Deliverable D2.1
Methodological framework

METRICS Metrological evaluation and testing of robots in international competitions

Lead contractor for this deliverable: LNE
Contributors: METRICS WP2 partners

Due date of deliverable: April 30, 2020
Submission date: April 30, 2020

Dissemination level : Public
871252 — METRICS — H2020-ICT-2018-20/H2020-ICT-2019-2

Contents

1 Introduction

1.1 Context

METRICS is a H2020 project which aims to organize evaluation campaigns so as to assess the technological maturity of robotic and Artificial Intelligence (AI) systems. The project is coordinated by the French national laboratory for metrology and testing (LNE) in partnership with sixteen European organizations specialized in the evaluation of intelligent systems and in the organization of competitions. Started in 2020, the project will last three years.

In recent years, robotics competitions have become increasingly popular in Europe, in particular thanks to the RoCKIn, euRathlon and EuRoC projects, whose methodologies have been harmonized and formalized within the RockEU2 project and have led to the European Robotics League (ERL) competitions, now supported by the SciRoc project. Within METRICS, partners from these projects have joined forces with organizers of other robotics competitions (RoboCup, Robotex, ROSE challenge, etc.) and AI competitions (Quaero, Repere, etc.), as well as metrologists specialized in intelligent systems and experts from the Digital Innovation Hubs (DIH).

The objective of METRICS is to jointly address a twofold challenge:

- Organize challenge-led and industry-relevant competitions in the four Priority Areas (PAs) defined by the European Commission: Healthcare, Inspection and Maintenance (I&M), Agri-Food, and Agile Production;
- Further develop the evaluation methodology to maximize the reproducibility of experiments and the repeatability of performance measurements, to serve as a reference in future competitions.

During the three years of the project, there will be two competitions per PA, per year:

- A field competition, in which the physical devices are tested in realistic operating environments (i.e. physical test-beds);
- A cascade competition, in which software is tested on data generated during the field competition.

All the competitions will be designed in a similar spirit: the first year is a dry-run that allows validating the evaluation procedure. After this, a competition will be organized once a year for the two remaining years. Participation to the METRICS competition is on a voluntary basis. METRICS participants are allowed to participate in one of the two evaluation campaigns without participating in the other.

1.2 Purpose of the document

This document describes the common evaluation framework that will be applied during METRICS competitions. **It outlines the points of importance to be respected in the creation of any evaluation plan for competitions**. It takes the form of a checklist, where each item offers methodological recommendation for the design and implementation of a rigorous evaluation process. These recommendations are presented in the final section of the document. These recommendations are meant to be general and non-specific to a type of competition, although they may be illustrated through examples from the competitions organized in METRICS.

This document is a first draft of the common evaluation framework built at the very beginning of the project. The objective of the work performed here is to set the methodology for the design of the framework, and competition organizers in each PA are expected to proof-use it when designing their first dry-run evaluation plan. The common evaluation framework is expected to evolve up to a consolidated framework at the end of the project (2023), by leveraging the expertise acquired throughout the years of competition. This document does not consider the overall organization of competitions (logistics, rules for participants, etc.), which is broached in the Deliverable D2.2 Good practice guide for competitions. This document does not supersede the evaluation plan expected for each competition.

2 METRICS overall concept

2.1 Competition, evaluation campaign, evaluation plan

In the present document, both expressions are used: competition and evaluation campaign.

A **competition** is an event that brings together different stakeholders of a field, around the evaluation of one or more characteristics of a product designed by competitors. This event attracts a public, that will either attend the competition if it takes place physically, or study the results if it is virtual. The objectives of the competition are set by the organization team and potentially sponsors or funders. An organization team design the competition procedure and is responsible for its implementation.

An **evaluation campaign** represents the process of evaluating products either vertically (by observing a range of products at a given time) and/or horizontally (by observing the evolution of the product over time). The competitions in METRICS host three evaluation campaigns - one per year - for each PA. METRICS proposes a framework that allows both vertical and horizontal evaluation.

The evaluation campaign must rely on an evaluation plan, a document that details the features of:

- One or more evaluation tasks that focuses on a device or software performing a specific action;
- Characteristics that need to be measured or estimated (performance, quality, safety, explainability, etc.);
- Metrics, that is to say formula that allow the production of scores (e.g. accuracy, precision, recall, F-measure);
- Test data or test environments (datasets or testbeds);
- Evaluation tools (software for data collection, visualization, comparison).

A rule book is drawn up for the attention of the participating teams, specifying the conditions of participation and the logistics specific to the competition. The rule book also includes all the elements of interest related to the evaluation presented in the evaluation plan.

2.2 METRICS objectives

METRICS competitions are expected to present several features, presented below, that highlight the excellence of the competitions:

- **Scientific**: While preserving the demonstration aspect typically associated with competitions, METRICS competitions are based on the scientific criteria of objectivity, repeatability and reproducibility, and respect the requirements of metrological rigor;
- **Benchmark-based**: The robots are evaluated through benchmarks, which means that they perform well-specified tests in realistic environments or on databases, and that their performance is assessed by applying quantitative metrics;
- **Modular**: METRICS does not only evaluate the robots as a whole. The elements constituting the robot's architecture are broken down into Functionalities (e. g. obstacle detection), which are combined to perform more complex Tasks (e.g. semantic navigation). The evaluation thus consists in Functionality Benchmarks (FBMs) which focus on the evaluation of specific capabilities with a limited utility when used alone, and Task Benchmarks (TBMs) that evaluate more complex activities;
- **Periodical**: METRICS competitions are organized as recurring events offering each time a similar evaluation framework (similar testbeds, similar testing datasets, same evaluation tools, etc.). It enables the monitoring of the technological progress of the community of developers on the whole;

- **Structured**: The competition is structured to optimize effort and maximize impact: each evaluation campaign is an event for public dissemination (demonstration value), as a matchmaking event to connect participants with complementary competencies (e.g., a research group and a company), and as a scientific endeavour (providing the scientific community with a stable set of benchmarking experiments, which enables objective comparison of research results and can act as the seed for the definition of standards);
- **Synergic**: METRICS builds on the well-established framework originally created by RoCKIn and subsequently validated, perfected and extended by RockEU2 and SciRoc. METRICS also builds on the methodological foundation and practical experience underpinning many successful competitions, such as Quaero and ROSE;
- **Open**: Through its partners, METRICS creates a network which stimulates and supports end-users and industry engagement in the design, implementation and evaluation of robotic benchmarks. METRICS will produce and make publicly available high-quality evaluation tools and annotated datasets that research and industry can use to develop and fine-tune their own algorithms, systems and products. Existing and prospective actors gain access to difficult-to-obtain data with associated ground truth and to validated evaluation tools. Importantly, these METRICS by-products benefit the competition and promote its long-term sustainability: users of the METRICS open data and tools will naturally be inclined to participate to the competitions, thus creating a virtuous circle enabling the success of the competitions.

Some of these features relate to the general organization of the competitions, while others are more specialized on the evaluation process. Procedures for verifying the proper implementation of these evaluation recommendations are provided in the following chapter.

3 Common evaluation framework

Each following section presents a topic relative to the evaluation and/or competition process, for which several mandatory aspects are presented. All these aspects, when properly addressed, have a positive impact on the metrological rigor of the evaluation process. The objective for the competition organizer is to validate all the mandatory aspects when designing the evaluation plan for the competition.

3.1 Organization of the evaluation

3.1.1 The first occurrence of the competition is a dry-run

Associated objective(s): Scientific, Benchmark-based

Description: The dry-run phase is essential for the development of a rigorous evaluation campaign. While the official evaluation campaign allows testing the systems, the dry-run allows testing the evaluation tools.

During the dry-run, the evaluation is simulated. This means that the result of the evaluation in itself does not have any significance. It allows testing the evaluation tools, through for example verifying the means of collecting the system hypothesis, the syntactic validation of the hypothesis format when used in the evaluation software. This also allows spotting potential flaws in the metrics (significance of the results in respect with the initial desired measure, computability of the score, etc.).

To achieve that, the full evaluation campaign must thus be simulated. Several discrepancies between the official evaluation campaign and the dry-run are accepted, and these discrepancies can be accumulated:

- Very small samples of data;
- Only one system, which output data will be split so as to simulate two or more systems;
- Fully simulated outputs of the data (although reasonably realistic).

However, the organizer must make sure that the simulation is as close as possible to the expected following evaluation campaign. Which means that the evaluation protocol must be reasonably ready for use (description of the measures and scores, objectives of the evaluation, etc.), the system outputs should be reasonably similar to real results, the qualification procedure of the test must be applied.

3.1.2 The evaluation plan is formalized

Associated objective(s): Scientific, Benchmark-based

Description: In terms of methodological and metrological rigour, reproducibility and relevance, it is essential that the competition be based on a rigorous scientific argument. This means that when the organizers design a competition, they performed choices. The choices may concern the relevance of the evaluation tasks selected (for example, a scientific state of the art or a sponsor has shown that the development of a particular characteristic needs to be boosted). The choices may also concern the observability of a phenomenon, its measurability or the calculability of a score.

For the sake of traceability and rigour, we therefore recommend the production of a document that summarises the scientific reasoning associated with the creation of the evaluation procedure, and any preliminary research and argument. We recommend that this document also focuses on describing the procedure of evaluation, addressed mainly to the competition organizers themselves as a work document. Its nature is scientific and aims at detailing the methodological and scientific rigor of the evaluation.

This information may appear in the rule book of the competition, or in a separate document, in a publication or in work documents private among the competition organizers; the dissemination level is relative to the nature of the competition (public or private, etc.) and is thus not constrained. In METRICS competition, it is requested that this information be made publicly available.

3.2 Evaluation tasks

3.2.1 Each evaluation task is relevant for industry

Associated objective(s): Open

Description: Competitions should allow estimating the current state of technological development of the solutions, and also boost their development. Therefore competition organizers must make sure that the competition matches the industrial trends, in terms of need and demand. Industrial relevance can be ensured in a number of ways.

First, competition organizers may produce proof-of-concept videos that highlight the expected Task or Functionality the competition is expected to benchmark. These videos can be presented to industrial stakeholders, who would then provide feedback, potentially rank the TBM or FBM they would want to see in the competition. In the case where no videos are available, competition organizers can offer story board, or any type of relevant descriptions.

Alternatively (or additionally), involving sponsors can also be a way to ensure industrial relevance, if sponsors are offered to take part directly in the definition of the evaluation tasks.

3.2.2 The dependent and independent variable of each evaluation are identified

Associated objective(s): Scientific

Description: To preserve the experimental rigour of the evaluations, it is essential that the independent and dependent variables of each evaluation (FBM or TBM) be identified and made explicit in the evaluation plan. The dependent variable is the element that is observed or measured. Its value will evolve according to the testing environments, that is constituted of several factors that are controlled and modified by the competition organizers (these are the independent variables).

Although the identification of independent and dependent variables are common experimental knowledge, their proper identification, and their formalization in the evaluation plan, is a guarantee that the factors of influence and their relation to the observed element are properly defined, which maximizes the chance of properly addressing them in the evaluation. Controllable influencing factors should be controlled, while the other factors should be measured in order to draw sensitivity curves.

For example, to assess the performance of an object detection system through computer vision, the competition organizer may present several types of objects, or one object with specific selected features. In this context, the features of the objects are the independent variables, and the result produced by the system (for example, the name of the object), may be considered the dependent variable.

3.2.3 The evaluation is modular (FBM+TBM)

Associated objective(s): Modular

Description: Evaluating the overall performance of a robot system while performing a task is interesting for assessing the global behaviour of the application, but neither does it allow the evaluation of the contribution of each component, nor does it put in evidence which components are limiting system performance. On the other side, the good performance of each element in a set of components does not necessarily mean that a robot built with such components will perform well: system-level integration has, in fact, a deep influence on this, which is not investigated at all by component-level benchmarking.

For these reasons, the evaluations shall include two groups of benchmarks: Functionality Benchmarks (FBMs) and Task Benchmarks (TBMs):

- Functionality Benchmarks (FBMs): A functionality is conventionally identified by researchers as a self-contained unit of capability, which is too low-level to be useful on its own to reach a goal (e.g. self-localization, crucial to most applications, but aimless on its own). A functionality can be provided by a single component or by a set of components, and usually involves both hardware and software. A FBM is a benchmark that investigates the performance of a robot component when executing a given functionality. A Functionality Benchmark is as independent as possible from the other functionalities of the system, so as to control the functionality under test as the sole dependent variable in the evaluation;
- Task Benchmarks (TBMs): A Task is an activity of a robot system that, when performed, accomplishes a goal that is considered useful on its own. A task always requires multiple functionalities to be performed (e.g. finding and fetching an object, which involves functionalities such as self-localization, mapping, navigation, obstacle avoidance, perception, object classification/identification, grasping). A TBM is a benchmark that investigates the performance of a robot system when executing a given task. TBMs are designed by focusing on the goal of the task, without constraining the means by which such goal is reached.

Combining a TBM with FBMs focused on the key functionalities required by the task provides a deeper analysis of a robot system and better supports scientific and technical progress.

3.2.4 The constraints are adapted to the objective of the evaluation

Associated objective(s): Scientific, Benchmark-based, Open

Description: As far as possible, the organisers must avoid restricting creativity and technological innovation. However, this freedom may be to the detriment of comparability. Indeed, how can we compare in detail solutions that use different functionalities to perform the same task?

The decision of constraining (or not constraining) must be taken with the utmost care by the competition organizers. In this objective, the competition organizers must consider:

- If one wants to identify the best way of achieving a task (e.g. what is the best way of weeding), the constraints on technology selection must tend to be loosened;
- If one wants to identify the best way of using a technology (e.g. the best algorithms to drive a platform), the constraints on technology will tend to be tightened.

Competition organizers should design a decision tree to find the best trade-off between creativity and comparability. The design of the decision tree can be assisted by a listing of what may be constrained in the evaluation task (autonomy, platform, source of energy, cost, time, etc.) and what cannot be constrained.

3.3 Testing environments

3.3.1 Repeatability and reproducibility of the observations are maximized

Associated objective(s): Scientific, Benchmark-based

Description: For the fairness and validity of the evaluation, competitors must be evaluated in the same conditions. However, due to the nature of their testing environment, some evaluations cannot be repeatable or reproducible. For instance, any outdoor setting will never be completely repeatable: clouds change in the sky, waves and tides modify the visibility underwater, etc. Therefore, repeatability and reproducibility of the observations, if not fully attainable, must at least always be maximized.

To achieve this in evolving environments, the competition organizer should first define all the factors that evolve and may alter the comparability of the results. In a second time, they estimate a range of validity in which they consider that the observations will reasonably be equivalent.

Several examples:

- To compensate for differences in visibility due to cloud cover, experimenters can define a range of luminosity in which system present an overall similar performance (proved formally or by testing), validated with light meters at the beginning of the test.
- To validate the efficiency of weeding, the crop line is destructed in the process, which means that several (and naturally different) lines must be used. In this context, all the test crop lines should fulfil some essential criteria to guarantee that the tests are comparable (e.g. overall density).
- Audio detection is sensitive to acoustic conditions, that may evolve according to the amount of public in the exhibition hall. In this context as well, a range of acceptable ambient noise level may be defined, and measured before each test.

One easily understand, through these examples, that there might never be favourable settings during the short time span of a competition (that lasts a few days at most). The competition organizers must then always try and find the most relevant trade-off between comparability and logistics feasibility, notably through a risk analysis of the environmental conditions that may impact the evaluation. If the risk of not being able to perform comparable observations is too high, the evaluation task must be redesigned.

3.3.2 The accessibility of the test beds is maximized

Associated objective(s): Open

Description: The competition organizers must maximize the accessibility of the test beds, which means that there will be as few limitations for participation as possible. Competition organizers must then list and consider all the limitations that may prevent either competitors to reach and use the competition site (in reasonable conditions), or to train for the competition in their own laboratory.

The following list is not exhaustive, but illustrates the categories of limitations that must be addressed:

- Logistic limitations:
  - The evaluation implies a potential danger when performing the test (e.g. manipulating dangerous chemicals) which would require some specific structure in the competitors' facilities;
  - The size or characteristics of the concerned robots constitutes an issue for travelling to the competition area (absence of adapted means of transport, accessibility of the testing area to big robots, to drones).
- Economic limitations:
  - The cost of building a reasonably equivalent test bed at the competitors' lab is prohibitive;
  - The testing procedure implies a costly destructive action that cannot thus be reproduced easily.

It is recommended that competition organizers present in their evaluation plan a description of the potential reasonable limitations that may prevent the participation to the competition, and reasonable remedial strategies that are consequently applied.

One potential remedial strategy, for example for costly testing structure, is to offer simulation tools or models. For example, if specific costly sensors are needed for the evaluation, competition organizers may suggest way of simulating their expected outputs.

We note that the cascade evaluation campaigns do not have all these limitations since data can be replicated and distributed without constraints of costs or distance (so long as intellectual property and licences are properly addressed). To ensure that these campaigns are open, it is recommended that evaluation tools and annotated datasets be made publicly available.

3.3.3 A qualification procedure is defined and implemented

Associated objective(s): Scientific

Description: The procedures applied during the evaluation must follow a strict protocol and minimize human error in order to produce reliable, repeatable and reproducible evaluation results. It is therefore essential to follow procedures that guarantee the quality of the operations carried out during the implementation of the competition. The main concern will be to ensure uniformity in the conditions under which the competition is carried out, be it test datasets, evaluation results, test environments or referees. It is the responsibility of the organiser to identify all elements for which poor management could lead to a loss of quality, fairness or impartiality.

This may consist in:

- Implementing a capacitation procedure for all the persons who perform measurement or observations on the field. Their competence must be proved in a formal way, such as a questionnaire or hands-on exercise ;
- Performing data verification through adapted procedures (cross-annotation, verification of the agreement between annotators, etc.).

All qualification activities must be documented so as to reinforce the traceability of the evaluation.

3.4 Scoring

3.4.1 Measurements and estimations are clearly identified

Associated objective(s): Scientific

Description: All competition organizers must make a good use of metrological knowledge in the design of scores. To achieve that, it is important to understand the difference between a physical measurement and an estimation. A measurement can be performed through a direct and quantitative approach, for example by using a calibrated tool. One can measure the sound level, or a distance. An estimation is an interpretation (through for example a mathematical formula), a result that is derived from several measurements and/or from qualitative observations.

To be fully objective, metrics better be based on physical measurement. However, this statement is easier to respect with the evaluation of basic functions. For example, to evaluate the ability of a mobile robot to stop one meter from a wall, a relevant and objective choice of metric is the measure of the distance between the robot and the wall. Nevertheless, when the evaluated tasks are more complex, the metric may include measurements, but also estimations. For example, the success of a natural language understanding task is assessed by the human referee who estimates, through their own understanding of languages, whether the system's output is valid or not.

An estimation is not necessarily less reliable than a measurement. However, since scores have a probative power about the performance of a system, it is important to understand and explain, when interpreting the results, what derives from a direct measurement, and what may be, for example, altered by a model that can not perfectly cover the observed phenomenon, or subject to subjectivity.

The competition organizers must thus define and explain, either in the evaluation plan or in the evaluation report, what is measured, and what is estimated. In this regard, it is expected that the estimated coverage is presented in the evaluation plan, as well as the potential uncertainties linked to the estimation.

3.4.2 Subjectivity is addressed in an appropriate way

Associated objective(s): Scientific, Benchmark-based

Description: The influence of the individual in the evaluation process can be a determining factor in the quality of the evaluation. For example, when the evaluation is based on observations performed by a referee on the field, or when a referee tests directly the robot by providing inputs (movements, speech, etc.). In this context, the reproducibility and repeatability may be fully altered in the absence of a proper procedure that guarantees the limitation of interpersonal variability.

As noted previously, referees must receive a capacitation (as defined in previous section), which will maximize the likelihood of obtaining relevant, consistent and reliable assessments. When the risk of interpersonal variations is high and cannot be reduced, it is recommended that a voting procedure between referees takes place.

When possible and scientifically relevant, simulation of the human must be used rather than resorting to a real human (for example, a sound sample of a spoken text can be sent directly to the system).

3.4.3 Metrics are properly designed

Associated objective(s): Scientific, Benchmark-based

Description: A metric is a mathematical formula that produces a score. This score represents the level of adequacy between the behaviour produced by the system and the expected ideal behaviour. This level of adequacy can represent the functional performance of the system, which means that the system performs correctly an action. This can also concern more general adequacy, such as economic, societal, or legal concerns. Whatever the purpose of the score produced, the construction of the metrics must respect certain fundamental points:

- The use of subjective parameters in the formula should be minimized as much as possible, since they may produce ambiguity and interpersonal variability. Non-avoidable subjective parameters must be dealt with as recommended in the corresponding section of this document;
- The notions of "success" and/or "error" must be strictly defined, to make sure that the chosen metrics really allows producing a relevant score;
- The metric must allow all systems to be evaluated fairly for the same evaluation task. If not, either the metric must be redefined or competition entry constraints must be set;
- The scientific literature already proposes peer-validated metrics for a vast majority of TBMs and FBMs, which may be the right inspiration.
- The evaluation of complex tasks can be broken down into simpler objective goals. The more goals achieved, the better the system.