Robotics and Autonomous Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Toward generalization of experimental results for autonomous robots

Francesco Amigoni *, Matteo Luperto, Viola Schiaffonati

Artificial Intelligence and Robotics Laboratory, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

ARTICLE INFO

Article history:
Received 26 August 2016
Accepted 31 August 2016
Available online xxxx

Keywords:
Autonomous robots
Experimental methodology
Generalization of experimental results
Representativeness of experimental settings

ABSTRACT

In this paper we discuss some issues in the experimental evaluation of intelligent autonomous systems, focusing on systems, like autonomous robots, operating in physical environments. We argue that one of the weaknesses of current experimental practices is the low degree of *generalization* of experimental results, meaning that knowing the performance a robot system obtains in a test setting does not provide much information about the performance the same system could achieve in other settings. We claim that one of the main obstacles to achieve generalization of experimental results in autonomous robotics is the low degree of *representativeness* of the selected experimental settings. We survey and discuss the degree of representativeness of experimental settings used in a significant sample of current research and we propose some strategies to overcome the emerging limitations.

1. Motivation

The call for a more rigorous experimental methodology currently plays a strategic role in the practical assessment of computing and represents a way for reflecting on its disciplinary status in between science and technology. This call puts attention on several questions: from the dispute on the name (should computing be called a science or not?) to the investigation of the sciences of the artificial, including the debate on whether and how traditional experimental principles (like control, comparison, repeatability, reproducibility, and generalization) could be applied to computing. However, few studies have systematically discussed the different ways the concept of experiment has been intended and employed in practice (including, for example [1]).

Here, we consider Artificial Intelligence and Autonomous Robotics as subfields of computing. How experiments are conceptualized and discussed in the case of intelligent autonomous systems represent no exception with respect to the above picture. In this paper, we focus on intelligent autonomous systems operating in physical environments, namely on *autonomous robots*, but many of our considerations and results hold also for other kinds of intelligent autonomous systems. Some solutions for developing reliable experimental methodologies have emerged in the practice of autonomous robotics, such as the use of data sets (like Radish [2] and Rawseeds [3]), the development of reliable and partially validated simulation tools (consider for example USARSim [4] and Gazebo [5]), and the development of benchmarks and scientific competitions (see, for example, RoCKIn [6]). However, systematic analyses on how the notion of experiment is used in the field and how good experimental practices could be developed, promoted, and adopted are still quite rare [7]. Besides methodological complications in adopting rigorous experimental protocols when dealing with artifacts such as intelligent autonomous systems (the most important one perhaps being the lack of independence of the experimenter, since usually the experimenter coincides with the designer), the experimental assessment of intelligent autonomous systems presents other difficulties, given the fact that these systems are often made to operate in physical environments and to interact with the real world. The focus on replicable and measurable robotics research, that has recently gained momentum in a part of this community, represents an important attempt to overcome some methodological, epistemological, and practical issues that slow down the industrial take-up of new solutions [8].

Among the limiting factors, we focus on the following ones that, although not exhaustive of the difficulties arising, offer a view on some of the issues often encountered in the experimental evaluation of autonomous robots.

- The interaction of robots with physical environments is largely unpredictable. Even if the possible data coming from sensors are finite (for example, a simple digital camera with 8 bits 720×720 pixels can return $8^{720 \times 720}$ possible images), their brute force enumeration and naïve analysis is far beyond the power of any conceivable computational device. Similar considerations hold for actuators.
- "Natural" metrics and protocols for evaluating the performance of autonomous robots are largely missing. For example, consider the difficulties in assessing the quality of

* Corresponding author.
E-mail addresses: francesco.amigoni@polimi.it (F. Amigoni), matteo.luperto@polimi.it (M. Luperto), viola.schiaffonati@polimi.it (V. Schiaffonati).

http://dx.doi.org/10.1016/j.robot.2016.08.016
0921-8890/

interactions between humans and robots [9]. Moreover, very often there exists no reference performance (such as ground truth) against which the performance of autonomous robots can be compared.
- The physical settings in which autonomous robots are tested are limited in number and nature and sometimes are not actually representative of other real environments. For example, a university building in which a service robot is developed and tested could be radically different from the residential apartments in which the same robot is intended to work.

In our opinion, these limiting factors (among others) negatively impact on the experimental evaluation of autonomous robots, both on a concrete level and on a more abstract one.

From a concrete point of view, when the aim is to evaluate the feasibility and the properties of proposed methods or systems, in particular against other alternative methods or systems, the above limitations imply that several resources (especially time) have to be spent to obtain in many cases only partial and weak results. Even if rigorous experimental protocols for evaluation of results have traditionally seldom played a major role in autonomous robotics, it is important to recognize their centrality and to work in order to increase their adoption in the field [7].

From a more abstract point of view, we can see the difficulties outlined above as a limit in achieving what are considered to be two of the main tenets of the experimental method: comparison and generalization [10]. *Comparison* concerns the capability not only of knowing what has been already done in the past, but also of comparing the new results with the old ones. This, for instance, requires full documentation and a sincerity principle in reporting, together with positive results, anomalies and negative results that can reveal important information. *Generalization* is the capability to interpret experimental results within a framework wider than the specific one developed for the experiment, in order to support more general outcomes. This requires not only to collect data, but also to interpret and explain them in order to derive the correct implications, given that interpretations and explanations are not easy to achieve and may not provide clear-cut results. It is important to point out that generalization is strongly related to two other principles which are traditionally considered at the core of experimental methodology: *reproducibility* and *repeatability*. Reproducibility is the possibility for different experimenters to achieve the same results by starting from the same initial conditions, using the same instruments and parameters, and adopting the same experimental techniques. Repeatability concerns the fact that a proper experiment must be the outcome of a number of trials to guarantee that experimental results are systematic and have not been achieved by chance. In particular, statistically significant patterns that can be identified over repeated trials of the same experiment play an important role for generalization.

These limitations in achieving comparison and generalization hinder both scientific development and industrial exploitation of the research results. In the first case, the reason is that large but unsystematic efforts are devoted to implement methods and to experimentally test them. In the second case, the main reason is that the proposed methods and systems cannot be easily and fairly ranked according to their features, and their applicability in settings different from those in which they have been developed is not guaranteed.

In the effort of improving the quality of experimental activities in autonomous robotics, some attempts have been made to take inspirations from how experiments are performed in traditional sciences, such as physics and biology, and to translate in the practice of autonomous robotics the general experimental principles of science [11]. However, dealing with artifacts, robotics shows a strong engineering component and cannot be fully assimilated to traditional scientific fields, where experiments are generally conducted for hypothesis testing purposes and with a strong theoretical background. In robotics, instead, experiments have mainly the goal of demonstrating that a given artifact is working with respect to a reference model (e.g., its expected behavior) and, possibly, that it works better than other similar artifacts. However, at the same time, the most advanced robot systems are extremely complex, and their behavior is hardly predictable, even by their own designers, especially when considering their interaction with the natural world. In this sense, experiments in autonomous robotics have also the goal of understanding how these artifacts work and interact with the world and, therefore, are somehow similar to experiments in the natural sciences.

With this paper, we would like to contribute to the discussion on experiments in autonomous robotics from the perspective of generalization, namely whether and how experimental results obtained in a specific case can be extended to other situations. We claim that one of the obstacles to achieve generalization of experimental results in autonomous robotics is the low degree of *representativeness* of the selected experimental settings. After an attempt to define representativeness and its importance for generalization, we survey the experimental settings adopted in the SLAM-related papers presented at ICRA 2014, showing that representativeness is currently not always considered as an important issue. We especially discuss the representativeness issues from the perspective of data sets, simulations, and competitions, which are some of the main means used to evaluate autonomous robots. Moreover, we propose some concrete strategies for improving the representativeness of the settings used in experiments within autonomous robotics.

The main original contributions of this paper, whose nature is more methodological than technical, are the recognition of the problem of generalization of experimental results and, in particular, of the representativeness of experimental settings, the identification of some current trends about representativeness, and the proposal of some strategies to improve representativeness.

2. Representativeness of experimental settings

In this section, we provide a tentative definition of representativeness.

Generally, experimental activities are conducted in settings (environments and configurations), which should be precisely specified. Experimental results will convey valuable information about the general performance of robot systems if these settings are enough similar to those in which the systems can possibly operate. The ideal goal in order to extend experimental results to other settings is to evaluate a robot in settings that are as much representative as possible, since the degree of generalization of experimental results depends on the degree of representativeness of the settings in which the results are obtained.

We could say that a particular experimental setting is as *representative* as much as its *features* are close to those of the class of settings where robots can operate. A feature is a distinctive characteristic of an environment, such as the presence of a loop of corridors in an indoor environment. The identification of the features of experimental settings and of the metrics to measure their similarity, which can be used to precisely define representativeness, is a largely open issue that depends on the specific areas of autonomous robotics.

Here, we do not attempt to provide any concrete proposal, but we list a number of (non exhaustive) requirements that enable the definition of such features and metrics and the evaluation, at least informal, of the representativeness of experimental settings.

Please cite this article in press as: F. Amigoni, et al., Toward generalization of experimental results for autonomous robots, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2016.08.016