

The Euron GEM Guidelines

Fabio P. Bonsignorio

Heron Robots & DIMEC Unige, Italy

John Hallam

Mærsk Mc-Kinney Møller Institute, South Denmark University, Denmark

Angel P. del Pobil

Robotic Intelligence Lab, Universitat Jaume I, Spain

- 'Look Ma, No Hands' syndrome?
- Replication of experiments
- Performance measure benchmarks to allow results comparison
- Needed to foster research advancement and enable practical application of research achievements



Look Ma,
no hands!



As the complexity of developed robotic and intelligent systems grows, it is more and more needed to define proper experimental approaches and benchmarking procedures.

Trustable benchmarks are needed in order to allow the comparison of the many research results in service robotics research and enable their industrial application.

Both replication and benchmarking are needed to foster a cumulative advancement of our knowledge of intelligent physical agents and even to correctly appreciate disruptive innovation in the science ((1) ???) and technology of robots.

Should we take inspiration from biology and medicine (2) ???

If robotics aims to be serious science, serious attention must be paid to experimental method.

What is an 'experiment' in robotics?

The research activities in the robotic fields are huge and it is huge the number of published papers.
In order to allow the exploitation of the many results obtained it is at least necessary to able to:

- validate the results by replicating them
- compare the results in term of the choosen performance criteria

Although some work is already carried on, a lot of open issues are still in front of us.

Replication&Falsification

As it is known K.Popper defined in a very tight way the requisites for a discipline to be considered 'scientific'. In social science, management and economics exact repetition is often seen as a limit case, experiments that systematically vary one or more input parameters of a system under study to see whether its output parameters remain stable or change according to the expected model in a predictable way.

Replication&Falsification

Only when the model fails clearly in a number of varied experimental setups it is considered 'not replicable'.
Nevertheless, as already noticed, all disciplines aiming to be considered 'scientific' incorporate a concept of experiment replication and a concept of 'falsification' of theory through experiments.

Replication&Falsification

A **clinical trial protocol** is the detailed written plan of a clinical experiment.

It may be inspiring looking at the US NCI guidelines for drafting a clinical trial protocol: the emphasis on signaling 'adverse events' , the definition of 'criteria for response assessment', the necessity of defining clearly principal and secondary hypotheses to be validated.

The **statistical section** of the protocol is asked to define how the data will be analyzed in relation to each of the objectives.

In particular it expects that an acceptable trial specify, with reference to the study objectives:

- * Method of randomization and stratification

Total sample size justified for adequate testing of primary and secondary hypotheses

- * Error levels (alpha and beta)
- * Differences to be detected for comparative studies
- * Size of the confidence interval of the estimates.

Replication&Falsification

It seems clear that in robotics the experimental methodology standards are currently in many cases weaker, and the syndrome 'it worked once, in my lab' could be more widespread than we may think.

Robotics papers come in many varieties. For example, a paper may present a new theoretical advance; it may describe a new system concept; it may advance an argument based on discussion; it may present comparisons between a set of known techniques; it may do more than one of the foregoing...

1. Is it an experimental paper?

An experimental paper is one for which results, discussion and/or conclusions depend crucially on experimental work. It uses experimental methods to answer a significant engineering or scientific question about a robotic (or robotics-related) system. To test whether a paper is experimental, consider whether the paper would be acceptable without the experimental work: if the answer is no, the paper is experimental in the context of this discussion.

2. Are the system assumptions/hypotheses clear?

The assumptions or hypotheses necessary to the function of the system must be clearly stated. System limits must be identified.

3. Are the performance criteria spelled out explicitly?

An experimental paper should address an interesting engineering (or scientific) question. Such questions will generally concern the relationship between system or environment parameters and system performance metrics. The performance criteria being studied must be clearly and explicitly motivated, and the parameters or factors on which they depend must be identified.

4. What is being measured and how?

The performance criteria being studied must be measurable; the paper must identify measurements corresponding to each criterion and motivate the choice of measurements employed. The data types of measurements should be clearly given or obvious — categorical (e.g. yes/no), ordinal (e.g. rankings), or numerical.

5. Do the methods and measurements match the criteria?

Measurement methods and choices must be clearly and explicitly described and, where appropriate, explained and justified. The paper must demonstrate (unless it is self-evident) that the chosen measurements actually measure the desired criteria and that the chosen measurement procedures generate correct data (for example, that implementations are plausibly correct).

6. Is there enough information to reproduce the work?

It is fundamental to scientific experimentation that someone else can in principle repeat the work. The paper must contain a complete description of all methods and parameter settings, or point clearly to an accessible copy of that information (which should be supplied to the paper's reviewers). Known standard methods need not be described, but any variations in their application must be noted. If benchmark procedures are used, they must be referenced, and any variations from the standard benchmark must be documented and justified.

7. Do the results obtained give a fair and realistic picture of the system being studied?

Care must be taken to ensure that experiments are properly executed: factors affecting measured performance that are not the subject of study must be identified and controlled for. In particular, uncontrolled variations in the system or the environment must be identified and dealt with by elimination, grouping techniques or appropriate statistical methods. The task tackled by the system must neither be too easy or too hard for the system being studied. Outlying measurement data may not be eliminated from analysis without justification and discussion.

8. Are the drawn conclusions precise and valid?

The experimental conclusions must be consistent with the experimental question(s) the paper poses, the criteria employed and the results obtained. System limits must be presented or discussed as well as conditions of successful operation. Conclusions should be stated precisely. Those drawn from statistical analysis must be consistent with the statistical information presented with the results.

Conclusions and future work

There is a widespread perception of the need of improving experimental practices in robotics, among many others world wide initiatives, the Euron SIG GEM(II and III!) is trying to address these needs.

It is thought that proper and widely accepted replication procedures and performance benchmarks are needed to allow the cumulative progress of robotic science and technologies and even to assess the value of new disruptive ideas.



check www.heronrobots.com/EuronGEMSig
for more...and to download the full guidelines
(beta!)