Report of the 8th Workshop on Empirical Requirements Engineering (EmpiRE 2023)

Vincenzo Gervasi
University of Pisa
Pisa, Italy
vincenzo.gervasi@unipi.it

Alessandro Marchetto
University of Trento
Trento, Italy
alessandro.marchetto@unitn.it

Maya Daneva
University of Twente
Enschede, the Netherlands
m.daneva@utwente.nl

ABSTRACT
The Eighth International Workshop on Empirical Requirements Engineering (EmpiRE 2023), co-located with the 31st IEEE International Requirements Engineering conference (RE 2023), was held on September 5, 2023 in Hannover, Germany. This report presents the workshop structure, the keynote speech, the themes of the presented papers, and the panel discussion.

Keywords
Empirical research methods, Requirements Engineering, Evidence-based Software Engineering, Artificial Intelligence

1. INTRODUCTION
Requirements Engineering (RE) has become a well-established discipline where a wide range of approaches, techniques, and tools have been proposed. Systematic attempts to evaluate and compare usefulness, effectiveness, and usability of such proposals resulted in a growing attention to methods for empirical assessment. Empirical Software Engineering (ESE) aims at applying the empirical research methodologies to the software engineering field. In other terms, it aims at studying and proposing qualitative and quantitative methods to collect and analyze evidence that helps evaluate software engineering approaches, techniques and tools. Design science, action research, case studies and experiments, hence, become indispensable and valuable ways to check proposals with respect to reality, thus allowing researchers and practitioners alike to understand the actual value, cost, and benefits of any proposed technique in a particular context. The long-term objective of the Workshop series on Empirical Requirements Engineering (EmpiRE) is to increase the cross-fertilization of ESE methods and RE by actively encouraging the exchange of ideas to understand why and how the empirical methods from ESE (and potentially other disciplines) can help to assess and improve existing or novel approaches in RE. EmpiRE 2023 builds on the success of seven previous editions, which were held in conjunction with RE'18 (Banff), RE'17 (Lisbon), RE'15 (Ottawa), RE'14 (Karlskrona), RE'13 (Rio de Janeiro), RE'12 (Chicago), RE'11 (Trento). EmpiRE grew upon the community's efforts and results of the CERE (Comparative Evaluation in Requirements Engineering) workshop series, which were held in conjunction with RE in 2003, 2004, 2006, 2007. With a strong emphasis on practical evaluation, but more comprehensive goals, EmpiRE is considered CERE's spiritual successor — thus marking the 20th anniversary since the inception of the series. The workshop has always been characterized by its focus on personal interaction and hands-on work, in addition to scientific presentations. Hence, the consequences of the COVID pandemic were particularly felt, with a three year hiatus that Empire 2023 is now closing. The goal of the 8th edition was thus to shape the next phase of cross-fertilization of RE and ESE, specifically: (i) to open up the interdisciplinary debate on the steadily moving frontiers in empirical RE, and (ii) to extend the network of RE and ESE researchers designing and conducting empirical studies in RE, which in turn will lead to the cross-fertilization between RE and ESE.

2. WORKSHOP THEMES
The topics in the EmpiRE 2023 Call for Papers included all those concerns related to the application of empirical research methods in RE. In particular: design science and action research as methods for doing research with practitioners and for practitioners; emerging research methods, e.g. leveraging data-centric intelligent systems; qualitative studies; approaches to evaluate validity of results of RE research. While we welcomed empirical research papers in all domains, we paid special attention to the research challenges in RE for Artificial Intelligence (AI), Machine and Deep Learning, Recommender Systems, and Natural Language Processing, following recent significant advancements in those areas in particular.

3. WORKSHOP PROGRAM
The EmpiRE 2023 workshop took place as a full-day workshop on the 7th of September 2023 at RE 2023, in Hannover, Germany. The workshop program included a keynote talk, two sessions paper presentations, and a panel discussion:

- Sjaak Brinkkemper (keynote): *Empirical Requirements Engineering for Smart-Life Applications: Experiences with Automated Medical Reporting in the Care2Report System*
- Rifat Ara Shams, Muneera Bano, Didar Zowghi, Qinghua Lu and Jon Whittle: *Exploring Human Values in AI Systems: Empirical Analysis of Amazon Alexa.*
- Jianwei Shi, Oliver Karras, Martin Obaidi and Malvin Tandun: *Can Videos as a By-Product of GUI Testing Help Developers Understand GUI Tests?*
- Gøran H. Strønstad, Ilias Gerostathopoulos and Emitza Guzman: *What's next in my backlog? Time series analysis of user reviews.*
- Jens Gulden and Alexander Rachmann: *The Square of Values for Modeling Human Values in Requirements Engineering.*
- Dan Berry, Sjaak Brinkkemper, Xavier Franch, Smita Ghaisas, and Oliver Karras (panel): *Standardizing validation methods: empirical evaluation of Large Language Model (LLM) based research.*

We invite readers to review the EmpiRE 2023 web site for further information at https://sites.google.com/unitn.it/empire2023/program and to read the full text of the workshop contributions in the Workshop Proceedings of the EmpiRE 2023 Conference [1].

3.1 The EmpiRE 2023 Keynote Talk
Sjaak Brinkkemper's keynote presentation brought a variety of perspectives on the question of how to do RE for smart life applications and how to collaborate with companies in this area. The talk reported researchers' experiences made in the team of the speaker in the design

and implementation of technology that reduces the administrative burden in routine healthcare processes (e.g. physiotherapy). The keynote used as its context the Care2Report research program (https://www.care2report.nl) whose goal is "to automate medical reporting processes based on multimodal (audio, video, bluetooth) recording of a consultation, followed by knowledge representation, ontological conversation interpretation, and finally the generation and uploading of the report in the electronic medical record system". In his keynote, Sjaak Brinkkemper presented the RE challenges the team encountered and the solution practices that were found to work in the specific research context of physiotherapists. The talk suggested that in those areas where innovative solutions are developed and RE for innovative designs needs to be done, it is hard to set apriori some specific evaluation goals and make a clear-cut selection of validation techniques that are expected to suit the context. It is also impractical at this point of time to initiate RE-community-wide standardization efforts (in terms of validation techniques), as the area of smart life application development – and in particular, the development of applications relying on generative AI technology – is so new. In turn, very little is known regarding what empirical validation techniques are the most promising and in what contexts. Instead, researchers should stay open to experimentation and be agile in order to do their best in the research context they confront. Researchers, therefore, need to adopt an 'entrepreneurial mindset' characterized by curiosity, creativity, risk-taking, and a passion for learning and growth.

Finally, the keynote ended with some practical advice on what empirical researchers should watch for if they want their collaborations with companies to be fruitful. First and most important is to establish long term trust. Second, the importance of carefully choosing the problem to be researched. This also means to assure the research questions are relevant for the industry partners and the research is beneficial for the company's bottom line.

3.2 Discussion on the Presentations

Rifat Ara Shams and colleagues [2] reported results from an exploratory study on human value requirements from the end-users' feedback for an AI system. These authors chose the Amazon Alexa app as their context and examined 1003 users' reviews to identify relevant human values and assess the extent to which these values are addressed or ignored in the app. The study revealed that out of 34 values of the end-users of Amazon Alexa, only one value is explicitly addressed while 23 are mostly ignored in the Alexa app. The user feedback collected via app reviews analysis provided mixed experiences (both addressed and ignored) on the 10 other values in the set of 34.

Jianwei Shi and colleagues [3] reported results of a comparative evaluation of two approaches to requirements-based testing of software systems with graphical user interfaces (GUIs). The authors wanted to evaluate the effectiveness and efficiency of these approaches in defect analysis, from the perspective of developers of GUI-based software systems as used in everyday life. The authors compared the use of screenshots and text against the use of videos as a by-product of GUI testing, integrating annotations and test outputs directly into the videos. The empirical evaluation findings indicated visible differences between the video and the screenshots in effectiveness and efficiency in defect analysis, but could not prove that the differences were statistically significant. The comparative evaluation concluded that both forms of multimodal documentation complement each other and that both types of documents are helpful.

Strønstad and colleagues [4] addressed the challenge of analyzing huge volumes of app reviews and presented a method that helps app analysts and requirements specialists to identify reviews that are worth looking at, and that inform the product backlogs with issues to fix or new requirements to consider. These authors' approach identifies such reviews by automatically detecting anomalies (unusual peaks) in time series of user reviews. The approach includes an automatic processing pipeline that ingests user reviews, aggregates them, and produces reports of which aggregates may contain valuable information for software evolution. The very first empirical evaluation suggested the approach is applicable and feasible in real-world contexts.

Finally, the work of Gulden and Rachman [5] posits that the discipline of software system design has to take into account how social values are reflected and incorporated in system design. In line with this, designers and analysts should be able to apply methodical means to consciously reason about the psychological and social values that are embedded implicitly or explicitly in software. The authors presented a proposal of a modeling approach (called the Square of Values) for visually explicating value constellations in ethical system design. The applicability of this proposal was evaluated by means of an experiment with students.

In the four discussed papers, a common theme emerged clearly, namely the importance of considering the human element in empirical RE research. All four papers investigate the role of humans, of their beliefs and values, of their reactions, of their potential for providing valuable input to analysts and developers. While this was certainly not planned for and not even expected, we believe it to be a distinctive feature and a good characterization of where the research interests of the EmpiRE community are heading to.

Furthermore, from a research-methodological perspective, another thread emerged: the first is about the use of theories. If human values are an integral part of empirical research then the use of theories from fields such as social sciences and psychology seems sensible, at least to inform empirical RE research interventions. The RE community made the observation a long time ago that other disciplines use theories [6], while, in RE, hypotheses often come from practice. As Dan Berry put it: "If it works, then I will use it, even if I have NO idea why it works!" This poses the question whether we as empirical researchers should do more or less of this. We considered this an important point for inclusion in upcoming editions of the EmpiRE workshop.

3.3 Panel on Standardizing Validation Methods: Empirical Evaluation of LLM-based Research

The five panelists of EmpiRE 2023 took turns to present their stands according to the alphabetical order of each one's last name, while the first author served as moderator.

Dan Berry presented RE for Artificial Intelligence (AI) as a "hairy task" [7] as we want AI to mimic humans while trying to avoid AI-caused mistakes that have far-reaching resonance. Sjaak Brinkkemper put forward that if we want an acceptable evaluation of a LLM-based system, then we need more task-specific metrics. For example, we need more medical domain specific metrics to judge acceptability of LLM-based systems, not just recall and precision. Xavier Franch emphasized the importance of the business case of LLM-based systems and the evaluation of these systems' level of accuracy and efficiency in practical contexts. Smita Ghaisas focused on the experience of Tata Consultancy Services in using LLM-based solutions for extracting legal requirements from contract obligations and the need to determine level of adequacy for LLM-based systems against context. Oliver Karras motivated a call for guidelines on how to report evaluation research on LLM-based systems and how to make such a piece of research more understandable for domain experts and non-technical users.

While we did not observe that any consensus emerged among the panelists on specific techniques, several interesting points were raised. Both the panelists and the attendants, in the lively discussion that followed the initial statements from the panel, agreed that metrics that were originally developed to standardize the measurement of performance of different techniques for attacking the same problem in theoretical or "lab" settings, are often not well suited to predict performance in real environments and in specific contexts, and that more situation-specific measures might be needed. In particular, the

moderator expressed his belief that framing complex real-world problems as instances of standardized tasks (i.e., classification, clustering, etc.), so that standards such as Precision and Recall can be used in validation, might betray the true nature of the problem, and that at times such forced framing is caused by the expectations and standards of the academic community, rather than by adherence to the real-world concerns. Smita Ghaisas concurred that while standardized measures are useful in a first exploratory phase, in practice performance on small-scale but real-world situations are used as drivers for adoption in the practitioners' world; other panelists with industrial experience confirmed her observation. Dan Berry also reiterated his position that even if standard metrics are used, their importance in a specific context has to be established based on the specific facets of the problem being considered.

There was general consensus that no set of standardized metrics can be defined at the current stage of development of generative AI techniques, but also that existing metrics that are widely acknowledged as relevant in the community, and originally developed for Information Retrieval tasks, are not well suited to cases where generative AI techniques (and especially, language models) are used. Instead, there should be more attention to validating proposals by "the proof is in the pudding" approaches, i.e. by performing (possibly smaller-scale) case studies of direct application.

4. NEXT STEPS

After gauging interest among the attendants of Empire 2023, a proposal has been put forward for organizing the 9th edition of the Workshop series. EmpiRE 2024 has now been approved as part of the program of the 32nd IEEE International Requirements Engineering 2024 conference, which will be held in Reykjavik, Iceland, on 24-28 June 2024. An open Call for Papers will thus be published in due time.

The audience also expressed interest in a longer-form venue (e.g., a journal special issue) where more fully developed contributions on the subjects addressed by the workshop could be published, and the organizers volunteered to explore the possibility.

5. ACKNOWLEDGEMENTS

The authors would like to thank to all those who have participated in the organization of the EmpiRE 2023 workshop, and in particular the members of the Program Committee, which was comprised of:

- Raian Ali, Hamad Bin Khalifa University, Qatar
- Dan Berry, University of Waterloo, Canada
- Faiza A. Bukhsh, University of Twente, The Netherlands
- Fabiano Dalpiaz, Utrecht University, The Netherlands
- Xavier Franch, Universitat Politècnica de Catalunya, Spain
- Andrea Herrmann, Herrmann & Ehrlich, Germany
- Irum Inayat, LERO, Ireland
- Mohamad Kassab, Penn State University, USA
- Nan Niu, University of Cincinnati, USA
- Oscar Pastor, Universidad Politécnica de Valencia, Spain
- Preethu Rose, RE Group - Tata Research Development and Design Center, India
- Angelo Susi, FBK, Italy
- Didar Zowghi, CSIRO's Data61, Australia

6. REFERENCES

[1] A. Marchetto, V. Gervasi and M. Daneva, "Welcome to the 8th International Workshop on Empirical Requirements Engineering (EmpiRE 2023)," 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), Hannover, Germany, 2023, pp. 136-137

[2] R. A. Shams, M. Bano, D. Zowghi, Q. Lu and J. Whittle, "Human Value Requirements in AI Systems: Empirical Analysis of Amazon Alexa," 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), Hannover, Germany, 2023, pp. 138-145

[3] J. Shi, O. Karras, M. Obaidi and M. Tandun, "Can Videos as a By-Product of GUI Testing Help Developers Understand GUI Tests?," 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), Hannover, Germany, 2023, pp. 146-153

[4] G. H. Strønstad, I. Gerostathopoulos and E. Guzmán, "What's Next in my Backlog? Time Series Analysis of User Reviews," 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), Hannover, Germany, 2023, pp. 154-161

[5] J. Gulden and A. Rachmann, "The Square of Values for Modeling Human Values in Requirements Engineering," 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), Hannover, Germany, 2023, pp. 162-167

[6] A. Hess, M. Trapp, O. Karras, and N. Seyff, "Welcome to the Fifth International Workshop on Learning from Other Disciplines for Requirements Engineering (D4RE'21)", RE Workshops 2021: 49-50

[7] D. M. Berry, Empirical evaluation of tools for hairy requirements engineering tasks. Empir. Softw. Eng. 26(5): 111 (2021)