

Benchmarking Intelligent (Multi-)Robot Systems

Workshop of the
19th European Conference on Artificial Intelligence (ECAI'10)
Lisbon, Portugal, August, 16th-20th, 2010.

<http://labrococo.dis.uniroma1.it/bimrs/>

The aim of this workshop is to focus the attention of researchers in Artificial Intelligence and Robotics on the need for evaluation and benchmarking methods that can provide quantitative reliable measurements of the ability of robots to perform complex tasks, that often cannot be achieved with simple reactive behaviors. A substantial basis to tackle this issue is given by robotic competitions and RoboCup in particular. While RoboCup competitions provide a first significant step and experience in benchmarking intelligent robot behaviors, we are still far from having a well-established performance evaluation methodology that can be widely accepted and generalized to other domains and tasks, and that is not restricted to robot subsystems, but rather tackles the full integrated system. Specifically, we believe that a significant contribution to this topic should be brought by the Artificial Intelligence community, both within competitions, such as RoboCup, and in other scientific endeavors.

There is a significant interest in the topic of benchmarking. IEEE RAS TC on “Performance Evaluation and Benchmarking of Robotic and Automation Systems” organizes workshops and events. The topic is also addressed within the workshop series “Performance Metrics for Intelligent Systems” PERMIS, which addresses the issue of performance evaluation in a wider perspective. However, the proposed workshop shall address the evaluation and benchmarking of intelligent robotic systems with a specific focus on the full integrated system that is designed to accomplish complex tasks, that often cannot be achieved with simple reactive behaviours/capabilities and/or involve multiple-robots. The former EURON SIG on Cooperative Robotics pointed out the relevance of the topics in a White Paper but no specific event on benchmarking has been organized. Moreover, the EURON SIG on Good Experimental Methodologies provide a web hub for people interested in benchmarking, objective performance evaluation and methodologies. This group also published a GEM SIG Review Guidelines paper and organized some events and workshops, however, the guidelines do not explicitly take into account multiple robots behavior and the related collective performance evaluation. Finally, a EURON Benchmarking WP worked on a survey of the current efforts in comparative robotics research. Members from the above cited communities are involved in the workshop organization, with the goal of creating an opportunity for the ECAI audience to meet them.

The goal of the workshop is to provide a forum for discussion of proposals and ideas on “Benchmarking Intelligent (Multi)-Robot Systems”. In particular, we aim at focusing the attention of the AI research community on the technical aspects performance evaluation of intelligent robotic systems, by presenting the state of the approaches, including (but not limited to) RoboCup competitions. Research contributions are therefore solicited on all aspects concerning the evaluation and benchmarking of intelligent robotics agents. Specifically, the focus will be on those methods and frameworks that aim at the evaluation of systems, possibly involving multiple-robots, with respect to complex tasks such as: soccer, disaster response, surveillance, home companion/helper, domestic and other environment services, etc...

Research contributions are therefore solicited on all aspects concerning the evaluation and benchmarking of intelligent robotics agents. Specifically, the focus will be on those methods and frameworks that aim at the evaluation of systems, possibly involving multiple-robots, with respect to complex tasks such as: soccer, disaster response, surveillance, home companion/helper, domestic and other environment services, etc...

Organizers

Xiao Ping Chen - Computer School, University of Science and Technology of China, Hefei, 230027, China

Adam Jacoff - Intelligent Systems Division, National Institute of Standards and Technology, Gaithersburg (MD), USA

Pedro U. Lima - Instituto de Sistemas e Robotica, Instituto Superior Tecnico, Lisboa, Portugal

Daniele Nardi - Dipartimento di Informatica e Sistemistica "Sapienza" Università di Roma, Rome, Italy

Program Committee members

Fabio Bonsignorio - Heron Robotics, Genoa, Italy

Daniele Calisi - DIS, Rome, Italy

Angel del Pobil - University Jaume I, Spain

John Hallam - Univ. Southern Denmark, Denmark

Yongmei Liu - Sun Yat-sen University, China

Raj Madhavan - NIST, USA

Elena Messina - NIST, USA

Martin Riedmiller - University of Freiburg, Germany

Peter Stone - University of Texas, USA

Satoshi Tadokoro - Tohoku University, Sendai, Japan

Mary-Ann Williams - Stanford Artificial Intelligence Lab, USA

Tijn van der Zant - RUG, Netherlands

Workshop program

The BIMRS workshop will take place on Tuesday, 17 August 2010. Here is the list of invited speakers and the program of their presentations.

Session 1	
9.00 - 9.30	Malte Helmert: Lessons learned from benchmarking in automated planning community
9.30 - 10.00	Francesco Maurelli: Benchmarking Autonomous Underwater Vehicles: the example of the SAUC-E competition
10.00 - 10.30	Daniele Calisi/Nardi: Performance evaluation of robot motion in rescue missions, service robots and autonomous vehicles
10.30 - 11.00	Coffee break

Session 2	
11.00 - 11.30	Alexander Kleiner: On Measuring and Disseminating SLAM
11.30 - 12.00	Gerhard Kraetzschmar: RoboCup@Home and RoboCup@Work: Scenarios for Benchmarking Service Robots in Residential and Industrial Settings
12.00 - 12.30	Manuela Veloso: Benchmarking Intelligent Systems through soccer games
12.30 - 14.00	Lunch

Session 3	
14.00 - 14.30	Johannes Pellenz: Benchmarking mobile robots using robot competitions
14.30 - 15.00	Fabio Bonsignorio: Experimental methodologies and performance evaluation criteria for benchmarking of intelligent (multi-)robot systems
15.30 - 16.00	Coffee break
16.00 - 17.00	All: Open Discussion

Lessons Learned from Benchmarking in the Automated Planning Community

Malte Helmert¹

1 INTRODUCTION

Automated planning is a classical area of artificial intelligence research that is concerned with the problem of planning a course of action for an agent (or set of agents) acting in a complex environment. In the frequently studied case of *classical planning*, e. g. in the propositional STRIPS formalism [1], the problem is that of finding a sequence of actions that achieves the goal of a single agent in a deterministic, static, fully observable world, given only a logical description of the initial state, goal, and preconditions and effects of actions. Automated planning is an active research area, represented at general AI conferences like IJCAI, AAAI and ECAI as well as in the annual ICAPS (*International Conference on Automated Planning and Scheduling*) conference series.

2 BENCHMARKING IN PLANNING

Up to the late nineties, most classical planning systems used their own, slightly or largely incompatible problem representations, which made it hard to compare them. This changed with the introduction of the PDDL modeling language [2], originally designed for the first International Planning Competition (IPC), which was organized by Drew McDermott as part of the AIPS 1998 conference. Between 1998 and 2008, the IPC has been repeated as a regular biennial event, which has helped establish PDDL as the problem representation used by essentially all classical planners today. Moreover, the benchmark problems used for the evaluation of planners at the IPC (which are growing with every edition of the competition, since an important aspect of the competition is to challenge planners on new, previously unseen problems) have become the “canonical” set of benchmarks for empirical evaluations of planning system. For example, all 46 papers with empirical evaluations of classical planners published in the ICAPS conference series in the period 1999–2005 use IPC benchmarks at least as part of their evaluation, and since 2003 the overwhelming majority uses IPC benchmarks *exclusively*.

3 LESSONS LEARNED

Few planning researchers would deny that the rise of a common input representation and benchmark set for classical planning has had a profound impact on the research community. There are both good and bad aspects to this, and no two researchers would completely agree on their assessment of these pros and cons and on what conclusions should be drawn from the experience. Hence, the following list of “lessons learned” should be regarded as highly subjective.

- *Common benchmarks push the state of the art.* The importance of benchmarks that allow directly comparing the approaches of different research teams cannot be overstated. Healthy competition can greatly accelerate the development of efficient algorithms.
- *Clear and common metrics are important.* In many cases, it is hard to decide which of two approaches performs better on a given problem since there are different, incomparable aspects to consider. Nevertheless, it is important to define clear, quantifiable metrics that make an unequivocal judgment. Trade-offs and design decisions in the development of such metrics should be highlighted; different metrics are needed for different purposes. Still, it is necessary to commit to a *small* set of core metrics, or else clear conclusions cannot be drawn.
- *Algorithm evaluators should not be algorithm designers.* Ideally, the evaluation of a system or algorithm should be performed by someone who has no stake in it, since it is often all too easy to come up with an experiment that makes a bad approach look good. A ubiquitous set of benchmarks with a commonly accepted evaluation methodology is one way to achieve this goal.
- *Overfitting is an issue, as is having a moving target.* No evaluation metric is perfect, and a research communities must continuously verify that it is still measuring the right thing. Algorithm designers quickly adapt their methods to optimize precisely what is measured, whether or not this captures the underlying research goal. At the same time, there are strong benefits to keeping evaluation metrics stable in order to reliably track progress over time.
- *Competitions are not scientific experiments.* Good engineering of poor ideas can often outperform bad engineering of good ideas. Understanding *why* an approach performs well is at least as important as knowing *that* it performs well. Benchmarking has its place within proper scientific methodology, but it can only be part of the story.
- *Maintaining benchmarks is work that must be incentivized.* Designing and maintaining a useful set of benchmarks is a significant amount of work that is often hard to justify because it does not produce the kind of scientific output (conference papers, journal articles, etc.) on which academics are evaluated. Research communities must find ways to actively encourage working on infrastructure for benchmarking and evaluation.

REFERENCES

- [1] Tom Bylander, ‘The computational complexity of propositional STRIPS planning’, *AIJ*, **69**(1–2), 165–204, (1994).
- [2] Drew McDermott, Malik Ghallab, Adele Howe, Craig Knoblock, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins, ‘PDDL – The Planning Domain Definition Language – Version 1.2’, Technical Report CVC TR-98-003, Yale Center for Computational Vision and Control, (1998).

¹ Albert-Ludwigs-Universität Freiburg, Germany, helmert@informatik.uni-freiburg.de

Benchmarking Autonomous Underwater Vehicles: the example of the SAUC-E competition

Francesco Maurelli and Yvan Petillot¹
Angelos Mallios and Pere Ridao²

1 Extended Abstract

One of the goals of robotic competitions is to evaluate the capabilities of different robotic systems. The rules and metrics are usually well defined and the results of different approaches can be clearly evaluated and compared. The underwater world makes no exception, with competitions organised in different parts of the world. SAUC-E stands for Student Autonomous Underwater Challenge - Europe. It is a competition for Autonomous Underwater Vehicles (AUV) which takes place in Europe on a yearly base. Held since 2006, it challenges the teams to design and build an underwater autonomous robot capable of performing realistic missions.

1.1 The tasks

During SAUC-E, each AUV has a slot of time to perform an autonomous mission. At present and in the past years, the only interaction is with the environment. No other vehicle is involved in the mission, neither the human can control or even have a feedback during the autonomous run. Wireless signal does not propagate underwater and the use of an acoustic modem is forbidden. The tasks can be divided into four main categories:

- **basic control capabilities:** this includes passing through a validation gate. No actual cognition of the gate concept is required, but only the capability of controlling depth and direction movement. A simple sequence *dive-forward-surface* is enough to demonstrate that the vehicle can be used in the underwater environment and the basic control problems have been solved. All the teams taking part in the last SAUC-E were able to demonstrate these capabilities.
- **sensor-reaction capabilities:** all the inspection tasks fall in this category. Both pipe inspection and wall inspection require sensor processing capabilities (video and/or sonar), plus an appropriate behaviour to react as a consequence of the data gathered. In the last competition, more than half of the teams were able to demonstrate at least one task in this category.
- **active interaction capabilities:** this includes the localisation and neutralisation of a mid-water target. The required behaviour for the AUV is to localise and free a buoy, cutting the rope, which links it to the bottom. None of the teams were able to cut the rope during the autonomous mission, but few of them have made an attempt.

- **advanced capabilities:** this includes the capability of locating an acoustic pinger. None of the teams was able to succeed in this task, neither to make a reasonable attempt.

1.2 Evolution of the competition

The 2010 competition has being organised for the first time by NURC, a NATO Underwater Research Centre located in La Spezia, Italy. The competition area was the NURC waterfront, a sheltered harbour with salt water and limited visibility. This location represents a significant change from previous competitions, held in indoor tanks. The requirement for the AUVs to perform in a realistic environment has risen the complexity of the competition. From year to year, the tasks have also changed progressively, in order to have more challenging tasks at each new edition. The ground target to be located has transformed into a pipe to be located and followed for a certain distance. Hitting a stationary buoy has transformed into following a moving buoy at a fixed distance and eventually into cutting the rope of a stationary buoy. Some completely new challenging tasks have been introduced. Autonomous docking has been introduced in 2009 (with only one team able to perform it) and acoustic localisation has been introduced in 2010 (however, with none of the teams able to demonstrate it).

1.3 Open points

Although a competition have clear rules, a single task can be carried on at different level of cognition. Performing a mission given a set of waypoints is different than performing a mission given a set of goals. For future competitions, it is important to understand the different levels of cognition from different teams performing the same task. For example, knowing the location of the pipe, the inspection could be performed just moving toward a specific direction, or continuously analysing the sensor data, calculating the pipe orientation with respect to the vehicle and, as a consequence, deciding where to move. Although both approaches would result in a success, the two systems are demonstrating different capabilities. Another key issue is the environment. To evaluate different systems, the operational conditions should be in theory as similar as possible. For outdoor missions, like in SAUC-E 2010, it is almost impossible to replicate the same exact conditions for all the teams (e.g. light conditions, salinity, waves). If the probability of condition change is high, this should lead to the building of robust systems. However, if it is low, although possible, a few teams, if any, would be prone to spend many resources in building a robust system.

¹ Ocean Systems Laboratory, Heriot-Watt University, EH14 4AS Edinburgh, Scotland - UK, email: f.maurelli@ieee.org

² Underwater Robotics Lab, Universitat de Girona, Escola Politcnica Superior - Campus de Montilivi - CP 17071 - Girona -Spain

MoVeME benchmarks for rescue robots, service robots and soccer robots

Daniele Calisi and Daniele Nardi¹

1 INTRODUCTION

Benchmarking and performance evaluation are key issues in every scientific field, in particular in those where different methods should be compared to each other and/or the evaluation is not straightforward or univocal.

In recent years, good experimental and evaluation methodologies have been applied in different areas of robotics. In this abstract, we focus on algorithms and methods that are directly related to the motion of wheeled robots or autonomous vehicles.

2 EVALUATING AUTONOMOUS MOTION SYSTEMS

The MoVeME framework [1, 2] is a collection of quantitative performance metrics and benchmark problems (in virtual environments). Both the metrics and the benchmarks have been designed to evaluate motion systems in critical and typical situations, without taking into account any particular need of specific scenarios and robot tasks. The rationale behind this choice was that the goal of the framework was to evaluate the performance of motion systems in its own, rather than with respect to some specific task. However, it is also true that specific tasks and scenarios raise important issues that must be addressed for the motion systems to be effectively used in real-world situations.

In this abstract, we consider some specific robot tasks and scenarios, in order to find the improvements and the extensions that can increase the effectiveness of the framework in objectively evaluating robot motion systems.

3 ISSUES RAISED BY SPECIFIC TASKS AND SCENARIOS

3.1 Rescue missions

Robot rescue missions present a number of difficulties that usually do not arise in other scenarios. In particular, real rescue missions environments present non-flat grounds and usually require 3D environment representations and navigation skills, as well as tracked, legged or aerial vehicles.

Nonetheless, difficulties for the motion system can arise even in those areas that can be explored by wheeled robots. For example, sensors can overlook low obstacles and uneven grounds, causing the motion system to fail without a known reason or to behave incorrectly. Specific situations can lead to this unwanted effects, and, if they are known and detected, should be avoided.

A suitable motion system should be robust to these issues and modify its behavior according to unexpected maneuver failures and a-priori knowledge about the environment.

3.2 Service robots and autonomous vehicles

Service robots are usually required to share the environment with other robots or human beings. This results in scenarios that are much

more dynamic, with respect to rescue missions. The issues that arise in this case are usually different from those mentioned above, and include: dynamic obstacles can require frequent trajectory re-planning and/or reactive behaviors; “predictable” movements are often preferred, when interacting or sharing the same space with humans; if the motion system controls an autonomous vehicle, additional issues about passengers comfort can arise.

3.3 Soccer robots

Path-planning and control in high-dynamic scenarios such as robot soccer matches involve also the possibility to evaluate the motion performance of the robots *as a team*, rather than individually. In order to accomplish some specific tasks, multiple goal configurations can be requested to be reached; in this case, the environment contains two kinds of dynamic obstacles: robots of the opponent team, whose behavior is mostly unpredictable and should be addressed using techniques similar to those applied for service robots, and robots from the same team, whose behavior can be the result of a distributed policy that is developed in order to minimize the possible conflicts between robots of the same team.

4 PROPOSED SOLUTIONS

In order to address the issues presented above, we propose to improve the MoVeME framework, both increasing the number of benchmark problems and improving the performance metrics. In particular:

- we need to address, from the evaluation point of view, the general problem of “configurability”: from the one hand, it is preferable that the motion system were adjustable in such a way to adapt to specific constraints (e.g., forbidden or preferred maneuvers, maximum speeds, curvature radius, etc.), from the other hand, difficult parameter tuning (especially when parameters do not have straightforward meanings or are not directly measurable) should be considered a disadvantage;
- we can generalize the “*Risk*” performance metric (that is currently related to the distance of the robot to obstacles), in such a way that it can also give a measure about the proximity to generic forbidden configurations, states or control commands;
- the set of experiments in the MoVeME framework should be extended with dynamic obstacles, unexpected behaviors and failures, in order to assess the robustness of the system to unforeseeable events; moreover, multi-robot scenarios (and multi-robot goals) should be considered.

REFERENCES

- [1] D. Calisi, L. Iocchi, and D. Nardi. A unified benchmark framework for autonomous Mobile robots and Vehicles Motion Algorithms (MoVeMA benchmarks). RSS Workshop on Experimental Methodology and Benchmarking, Zurich, Switzerland, June 2008.
- [2] Daniele Calisi and Daniele Nardi, ‘Performance evaluation of pure-motion tasks for mobile robots with respect to world models’, *Autonomous Robots*, 27(4), 465–481, (September 2009).

¹ Department of Computer and System Sciences, “Sapienza” University of Rome, email: <firstname>. <lastname>@dis.uniroma1.it

On Measuring and Disseminating SLAM

Alexander Kleiner ¹

Abstract. The problem of mapping an unknown environment with uncertain sensor measurements is often referred to as the simultaneous localization and mapping (SLAM) problem. SLAM is considered to be a complex problem because to localize itself a robot needs a consistent map and for acquiring the map the robot requires a good estimate of its location. This mutual dependency among the pose and the map estimates makes the SLAM problem hard and requires searching for a solution in a high-dimensional space.

Whereas dozens of different techniques to tackle the SLAM problem have been presented, there is no gold standard for comparing the results of different SLAM algorithms. Currently, the evaluation of robotic maps is based on qualitative analysis (i.e. visual inspection). This procedure turns out to be suboptimal, particularly when applied to large-scale maps. Furthermore, it does not allow for better understanding of what errors specific systems are prone to and what systems meet the needs.

Although a rich amount of SLAM techniques have been presented in the past, they are almost never found on commercial robot platforms used by first responders in the field.

In this talk I report results from recent international efforts on measuring the accuracy of SLAM. These include results from the standard test methods for response robots developed by the ASTM International Committee on Homeland Security Applications (E54.08.01), the benchmarking and scoring of teams during competitions, such as RoboCup Rescue, and several workshops and special issues dedicated to this issue. Furthermore, I introduce a simple mapping system for supporting first responder teleoperation that has been demonstrated in the response robot exercise in Disaster City and is currently integrated into a commercial robot platform used by bomb squads.

¹ University of Freiburg, Germany, email: kleiner@informatik.uni-freiburg.de

RoboCup@Home and RoboCup@Work: Scenarios for Benchmarking Service Robots in Residential and Industrial Settings

Gerhard Kraetzschmar¹

The large amount of research performed in robotics in the past 20 years has resulted in an awesome collection of algorithms, methods, architectures, and systems designs, which make it difficult for developers of new systems to choose the right approach. In a 2008 workshop talk, I coined the expression “1001 architectures for 1001 robots” for this problem. In the absence of widely accepted consensus on what the right approach is, benchmarking could provide some help in making the right decisions. However, due to its intrinsic interaction with the environment, benchmarking in robotics has proven to be quite difficult. While good benchmarks exist for algorithms and methods in some specific areas like vision and SLAM, benchmarking component assemblies, systems, and applications is still a largely open problem. RoboCup@Home is a RoboCup competition that targets residential robotics applications, while RoboCup@Work is a recently proposed new RoboCup competition that transfers some of the RoboCup@Home ideas to industrial settings. The talk will present the structure as well as some details on both competitions and outline their potential to serve as testbeds for scientific benchmarks of both specific components and complete systems.

¹ Bonn-Rhine-Sieg University, Germany, email: gerhard.kraetzschmar@h-brs.de

The Challenges of Evaluating Multi-Robot Systems: A Perspective from the RoboCup Experience

Manuela Veloso¹

For more than 10 years, we have been pursuing multi-robot competitions and symposia within the RoboCup initiative. We have learned a lot about evaluating research by trying to compare solutions in common benchmark problems. In this talk, I will overview the different RoboCup leagues, present how they were initially introduced, and explain their scientific goals. I will detail the challenges with the evaluation and testing of approaches in a competition scenario, and discuss the remarkable advances that we have experienced enabled by the competitions. I will conclude with a proposed set of guidelines to create successful benchmarks.

¹ Carnegie Mellon University, USA, email: veloso@cs.cmu.edu

Benchmarking mobile robots using robot competitions

Johannes Pellenz¹

Newly proposed solutions in robotics are often difficult to compare with other approaches already known in the community: The results presented in publications might work only in the own lab environment of the research group and maybe only under certain circumstances. For the specific task of robotic mapping, repositories containing standard datasets with 2D² and 3D³ laser scan data give the chance to compare the results on the same data. Further more, some SLAM algorithms are available as open source⁴. However, prerecorded data does not help to evaluate planning strategies or navigation solutions. Even more difficult is the evaluation of new sensor designs – here the generated data by itself is the output of the research effort.

Scientific robot competitions – such as the RoboCup – provide a great opportunity to demonstrate new approaches and to benchmark their strength in direct comparison to other systems.

Mainly known for the soccer competitions, RoboCup provides also competitions in non-standardized environments: In the RoboCup@Home scenario, the robots have to solve problems in typical household environments. The RoboCup Rescue competition is even more challenging: It also provides an unknown area but also features a non-flat ground. Here, the challenge is to navigate through an arena with different levels of difficulty (such as uneven underground, ramps and stairs) with a remote controlled or an autonomous robot. During the main competition, the robot or team of robots can score points by finding survivors and by collecting sensor data that describes the state of the simulated victims.

Due to the always changing environment (with a design that is beyond the influence of the participating teams), the robots can show their ability to work in unknown environments. The winner of the whole competition must show a strong combined skill in mobility, navigation, map building, victim detection and system stability.

However, groups that research on specific topics only (which is common for University teams) can get rewarded by focus on one of the "Best in Class" competitions in RoboCup Rescue: Best in class trophies are given for the most mobile robot (ignoring the aspects of autonomy and victim identification), the best autonomous robot (ignoring its mobility performance) and the robot with the best manipulation skills. Since this year, a separate (inter-league) mapping challenge is organized, where the sensor data of a particular robot or measuring system is provided to all participating teams, and therefore only the mapping software determines the result of that challenge.

In addition to the actual competition, the RoboCup Symposium provides a forum to present the results of the ongoing research.

We have found that the solutions that were used by the successful teams (in the main competition, as well as in the best in class challenges) were adopted quickly by the other teams in the same or the following year. Therefore, the RoboCup competitions help not only to evaluate good solutions and to benchmark the performance of the robots, but also to move forward the research community as a whole.

¹ University of Koblenz-Landau, Germany, email: johannes.pellenz@uni-koblenz.de

² see <http://radish.sourceforge.net/>

³ see <http://kos.informatik.uni-osnabrueck.de/3Dscans/>

⁴ see <http://openslam.org/>

Experimental methodologies and performance evaluation criteria for benchmarking of intelligent (multi-)robot systems

Fabio P. Bonsignorio¹

As the complexity and variety of required tasks and of the environments targeted by robotics research grows, it becomes necessary to develop well defined and rigorously founded procedures and methods that allow quantitative comparison of the solutions provided by the research activities.

In order to facilitate exchanges of methods and solutions between different research groups and to be able to assess of the state of the art with the required objectivity.

In order to address these needs the European Robotics Network of Excellence (EURON) has funded a Special Interest Group on Good Experimental Methodology and Benchmarking. The European Commission sponsored the EURON network (EURON I and EURON II) till the end of April 2008. Since then, EURON 3 continues on the momentum that it gained during almost a decade of networking. All the activities of Euron 3, including this SIG, continue without public funding and on voluntary basis.

This group has organized more than ten successful workshop on the topic of benchmarking intelligent and robot systems and the related topic of Refutable Robotics Research at such conferences as IROS, RSS and recently ICRA with significant attendance. This talk will focus on principles, methods, and applications expanding beyond the current limits of robotics applications and/or embodied AI systems in terms of cognitive capabilities and autonomy.

These issues are discussed with reference to a research and reporting context, where the “robotics experiments” themselves are difficult to “replicate”. It will be explained what is meant as ‘Refutable Robotics Research’. A survey of existing activities and developing methods for benchmarking and to cope with the related experimental replication issues will be given.

We are facing several, not only, methodological but also platform and epistemological issues, there is a growing awareness that replication of reported results and their comparison by means of benchmarks are needed to foster a quicker cumulative advancement of our knowledge of intelligent physical agents and even to correctly appreciate disruptive innovation in robotics and ‘intelligent systems’ research. Is New AI better than GOF AI? Should we take inspiration from biology and medicine? Are replication and benchmarking really possible in robotics? Are current limits in reporting a bottleneck to industrial exploitation? Is there an impact on robotics research funding policy evaluation?

¹ Heron Robots s.r.l., email: fabio.bonsignorio@gmail.com