-driven Out of ODD zone fits with a portion of the Model Out of ODD zone. The Model performance is not trusted in this zone, even in an opportunistic ODD zone such as (G).

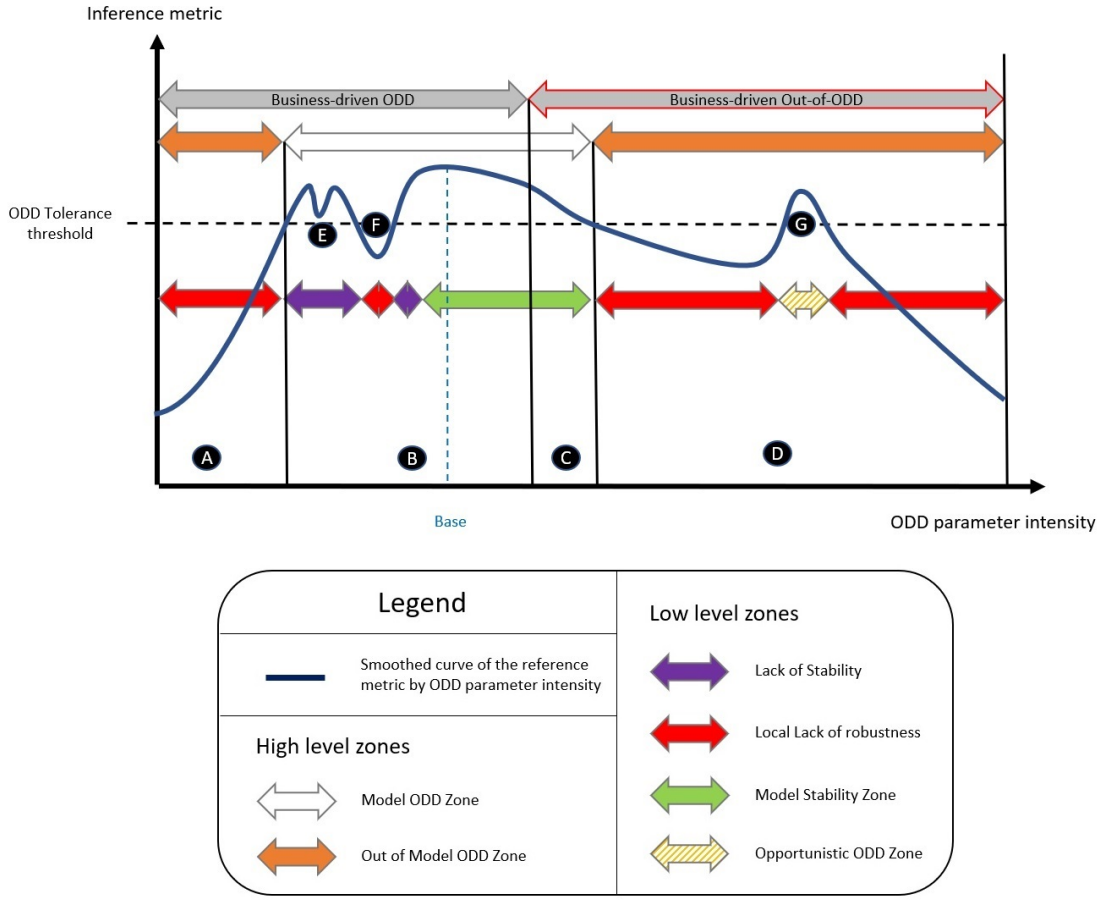


Figure B.2: Model ODD zones compared with Business-driven zones

# C. Conclusion

This method shows an easy way to characterize of the Model ODD of a given ODD parameter. It can be implemented with simple rules or a mix of rules and statistics to automatize the definition of zones of ODD parameter intensity where the model can infer safely on input data. It can be used for AI systems comparison by comparing their Model ODD. It can also be used to verify if a given ODD parameter requires to be monitored in the input of the AI system. Another usage is to automatize rule-based monitoring calibration as detailed in the document 321BA.

# Bibliography

SAE J3016 (2018). Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems.

Confiance ai

Title: Methodological Guideline for Model ODD Characterization

Keywords: ODD, rule-based

This guide presents a rule-based and top-down approach to characterize the ODD from a prediction model and its training dataset. This approach can be cut in several steps: choice of anomalies (ODD parameters), definition of anomalies levels, creation of datasets containing those anomalies levels, inference on those datasets and inference results analysis from predefined rules to characterize zones of anomalies levels where the model prediction can be trusted or not.

Our partners:

Air Liquide AIRBUS Atos cea Inria
NAVAL GROUP GROUPE RENAULT SAINT EXUPÉRY SAFRAN sopra steria
SystemX THALES Building a future we can all trust Valeo