1st International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020

In conjunction with the 24th European Conference on Artificial Intelligence - ECAI 2020

Santiago de Compostela, Spain

AI evaluation campaigns during robotics competitions: the METRICS paradigm

Guillaume Avrin and **Virginie Barbosa** and **Agnes Delaborde**[1]

Abstract. Competitions are a proven and cost-effective method to quickly develop new disruptive technologies for new markets. However, artificial intelligence and robotics competitions are usually organized independently, in silos. To get the most out of competitions, evaluation of robots should be modular (tasks are designed to assess independently different technological building blocks of the robotic systems) and open (evaluation tools and data must be publicly and freely available online). This not only meets the benchmarking needs of end users (mostly interested in performance measurements of the complete system), but also those of integrators and developers of intelligent components (camera, lidar, radar, actuators, etc.).

With this objective in mind, the H2020 METRICS project (2020-2023) organizes competitions in four application areas (Healthcare, Infrastructure Inspection and Maintenance, Agri-Food, and Agile Production) relying on both physical testing facilities (field evaluation campaign) and virtual testing facilities (data-based evaluation campaign) to mobilize, in addition to the European robotics community, the artificial intelligence one.

This article presents this approach and pave the way for a new robotics and artificial intelligence competition paradigm.

1 INTRODUCTION

Robotics competitions contribute to matching supply and demand, by encouraging innovation on the one hand, and validating new technologies on the other. The rigorous assessment procedure carried out during competitions helps reducing technical and commercial risks for manufacturers and buyers, thanks to the evaluation of robotics capacities and functionalities, and the verification of compliance with regulatory, economic, social and societal requirements.

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 [13, 16]. In parallel, and in a disjointed way, IA competitions structure the development of these systems [15, 18]. Within METRICS, partners from these projects have joined forces with organizers of other robotics competitions (Robocup, Robotex, ROSE challenge, etc.) [12, 3] and Artificial Intelligence (AI) competitions (Maurdor, Quaero, Repere, etc.) [14, 9, 11], as well as metrologists specialized in intelligent systems and experts from the Digital Innovation Hubs.

Started at the beginning of 2020, the H2020 METRICS project[2] plans to jointly address a twofold challenge:

- organize challenge-led and industry-relevant competitions in the four Priority Areas (PAs) identified by the European Commission: Healthcare, Infrastructure 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.

This article presents in Section 2 the trends and findings underlying the need for a new paradigm for the organization of robotics and AI competitions. Sections 3 and 4 present the METRICS project's response to this need, first introducing the concept of field and data-based evaluation campaigns and then outlining their interrelationship.

2 ROBOTICS COMPETITIONS TRENDS AND NEEDS

For more than twenty years, competitions have been pushing forward research and development in the robotic community:

- Fostered by the development of Information and Communication Technologies (ICT), the evolution of AI and the democratization of robotics components, competitions gather each year hundreds of participants and sponsors, and millions of spectators;
- Competitions are mainly organized by networks of expert roboticists (industrials, academics) who set the objectives of the contest and coordinate the definition of the rules and the evaluation process;
- Several competitions are explicitly dedicated to scientific and technological emulation and public dissemination of robotics, and rely on the organization of worldwide events (e.g. RoboCup), while others are dedicated to a more specialized audience of researchers and industrials in robotics (e.g. the DARPA Robotics Challenge);
- Competitions are encouraged to be more than “one-off demonstrations” [2] and adopt the form of repeatable and reproducible scientific experiments based on metrological practices so as to assess their real benefit on robotics development and control (e.g. the European Commission (EC) recommends that autonomous systems be subject to rigorous, quantified and objective scrutiny): indeed, several European

[1] Evaluation of AI Departement, Laboratoire national de métrologie et d’essais, France, email: guillaume.avrin@lne.fr

[2] https://metricsproject.eu/

projects (RoCKIn, euRathlon, RockEU2, SciRoc) pursued this approach, currently embodied by the ERL competition [1];

- Recently, reports issued by public and independent scientific committees (MAR, Delvaux report, EGE, AI HLEG, etc.) stressed the need to urgently structure and mobilize the robotics and AI community [17, 5, 8, 6];
- The reliability and validity of AI algorithms are at the heart of European concerns (2019 Council of Europe conference and reports cited above) [4, 7]. This requires that competitions address both the behavior of the robot in a physical environment, and the behavior of its AI algorithms when confronted with sets of properly qualified and controlled data. To our knowledge, the combination of robotics and AI competitions, aimed at mobilizing both communities to develop smarter and more efficient robots, has not been addressed in previous European robotics competitions.

3 FIELD AND CASCADE EVALUATION CAMPAIGNS

For each competition, METRICS is organizing both evaluation campaigns in physical environments with physical testbeds, and evaluation campaigns based on testing datasets called “cascade evaluation campaigns”. This section presents these two concepts.

3.1 Field evaluation campaign

For each competition, half of the evaluation campaigns are carried out in physical testbeds and in physical environments (control apartment, production line, agricultural plot, etc.). These evaluations (similarly to ERL competitions) include two groups of benchmarks:

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 is as independent as possible of the other functionalities of the system, so as to control it 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.

Evaluating the overall performance of a robot system while performing a task is interesting for assessing the global behavior 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, 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. The objective is to address the evaluation needs of end-users, integrators and equipment manufacturers. As AI matures, these benchmarks can also be enriched with ability-oriented evaluation [10].

3.2 Data-based cascade evaluation campaign

Robots participating in field evaluations generate data (including images or audio for diagnosis assistance, object or order recognition, etc.). These datasets are collected, annotated and qualified. To evaluate the robots participating to the field competitions, METRICS compare the data annotated by human experts (called “references”, such as annotated images or audio files) with the automatic annotations generated by robots. The evaluation process is presented Figure 1.

These annotated testing datasets are then re-used to launch cascade evaluation campaigns, targeting participants from the European AI community, either specialized in robotics or not. In the context of these campaigns, METRICS evaluators compare the references with the automatic annotations performed by the competing AI algorithms.

Evaluations in physical environments and on databases are very complementary. When evaluations in physical environments are dynamic and closed-loop, the AI can control the robot so as to maximize the performance of the detection (sensor adjustment, robot movements, etc.). On the other hand, these modifications are to the detriment of reproducibility (it is very difficult to duplicate exactly the same experiment in a physical environment, with the same environmental conditions) and the physical nature of these tests implies an increase in the cost of the evaluation (it is necessary to organize physical appointments involving technical teams and potentially large test environments). On the contrary, evaluations on databases are perfectly reproducible and at lower costs (only computers are needed, and the test dataset can be copied and distributed at will). However, these test data become obsolete as robotic technologies develop (sensors used, frequency and nature of information gathering, etc.). Moreover, they only allow evaluations without feedback loop: the test data are static and the algorithm cannot adapt the information capture to its needs.

Evaluation campaigns in physical environments therefore make it possible to produce dynamic scenarios involving feedback loops.

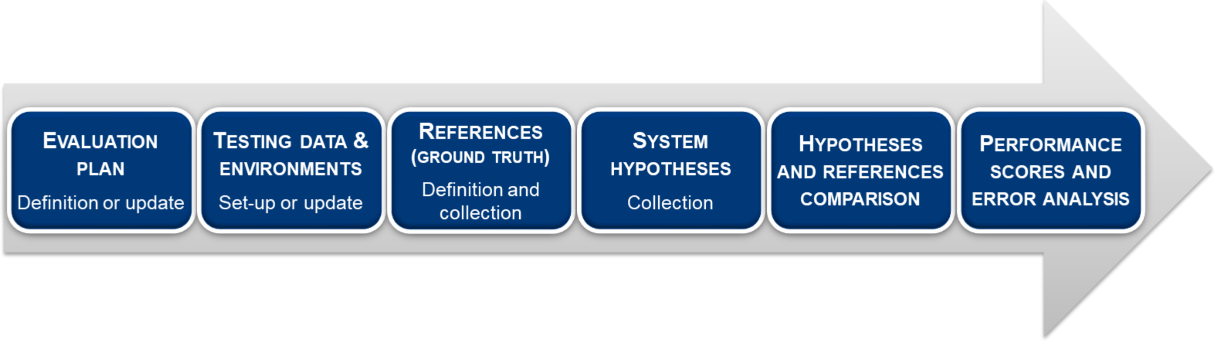
EVALUATION PLAN
Definition or update
TESTING DATA & ENVIRONMENTS
Set-up or update
REFERENCES (GROUND TRUTH)
Definition and collection
SYSTEM HYPOTHESES
Collection
HYPOTHESES AND REFERENCES COMPARISON
PERFORMANCE SCORES AND ERROR ANALYSIS

Figure 1 Steps of an evaluation campaign (a METRICS competition comprises several campaigns)

They also make it possible to regularly generate new test data that can be used during cascade evaluation campaigns aimed at evaluating AI algorithms, particularly recognition algorithms, under the best conditions of reproducibility. Furthermore, the cascade evaluation campaigns aim at maximizing the scientific and technological impact of the METRICS competitions by mobilizing the AI European community in addition to the robotics community. Contrary to testbeds competitions, where the effort required for a developer to enter the competition is very high and precludes the participation of a high number of teams, cascade evaluation campaigns have a lower barrier to entry and maximize the participation of large numbers of participants as they can be scaled up easily and at a low cost (no travel costs for robots, copies and transmission of test data almost free of charge, etc.).

Each cascade evaluation campaign is interspersed between two physical environment evaluation campaigns.

3.3 Evaluation framework

METRICS evaluation paradigm consists in comparing reference data (the “ground truth” annotated by human experts or provided by measuring instruments in the test facility) with hypothesis data (the behavior or output produced automatically by the intelligent system). This comparison allows the estimation of the performance, the reliability and other characteristics such as the efficiency of robots. The evaluation can concern the entire system (during TBM) or the main technological bricks taken independently (during FBM), as shown in Figure 2.

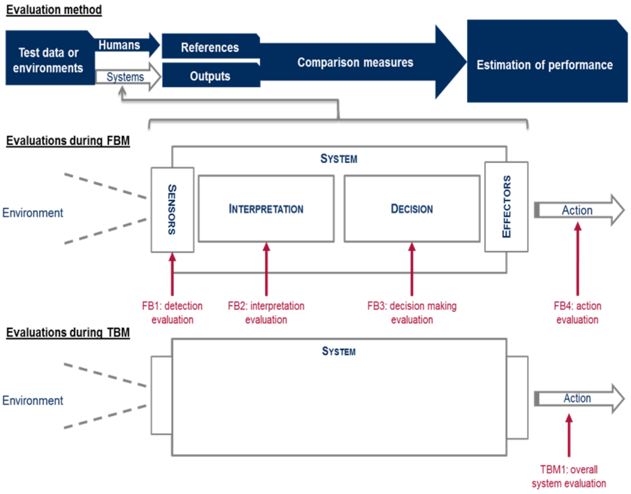

Figure 2 Evaluation method used in METRICS

4 COMPETITION ORGANISATION

During METRICS, four competitions are launched (one per PA). These competitions will last three years, with one dry-run and two official evaluation campaigns. Participants are free to take part to one campaign without participating to the others. Several call for participation will be launched by METRICS organizers during the project, broadcast broadly through community mailing lists, social media and on METRICS official website.

The repeated evaluations allow the sponsor to assess the effectiveness of the funding granted for the organization of the competition (for example to estimate the performance of potential technological solutions that address its use case). For developers, this allows them to update the technological components of the robotic system according to the quantitative results obtained.

The dry-run phase guarantees the smooth implementation of the competition: it allows the organisers to make sure that their evaluation plan is both realistic with respect to the capabilities of the systems, and fair among the different technologies used by the teams.

The dry-run campaign is followed by two official evaluation campaigns (around 12 months each), both for testbed competition and dataset (cascade) competition, aimed at objectively measuring participating robots progress in real field conditions. To this end, the evaluation plan is meant to be adapted throughout the competition so as to accompany the evolutions of the teams’ technological solutions. The steps of an evaluation campaign are presented in Figure 3.

5 CONCLUSION

Incentives for the system developers to participate in competitions are scientific or commercial recognition. Rigorous evaluation brings supply and demand together. It helps the potential users to identify the most promising solution for their needs through objective and reliable benchmarks. It allows the developers to position their products in relation to the competitors, identify the technological and commercial maturity of the product and assess the R&D effort that remains to be done before a viable product can be designed. Thus, the evaluation generates a very strong appeal in the general public, increases the visibility of the technical solutions and maximizes marketing benefits.

This is what the competition paradigm proposed by METRICS aims to provide to the AI and robotics community united in a joint initiative to advance these intelligent technologies, starting with four priority application areas for Europe.

ACKNOWLEDGEMENTS

The authors wish to thank the METRICS partners for their valuable input and collaboration.
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 871252. The contents of this document are the sole responsibility of the authors and can under no circumstances be regarded as reflecting the position of the European Union.

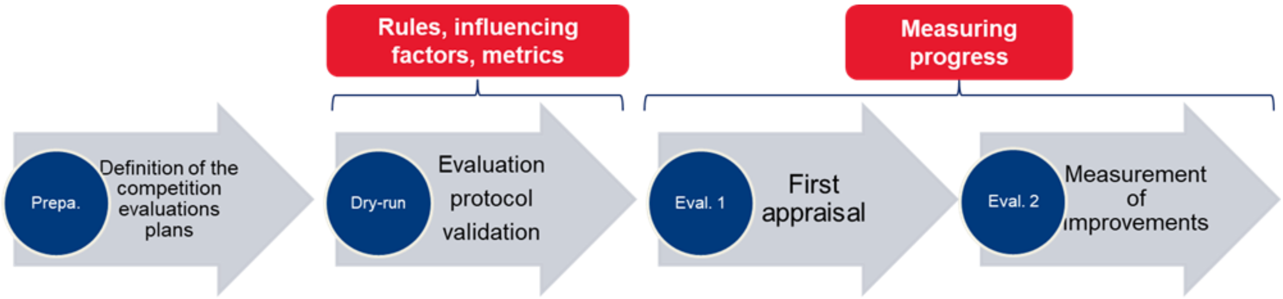
Rules, influencing factors, metrics
Measuring progress
Prepa.
Definition of the competition evaluations plans
Dry-run
Evaluation protocol validation
Eval. 1
First appraisal
Eval. 2
Measurement of improvements

Figure 3 The four steps of a METRICS competition

REFERENCES

[1] F. Amigoni, E. Bastianelli, J. Berghofer, A. Bonarini, G. Fontana, N. Hochgeschwender, L. Iocchi, G. Kraetzschmar, P. Lima, M. Matteucci, P. Miraldo, D. Nardi, V. Schiaffonati. ‘Competitions for Benchmarking: Task and Functionality Scoring Complete Performance Assessment’, *IEEE Robotics & Automation Magazine*, 53-61, (2015).

[2] M. Anderson, O.C. Jenkins and S. Osentoski, ‘Recasting robotics challenges as experiments’, *IEEE Robotics and Automation Magazine*, 10-11, (2011).

[3] G. Avrin, A. Delaborde, O. Galibert and D. Boffety, ‘Boosting agricultural scientific research and innovation’, *Proceedings of the 3rd RDV Techniques AXEMA*, SIMA, (2019).

[4] Council of Europe and Finnish Presidency of the Committee of Ministers, ‘Conclusions of the two-day conference Governing the Game Changer – Impacts of artificial intelligence development on human rights, democracy and the rule of law’, (2019).

[5] M. Delvaux, ‘REPORT with recommendations to the Commission on Civil Law Rules on Robotics’, 2015/2103(INL). PE 582.443v03-00, A8-0005/2017, (2017).

[6] European Commission’s High-level Expert Group on Artificial Intelligence, ‘Ethics Guidelines for Trustworthy AI’, (2019).

[7] European Commission, ‘White Paper on Artificial Intelligence: a Euro-pean approach to excellence and trust’, Technical report, (2020).

[8] European Group on Ethics in Science and New Technologies, ‘Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems’, European Union, ISBN 978-92-79-80329-1, (2018).

[9] O. Galibert, S. Rosset, C. Grouin, P. Zweigenbaum, and L. Quintard, ‘Structured and extended named entity evaluation in automatic speech transcriptions’, *Proceedings of 5th International Joint Conference on Natural Language Processing*, 518-526, (2011).

[10] J. Hernández-Orallo, ‘Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement’, *Artificial Intelligence Review*, 48(3), 397-447, (2017)

[11] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly, ‘A presentation of the REPERE challenge 2012’, *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, IEEE, 1-6, (2012).

[12] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and E. Osawa, ‘Robocup: The robot world cup initiative’, *Proceedings of the first international conference on Autonomous agents*, ACM, 340-347, (1997).

[13] P. Lima, D. Nardi, G. Kraetzschmar, R. Bischoff and M. Matteucci, ‘Rockin and the european robotics league: building on robocup best practices to promote robot competitions in europe’, *Robot World Cup*, Springer, 181-192, 2016.

[14] I. Oparin, J. Kahn and O. Galibert, ‘First maurdor 2013 evaluation campaign in scanned document image processing’, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 5090-5094, IEEE, (2014).

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, , S. Ma, ... and A. C. Berg, ‘Imagenet large scale visual recognition challenge’, *International journal of computer vision*, 115(3), 211-252, (2015).

[16] F.E. Schneider, D. Wildermuth and H.L. Wolf, ‘ELROB and EURATHLON: Improving search & rescue robotics through real-world robot competitions’. *Proceedings of the International Workshop on Robot Motion and Control*, IEEE, 118-123, (2015).

[17] SPARC, ‘Robotics 2020 Multi-Annual Roadmap for Robotics in Europe’, Technical Report SPARC, (2015).

[18] X. Yang, Z. Zeng, S. G. Teo, L. Wang, V. Chandrasekhar, and S. Hoi, ‘Deep learning for practical image recognition: Case study on kaggle competitions’. *In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 923-931, (2018).