Toward Replicable and Measurable Robotics Research

By Fabio Bonsignorio and Angel P. del Pobil

The famous experiment by Galileo—one of the founders of modern science—in Pisa's Cathedral in 1582, was one of the very first examples of a scientific experiment validating a scientific result: the discovery of the pendulum law. Galileo measured the variations of the oscillation period of a lamp in the dome by his own heart rate. From those times, experiment replication and experiment replication and reproducibility of results are at the cornerstone of the scientific method. Yet in robotics, artificial intelligence, and automation, the reproduction of results from conference and journal papers, as they are today, is quite often very difficult, if not impossible. This situation is bad for science, as it becomes difficult to objectively evaluate the state of the art in a given field, and also it becomes problematic to build on other people's work, thus undermining one of the basic foundations of scientific progress.

Moreover, it is detrimental to the industrial exploitation of results, for which we need to compare the effectiveness and efficiency of different methods proposed to solve the same technical or scientific problem, for example, from the computational and energy-consumption standpoints. This difficulty in reproducing results, however, makes this comparison usually very cumbersome and without trustable outcomes. This situation hampers and slows down the industry take-up of research results, and there are many more than those already exploited that are likely to benefit our daily lives.

The community has been aware of this issue for a long time. In 2007, we, with John Hallam, created the European Robotics Research Network (EURON) Good Experimental Methodology (GEM) and Benchmarking Special Interest Group (SIG) within the EURON Network of Excellence (NoE), a NoE is a networking-oriented European-funded project. In 2006, one of us, Angel P. del Pobil, organized a workshop on benchmarking at the IEEE/Robotics Society of Japan International Conference on Intelligent Robots and Systems (IROS) in Beijing, China, as an activity of the EURON NoE work package devoted to benchmarking, and the first website on survey and inventory of current efforts in comparative robotics research was established (http://www.robot.uji.es/EURON/en/index.htm).

The GEM guidelines [1] were one of the major outputs of the SIG's early activities. Although, initially, the guidelines were focused on more careful reviews, mainly thanks to one of us (Fabio Bonsignorio), the real problem became clear: the core issue is the reproducibility/replicability of experimental results. In 2009, at the International Conference on Robotics and Automation (ICRA) in Kobe, Japan, the IEEE Robotics and Automation Society (RAS) Technical Committee on Performance Evaluation and Benchmarking of Robotic and Automation Systems was founded, with similar objectives. In parallel, the Performance Metrics for Intelligent Systems conference series focuses on performance measurement challenges arising from the application of robotics and automation technologies to practical problems in the commercial, industrial, homeland security, and military domains. More information can be found at http://www.nist.gov/el/isd/permis2012.cfm.

There has been a long series of workshops at various conferences, such as IROS, ICRA, and the Robotics Science and Systems Conference, in which more than 200 people have participated so far. We mostly organized them with Elena Messina and John Hallam, but there have also been some organized by others, and there have been a number of competitions and publications aiming at finding a way out of a situation that is considered by many as unsatisfying (see http://www.ieee-ras.org/performance-evaluation and http://www.heronrobots.com/EuronGEMSig/).

When EURON joined euRobotics Association Internationale Sans But Lucratif (AISBL), the private part of the European Public–Private Partnership on Robotics, the activities of the former EURON GEM SIG became part of the Topic Group on Evaluation and Assessment of Research Results, also known as Benchmarking and Competitions. A solid example of benchmarking methodology is proposed in "Benchmarking in Manipulation Research" by Berk Calli, Aaron Walsman, Arjun Singh, Siddahrta Srinivas, Peter Abbel, and Aaron Dollar. There are experimental setting where this approach is difficult to implement. In those cases competitions,

Digital Object Identifier 10.1109/MRA.2015.2452073
Date of publication: 11 September 2015

or, better, scenario-based evaluation procedures, have been recognized as a component of the recipe for the benchmarking of results, particularly when intelligent behaviors are involved. The extent to which competitions can be regarded as scientific experiments, and which ones, is still a matter of discussion. An article in this issue, "Competitions for Benchmarking," by Francesco Amigoni, Emanuele Bastianelli, Jakob Berghofer, Andrea Bonarini, Giulio Fontana, Nico Hochgeschwender, Luca Iocchi, Gerhard K. Kraetzschmar, Pedro Lima, Matteo Matteucci, Pedro Miraldo, Daniele Nardi, and Viola Schiaffonati, may provide some hints.

Methodological, Practical, and Epistemological Issues

Although the number of robotics papers published in journals and conferences is constantly growing, the possibility of reproducing results is left to the good will of some authors. The number and nature of envisioned applications and proposed methods are vast and also steadily increasing. As a consequence, some members of the community believe that the comparison of results would not be practically possible. A remarkably varied set of robotic applications is approached by a significantly disparate set of methods, sometimes based on notably different principles, with different hardware (HW)/software (SW) architectures in different environments. On the one hand, the explosive growth of research results shows that the community is becoming larger and increasingly active; on the other hand, it raises some serious problems when you have to objectively evaluate the actual relevance of the results and the actual state of the art in a given field.

As previously stated, the difficulty of reproducing results—let alone comparing different methods and solutions—slows down the industrial take-up of new solutions. Basic research is also hindered, since it is very difficult for a research group to build on the results of another one, leading to a very limited cross-exploitation of results between different groups, and a general prevalence of exploration over exploitation. Many new solutions are proposed, but the community often does not go deep into the analysis of most of them.

The EURON GEM guidelines [1] are essentially an adaptation to the robotics and automation domain of the general methodology applied in science and engineering that was pioneered by Galileo and Boyle. Today, as discussed in [2], only a limited subset of published results follow those methods and usually not completely. Of course, not every paper should follow a rigorous experimental protocol: position papers, concept papers describing upcoming research, papers concerning algorithms, or survey papers do not need to comply with a rigorous and epistemologically sound experimental methodology. Still, many papers that claim to have solved a problem (say, autonomous driving) based on simulations or field experiments should comply. Robotics, artificial intelligence, and automation are not pure mathematics. The proposed solutions need to be able to work in the set of environments and for the set of tasks for which they have been studied. There are scientific aspects in robotics, for example, related to the unbundling of the brain-body nexus in humans and animals, but even when we are closer to pure engineering applications, experimental proofs of the effectiveness of the proposed solutions are needed.

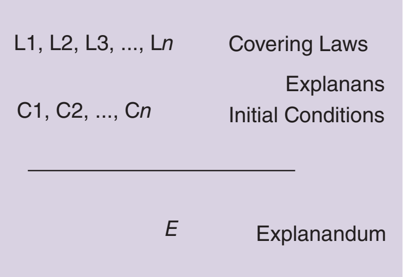

Figure 1. The Hempel–Oppenheim model of scientific knowledge. In the conceptual schema represented in this figure, which summarizes the Hempel–Oppenheim model of scientific knowledge, all the logical enunciates have a probabilistic truth value. We need a precise and complete list of laws invoked for the explanation, a precise and complete list of initial conditions (system HW/SW architectures, environments, tasks), a precise definition of what is explained or proved. In addition, we must accept the fact that our theoretical claims, enunciate, have to be of probabilistic nature, since we operate in open-ended stochastic environments. (Figure adapted from G. Boniolo, "A Contextualized Approach to Biological Explanation," *Philosophy,* vol. 80, pp. 219–247, 2005).

We should at least be able to

- validate the results by replicating them
- compare the results in terms of the chosen performance criteria.

This holds true for both purely scientific issues and real-world applications. The fact that robotics research deals with very diversified problems should not be seen as a serious obstacle. Indeed, medicine and life science, for instance, where the complexity and variety of the studied objects are not smaller than in robotics, have developed rigorous experimental protocols. We should take inspiration from them. An epistemological model of biological science was proposed by Hempel and Oppenheim; see Figure 1.

> **We should not be so surprised by the fact that we are struggling to define valid and shared benchmarking procedures for intelligent robots.**

We can expect that having replicable and measurable results will affect the content of the results, not just their reporting. We should not be so surprised by the fact that we are struggling to define valid and shared benchmarking procedures for intelligent robots. Their development uncovers a lot of practical, publishing, and also epistemological issues. A more detailed discussion of this topic can be found in [7] and will be the main topic of a future publication. Besides the so-far unsatisfactory, in this respect, experimental and reporting practice, an important reason could be the limited scientific understanding of intelligence and cognition in natural and artificial systems. The practical issues span from modeling, to statistical significance assessment, to the mechatronic design and construction of specific test equipment, and to the actual replication procedures, the experimental protocols, and the necessity to provide the data, time series, and HW/SW description. The epistemological issues, with respect to paradigm examples of science, like physics, are many: multilevel causality, the large number of preconditions and laws involved, and probabilistic relations between causes and effects. Robotics has many problems in common with biology and medicine. We have comparison and evaluation criteria for cars and many other machines and appliances; we are just starting to develop those for robotics and intelligent systems. The articles in this issue show that it is possible, and we already have some promising proposals. After several years of discussions and attempts presented in a long series of workshops and elsewhere, a new kind of replicable paper in robotics has become mature.

> **This situation is bad for science, as it becomes difficult to objectively evaluate the state of the art in a given field.**

State of the Art?

There has been a growing awareness about these issues in the community. Yet, it is still very difficult to find examples of replicable papers in robotics and automation. It is now possible to attach supplemental materials to papers in the most important journals of the field. Increasingly, authors share data sets and code, in particular, in the simultaneous localization and mapping (SLAM) community, and shared data sets and libraries, like Peter Corke's MATLAB libraries, are made available. But despite the progress in defining replication protocols, we are, in this respect, at the very beginning. Years ago, Amigoni et al. [8] showed that not a single paper among the top cited ones in SLAM and navigation met all the basic criteria listed in the GEM guidelines. We may have clearly improved since then, but probably not enough.

Competitions have also matured in the direction of becoming experiments on the most elusive intelligent behaviors. You will not find the real state of the art here, either in this editorial or in this issue, as far as replication of results is concerned. The reason is straightforward: this issue is the first example of a publication including a list of replicable research results. To a certain extent, the state of the art coincides with this issue of *IEEE Robotics and Automation Magazine* (*RAM*).

It is also interesting to note that, in more established areas of research with more mature experimental methodology, like clinical research, there have recently been serious concerns about the replicability of published research and the consequent negative impact on research and even new drug development and health care [3]–[5]. The idea that the publishing process should evolve is widespread. Published research reporting should provide enough information to allow the replication of the results. The web, and the easier distribution of information that the web makes possible, might be part of the solution.

On the one hand, this new possibility was identified several decades ago [6]. On the other hand, the practice of sharing research is already evolving, as shown by the success of preprint e-publishing platforms like arxiv (www.arxiv.org) or some recent experiments of open review on the web (see http://openreview.informatik.uni-freiburg.de) as well as by this special issue.

Contribution of this Issue

After many discussions and attempts, a new kind of paper seems to be necessary. This new kind of paper should include the following:

- *description*—a journal paper with text, figures, and multimedia, according to GEM guidelines (or similar)
- *data sets*—similar to the option provided by various journals and magazines, included this one
- *code identifiers*—complete code identifiers and/or downloadable code (executable files may be enough)
- *HW identifiers*—HW description or HW identifier (if it is identifiable).

This special issue of *RAM* is the very first example of a collection of replicable robotics reports covering a remarkably wide area of diverse robotics subfields. The articles in this issue provide a living example of the viability of replicable research in robotics. They span a wide and diverse set of areas of research in robotics, thus countering the idea that this field is too diversified to allow a rigorous shared methodology.

The articles report replicable experiments, benchmarking methods, and a couple of exemplary surveys on competitions ("Humanoid Robots in Soccer," by Reinhard Gerndt, Daniel Seifert, Jacky Baltes, Soroush Sadeghnejad, and Sven Behnke) and on the new and important field of soft robotics ("Deformation in Soft-Matter Robotics," by Liyu Wang and Fumiya Iida). We have a very interesting article about how competitions can be given a rigorous scientific meaning in the ("Competitions for Benchmarking," by Francesco Amigoni, Emanuele Bastianelli, Jakob Berghofer, Andrea Bonarini, Giulio Fontana, Nico Hochgeschwender, Luca Iocchi, Gerhard K. Kraetzschmar, Pedro Lima, Matteo Matteucci, Pedro Miraldo, Daniele Nardi, and Viola Schiaffonati). The set of replicable research examples covers wearable systems ("Wearable Inertial Sensors," Barbara Bruno, Fulvio Mastrogiovanni, and Antonio Sgorbissa) and manipulation ("Benchmarking in Manipulation Research," by Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar). We have three papers on different aspects of marine robotics ("Tracking Divers," by Nikola Mišković, Đula Nađ, and Ivor Rendulić; "Exploring 3-D Reconstruction Techniques," by Javier Pérez, Jorge Sales, Antonio Peñalver, David Fornas, José Javier Fernández, Juan Carlos García Sánchez, Pedro J. Sanz, Raúl Marín, and Mario Prats; and "Testing the Waters," by Andrea Sorbara, Andrea Ranieri, Eleonora Saggini, Enrica Zereik, Marco Bibuli, Gabriele Bruzzone, Eva Riccomagno, and Massimo Caccia). And then we cover motion planning ("Benchmarking Motion Planning Algorithms," by Mark Moll, Ioan A. Şucan, and Lydia E. Kavraki), bipedal locomotion ("Benchmarking Bipedal Locomotion," by Diego Torricelli, Jose Gonzalez, Jan Veneman, Katja Mombaur, Nikos Tsagarakis, Antonio J. Del-Ama, Angel Gil-Agudo, Juan C. Moreno, and Jose L. Pons), and last but not least the requirements for replicable simulation experiments ("RoboCup Simulation Leagues," by David M. Budden, Peter Wang, Oliver Obst, and Mikhail Prokopenko).

You should read the articles from various standpoints: the novelty of the content, the significance and viability of the proposed benchmarks, the approaches that the authors have chosen to allow the replication of their results. The first question to ask is: are these results reproducible? You will notice that the articles have different focuses and that the approaches are different. What are the strengths and weaknesses of the various approaches? Some authors, like Perez et al., seem more focused on the definition of the benchmarking criteria, others, like Moll et al. or Sorbara et al. on the replicability of the benchmarks. Some, like Bruno et al., rely on third-party repositories like github or sourceforge, some have designed a dedicated website. Some use an XML-based description, some do not. Are the experiment statistics always managed in the best way? How should the statistical significance of the experiments be evaluated and the related metrics reproduced? Have a look at Figure 1 in Moll et al.; to be able to replicate the experiments, we will need to structure systems like that. What is the best way to implement them? You may wish to compare with Figure 2 in Sorbara et al. (for example). This collection of very interesting articles inspires a long list of thought-provoking questions and provides many possible solutions and insights.

Of course, this is just a starting point. Hopefully, the practical replication of the results by the community will show the best ways to provide information to make the results of robotics papers reproducible.

Road Ahead

We will need to foster the proper attitudes toward replication of results in the community. We should not think that scientific publishing could not further evolve. Replicable papers can be a valuable addition to the current scientific publishing landscape. In this new context, the initial severe peer-review preceding the publication of papers will be just a prerequisite for the real peer-review based on the active reproduction of the published results by the community at large. This will also make easier the understanding of the still open scientific problems related to intelligent, animal-like, and cognitive behaviors.

Another thing to consider for the future is that the authors of the articles in this issue, while usually providing the information necessary through their own websites, also had to upload the data needed for replication to the magazine website as attachments to this article. We think that, in the future, we will need a more structured approach; in this sense, the website structures for this issue will also contribute to the definition of a new publishing set of conventions to present replicable papers, not as just attachments. This is what is available now, and it is useful to have a single self-contained entry for all the articles in the special issue.

We would like to see the results of many of the articles here reproduced as they are in other articles commenting on these issues and suggesting improvements. Although many challenges are still ahead, we believe we are heading in the right direction: back to the basics of the scientific method.

References

[1] F. Bonsignorio, J. Hallam, and A. P. del Pobil, Eds. (2008). GEM Guidelines. Euron GEM Sig Report. [Online]. Available: http://www.heronrobots.com/EuronGEMSig/

[2] F. Bonsignorio, A. P. del Pobil, and E. Messina, "Fostering progress in perfomance evaluation and benchmarking of robotic and automation systems [TC spotlight]," *IEEE Robot. Automat. Mag.*, vol. 21, no. 1, pp. 22–25, 2015.

[3] Challenges in Irreproducible Research. Nature Special. [Online]. Available: http://www.nature.com/nature/focus/reproducibility/

[4] *How Science Goes Wrong: Scientific Research has Changed the World.* Now it needs to change itself, The Economist, 2013.

[5] *Trouble at the Lab: Scientists Like to Think of Science as Self-Correcting.* To an alarming degree, it is not, The Economist, 2013.

[6] J. Claerbout, "Electronic documents give reproducible research a new meaning," in *Proc. 62nd Annu. Int. Meeting: Society Exploration Geophysics*, 1992, vol. 92, pp. 601–604.

[7] F. Bonsignorio, J. Hallam, and A. P. del Pobil, "Defining the requisites of a replicable robotics experiment," in *Proc. RSS Workshop Good Experimental Methodologies Robotics*, 2009.

[8] F. Amigoni, M. Reggiani, and V. Schiaffonati, "An insightful comparison between experiments in mobile robotics and in science," *Auton. Robot.*, vol. 27, pp. 313–325, 2009.

IMAGE LICENSED BY INGRAM PUBLISHING