Can we Reproduce it? Toward the Implementation of good Experimental Methodology in Interdisciplinary Robotics Research

Florian Lier[1], Phillip Lücking[1], Joshua de Leeuw[2], Sven Wachsmuth[1], Selma Šabanović[3] and Robert Goldstone[4]

***Abstract*— The insufficient level of reproducibility of published experimental results has been identified as a *core issue* in the field of robotics in recent years. Why is that? First of all, robotics focuses on the abstract concept of computation and the *creation* of technological artifacts, i.e., software that implements these concepts. Hence, before actually reproducing an experiment, the subject of investigation must be *artificially created*, which is non-trivial given the inherent complexity [5]. Second, robotics experiments usually include expensive and often customized hardware setups (robots), that are difficult to operate for non-experts. Finally, there is no agreed upon set of methods in order to setup, execute, or (re-)conduct an experiment.**

To this end, we introduce an interdisciplinary and geographically distributed collaboration project that aims at implementing good experimental methodology in interdisciplinary robotics research with respect to: a) reproducibility of required technical artifacts, b) explicit and comprehensible experiment design, c) repeatable/reproducible experiment execution, and d) reproducible evaluation of obtained experiment data. The ultimate goal of this collaboration is to reproduce the *same* experiment in two different laboratories using the same systematic approach which is presented in this work.

I. INTRODUCTION

Reproducibility has been identified as a core issue in robotics [6]. Individual scientists have been working on this often neglected topic for a long time. Already in 2007, Fabio Bonsignorio, Angel P. del Pobil and John Hallam initiated the Good Experimental Methodology (GEM) and Benchmarking Special Interest Group (SIG) within the EURON Network of Excellence. Their GEM guidelines [1] were one of the major contributions of this special interest group with respect to reproducibility. Only recently, Bonsignorio et al. published a special issue of the IEEE Robotics & Automation Magazine dedicated to reproducibility in Robotics. In their article [2], Bonsignorio et al. demand a new kind of paper regarding reproducibility: A journal paper with text, figures, and multimedia, according to GEM or similar guidelines, data sets, complete code identifiers and/or downloadable code, and lastly, hardware descriptions or identifiers. This demand is perfectly aligned with the goals (cf. abstract [a-d]) of our work. In this paper, we will briefly describe our collaboration scenario, the issues we are tackling, and introduce our **systematic approach** to foster reproducibility of robotics experiments.

[1]CITEC (EXC 277) at Bielefeld University, Germany flier,plueckin,swachsmu@techfak.uni-bielefeld.de

[2]Cognitive Science Department, Vassar College, USA jdeleeuw@vassar.edu

[3]School of Informatics and Computing, Indiana University, USA selmas@indiana.edu

[4]Dep. of Psychological and Brain Sciences, Indiana University, USA rgoldsto@indiana.edu

Reproducible Research in Robotics: Current Status and Road Ahead
ICRA 2017, Singapore

II. THE SCENARIO AND INHERENT ISSUES

Our scenario is typical for modern interdisciplinary Robotics research. Scientists from Indiana University, with a scientific background in HRI and psychology, and researchers from Bielefeld University, with expertise in software engineering and robotics, are interested in cooperatively conducting HRI experiments using the same Robotic platform. This cooperation faces **the same issues** that can be observed in the broader community with respect to reproducibility [7] [8]. Our ultimate goal is to conduct the same experiment at Bielefeld and Indiana to cross-validate results and to learn about the requirements to accomplish full *practical* reproducibility. Another ambition is to conduct experiments independently, i.e., without sending project members around the globe in order to oversee experiment setup, execution, and evaluation.

In our project we are facing the following issues: i) How can we technically reproduce the utilized system in both laboratories using an identical soft- and hardware execution environment and exactly the same parameters, e.g., sensor frequencies, middleware settings, etc? ii) How can we ensure, that the experiment protocol is consistently adhered to at both sites? This includes the actual *real world interaction* of the test subjects with the system, as well as the operation (starting & runtime verification) of the soft- and hardware stack by non-experts iii) How can we ensure that the independently collected data is consistently evaluated, e.g., with the same evaluation method? iv) Lastly, how can we also enable other researchers to reproduce our experiment without additional labor, such as collecting and "packaging" all necessary artifacts retrospectively?

III. A SYSTEMATIC APPROACH TO FOSTER REPRODUCIBILITY

In order to solve issues i, iii, & iv we base our experiment upon a software tool chain that has been designed to foster reproducibility of software intensive experiments in robotics. The tool chain is called **Cognitive Interaction**

Toolkit (CITK). In the following, we will briefly outline its capabilities; more technical details are explained in [3]. Issue ii will be covered by **jsPsych** [4], which has been developed for behavioral sciences like psychology.

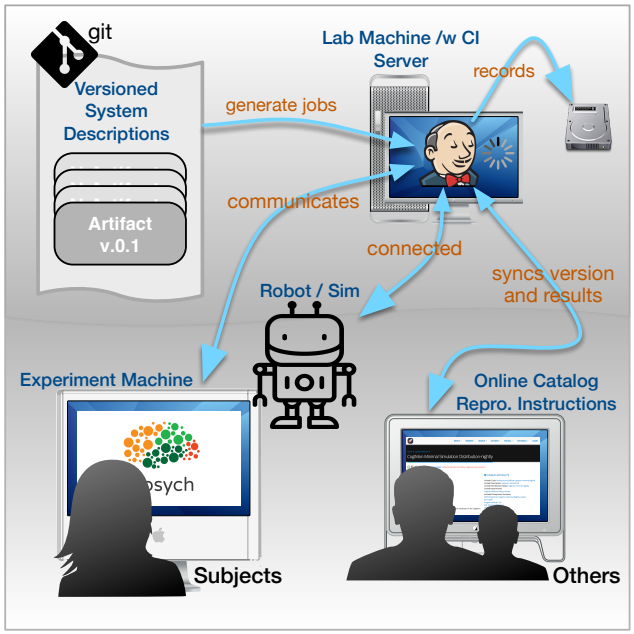

Fig. 1. CITK and jsPsych setup for reproducible Robotics experiments

Figure 1 depicts our systematic approach. The CITK provides a template-based description framework in order to define the technical aspects of a robotics system. There are two types of descriptions. The first type is called a *recipe* [9]. It describes required system artifacts: software components, downloadable data sets, or system configuration files. Templates for new types of artifacts can be added on-the-fly. The second type is called a *distribution* [10]. A distribution is a composition of N recipes and hence specifies an entire system. Distributions, as well as recipes, always reference **versions**, e.g., tags, branches, or commit hashes of an artifact. A distribution reflects a versioned instance of a system. Recipes and distributions are publicly available in our GIT repository [10]. The CITK also provides a pre-packaged (download and run it, no configuration required) Continuous Integration Server [11] [12]. The CI Server is capable of compiling, deploying, and running entire software systems. In order install and run a system, the CITK implements a generator-based approach. A so-called job-configurator tool automatically creates all required build-jobs (for every recipe!) on the CI Server. A CITK user only selects the desired distribution file. It is also possible to connect a physical robot to the machine that runs the CI Server in order to control/actuate it. Lastly, the CITK provides a framework to automatically start, stop, and introspect a robotics software system [13]. Running a system merely means triggering a special build-job on the CI Server. Data that is acquired/logged during each system run is stored (and timestamped) on the CI Server. By utilizing this part of our structured approach we can ensure technical reproducibility of all required artifacts and also repeatable experiment execution (re-trigger the corresponding build-job) regarding the software side of an experiment. A CITK showcase video can be watched here: `https://vimeo.com/205541757`

With respect to the "human side", i.e., the actual experiment design and orchestration plus the adherence of its protocol, we make use of a framework called jsPsych. jsPsych is a JavaScript library for creating behavioral experiments in a web browser. To use jsPsych, a researcher provides a description of the experiment structure in the form of a timeline. It handles which trial to run next and storing the obtained data. jsPsych uses plugins to define what to do at each point on the timeline. We extended the functionality of jsPsych in order to a) trigger an experiment/system run on the CI Server and b) execute experiment-specific behaviors on the robot, e.g, based on the current state of the timeline in jsPsych. In our experiment, the test subject will be able to see the robot while **jsPsych guides him or her** through the experiment. Thus, we can also resolve issue ii by implementing a structured & repeatable experiment timeline that is coupled and synchronized with our robot ecosystem. Furthermore, since jsPsych has been written for psychologists and our "Robotics-CI" extension/plugin is transparent for end-users, experiments for robotics can be designed by non-technical staff. No additional expertise is required. After each experiment trial the next step in our approach is the evaluation and analysis of the obtained data. Since the robot-centric data is stored on the CI Sever, e.g., actuator configuration at any given time, and also the human-centric data (sent by jsPsych), the merging can be done automatically. This is, again, achieved by creating recipes for data analysis and the generation of a corresponding build-job as explained earlier. Here, psychologists may provide scripts written in R or Python that implement *established statistical methods*. Thus, consistent, transparent and comprehensible evaluation of results can be traced back for each trial.

Lastly, the CITK tool chain features an online catalog (`https://toolkit.cit-ec.uni-bielefeld.de/`) for scientific experiments in robotics. This catalog implements a human readable and browsable representation of recipes, distributions, experiments and collected data sets. The catalog can be automatically updated by a running CITK CI Server instance. Moreover, the catalog provides detailed instructions about how to use the CITK in order to **technically reproduce** any system that has been uploaded to the catalog. The catalog also serves as a landing page for "reproducible papers" in order to provide the additional information as demanded by Bonsignorio et al. and other scientists working on reproducibility in computational sciences [14] [15] [16]. This part of our approach will tackle issue iv: how can we also enable other researchers to reproduce our experiment? By using the CITK, in combination with jsPsych, all required information is at hand in a versioned, traceable, well-structured and

"ready to deploy" representation — even during the development phase since build-jobs can be updated on-the-fly. In the upcoming weeks we will determine the final research hypothesis including a pilot experiment. The experiment will be deployed and executed at IU and BU independently using the CITK. We will report on the lessons learned and upload our experiment to the catalog in order to also enable other researchers to reproduce it. We are planning to integrate data sets and methods that have been developed in the benchmarking community, e.g, in [17] [18].

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 and A. P. del Pobil, "Toward Replicable and Measurable Robotics Research [From the Guest Editors]," in IEEE Robotics & Automation Magazine, vol. 22, no. 3, pp. 32-35, Sept. 2015.

[3] Lier F, Hanheide M, Natale L, et al. Towards Automated System and Experiment Reproduction in Robotics. In: Burgard W, ed. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

[4] de Leeuw, Joshua R. "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser." Behavior Research Methods 47.1 (2015): 1-12.

[5] Brugali, Davide, and Azamat Shakhimardanov. "Component-based robotic engineering (part ii)." IEEE Robotics & Automation Magazine 17.1 (2010): 100-112.

[6] Osentoski, Sarah, et al. "Brown ROS Package: Reproducibility for Shared Experimentation and Learning from Demonstration." Enabling Intelligence through Middleware. 2010.

[7] Krishnamurthi, Shriram, and Jan Vitek. "The real software crisis: Repeatability as a core value." Communications of the ACM 58.3 (2015): 34-36.

[8] Osentoski, Sarah, et al. "Robots as web services: Reproducible experimentation and application development using rosjs." Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011.

[9] Lier, Florian. "Cognitive Interaction Toolkit Recipe File (JSON)" https://goo.gl/NJOHNC Cognitive Interaction Toolkit - Research for Cognitive Interaction, 13 Nov. 2013. Web. 24 Feb. 2017.

[10] Lier, Florian. "Cognitive Interaction Toolkit Distribution File (JSON)" hhttps://goo.gl/UlBKdi Cognitive Interaction Toolkit - Research for Cognitive Interaction, 18 Nov. 2016. Web. 24 Feb. 2017.

[11] Fowler, Martin, and Matthew Foemmel. "Continuous integration." Thought-Works) http://www. thoughtworks. com/Continuous Integration. pdf (2006): 122.

[12] Duvall, Paul M. Continuous integration. Pearson Education India, 2007.

[13] Lier, F., Ltkebohle, I., & Wachsmuth, S. (2014). Towards Automated Execution and Evaluation of Simulated Prototype HRI Experiments. HRI '14 Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, 230-231. doi:10.1145/2559636.2559841

[14] Amigoni, Francesco, Monica Reggiani, and Viola Schiaffonati. "An insightful comparison between experiments in mobile robotics and in science." Autonomous Robots 27.4 (2009): 313.

[15] Mesirov, Jill P. "Accessible reproducible research." Science 327.5964 (2010): 415-416.

[16] Howison, James, and Julia Bullard. "Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature." Journal of the Association for Information Science and Technology (2015).

[17] F. Bonsignorio, A. P. D. P. Del Pobil and E. Messina, "Fostering Progress in Performance Evaluation and Benchmarking of Robotic and Automation Systems [TC Spotlight]," in IEEE Robotics & Automation Magazine, vol. 21, no. 1, pp. 22-25, March 2014. doi: 10.1109/MRA.2014.2298363

[18] Ceriani, Simone, et al. "Rawseeds ground truth collection systems for indoor self-localization and mapping." Autonomous Robots 27.4 (2009): 353.

ACKNOWLEDGMENT

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

This research has been conducted in the context of the Thematic Network Interactive Intelligent Systems, which is supported by the German Academic Exchange Service (DAAD) and sponsored by the German Federal Ministry of Education and Research (BMBF).

We would like to thank Martin Wiechmann for his work on the technical realization, as well as various software integration tasks.