Replicable Theory

Benjamin Lucien Kaminski[1,2]

[1] Saarland University, Saarland Informatics Campus, Germany
[2] University College London, United Kingdom
kaminski@cs.uni-saarland.de

Abstract. In addition to the difficulties with replicating experiments or systems from some given theoretical description, we discuss *the possibility that already the theory itself is poorly replicable*. After explaining what we understand by theory replicability, we propose to scientifically evaluate whether or not the broader field of *Logic, Semantics, and Verification in Computer Science* suffers from systematic theory replicability problems.

Keywords: Replicability · Theory · Intuition

If you can't explain it simply,
you don't understand it well enough.

Albert Einstein [9]
(misattributed and misquoted)

1 Introduction

Replicability is one of the cornerstones of the scientific method. Even theory-driven fields like theoretical physics depend in one way or another on replicability: Trust in a theory is time and again established by conducting many different experiments that repeatedly confirm the theory with ever increasing precision. In order to conduct experiments that faithfully confirm a theory, the theory has to be well understood by the experiment designers.

In many experiment-driven sciences, replication crises have been discussed [8]: It has been found that many scientific findings are difficult or impossible to replicate. Recently a possible replication crisis has also been brought forward in a more formal science, namely mathematics [2]. In this article, we want to discuss replicability of theories in theoretical computer science. We first explain what we understand by replicability of a theory, show possibly detrimental situations arising from poor replicability, and finally propose to conduct an experimental evaluation of how replicable the findings in theoretical computer science are.

2 Reproduction and Replication of Theories

> *At least in my own case, understanding mathematics*
> *doesn't come from reading or even listening.*
> *It comes from rethinking what I see or hear.*
> *I must redo the mathematics in the context*
> *of my particular background.* [...]
> *When I have reorganized the mathematics in my own terms,*
> *then I feel an understanding, not before.*
>
> Stephen Smale [6]

According to the Stanford Encyclopedia of Philosophy, we can distinguish between *reproducibility* and *replicability* as follows [4]:

> *"Reproducibility is the reproducibility of an experiment, given a fixed theoretical description.* [...] *Replicability* [...] *is where experimental procedures differ to produce the same experimental result."*

In this short paper, however, we do not consider reproducibility or replicability of *experiments* but rather that of the *theoretical descriptions*. And so we ask:

When is a theoretical description reproducible or replicable?

Untrained in philosophy, we would make the following (perhaps simplistic) argument: Consider a theoretical description which is printed on paper. To *reproduce* that description, one could, for instance, *make a copy* of the paper. This process indeed (re)produces the same description and can be performed *entirely without any understanding* of the theory. We would thus argue that *any theoretical description is reproducible* in practice.

Replicating the theoretical description, on the other hand, is something else entirely. For a true replication in the sense of the Stanford "definition", we argue that one would need to *read* the description, *understand* the description and *develop an intuition* for the theory, *rethink* the theory, indeed almost *reinvent* the theory, and then *formulate a new description of the same theory*, but in one's own words. Revisiting the epigraph of this section, we coin this procedure *theory replication in the Smalian sense*. With Smalian theory replication, the *cognitive procedures to produce an equivalent*[3] *description of the same theory will differ*, if only because the persons conducting the thinking may have very different scientific backgrounds.

Assuming that we accept the Smalian notion of what theory replication constitutes, we claim that *not* every description of a theory is replicable, even if the theory and its description are sound. Indeed, we believe that the *degree of replicability* varies greatly, and can and should be considered *a* suitable and important measure of the description's quality.

[3] It is of course virtually impossible that two persons replicating a theory would arrive at exactly the *same* wording.

3 Consequences of Inaccessibility and Poor Replicability

In this section, we present two pieces of evidence that poor comprehensibility of theoretical contributions may have detrimental consequences.

Inter-universal Teichmüller Theory. Our first piece of evidence is the notable example of Shinichi Mochizuki's *Inter-universal Teichmüller Theory* [5]. Its most striking application would be to provide a proof for many outstanding conjectures in number theory, most centrally the *abc conjecture*. Alas, Mochizuki's theory is considered incomprehensible widely across the mathematical community and thus *abc* remains a conjecture to most mathematicians [7]. Still, as the implications of Mochizuki's theory would be so profound, many mathematicians, among them at least one Fields medalist, have spent significant time trying to understand Inter-universal Teichmüller Theory. The total amount of time dedicated to this endeavor is estimated to have already exceeded 30 researcher years [2], and efforts continue to this day.

The BITA Conference. The following is anecdotal evidence. While it is based on true events, all names were anonymized.

Alice served on the program committee of BITA and she was assigned to review a paper about progress on the ALAM framework.[4] The theoretical development in this paper was mostly inaccessible to Alice and there was, by her judgement, no way she could have replicated this paper, not even within an unreasonable amount of time. It emerged from the PC discussion that the paper was also rather inaccessible to the other reviewers. The reviewers agreed that it would need an ALAM expert to properly judge this paper. But finding an expert reviewer proved to virtually impossible, because all ALAM experts ended up being conflicted with one another, and in particular conflicted with the authors.

How should the reviewers have decided in such a situation? Accept the paper in the spirit of *"did not fully understand, but looks fine to me"*? Or reject the paper in the spirit of *"I will reject whatever I do not understand"*?

This whole predicament would likely have been prevented, were the paper accessible to the broad BITA audience, not solely to ALAM experts. What is more: Were the paper accessible to the whole BITA audience, then

(1) non-ALAM experts could still properly and fairly judge the paper, even if only from the perspective of an ALAM outsider, and
(2) – much more importantly – once the paper is published (be it at BITA or elsewhere), more people have access to the knowledge that the ALAM-paper authors produced.

Presuming that ALAM is any good, more people being able to replicate the ALAM theory will likely increase the number of ALAM experts over time, which would ultimately benefit the ALAM *and* the BITA community.

[4] Alice's gender, the conference name, and the framework name were randomly chosen/generated using random.org.

4 Proposed Experimental Evaluation

True to the motto *"The first step in solving any problem is recognizing there is one"*, we propose to experimentally evaluate whether research in theoretical computer science, in particular in the field of *logic, semantics, and verification* suffers from poor replicability or not. Such an experiment could be conducted with the program committee members of a major theory-driven yet broad conference like CAV, CSL, ESOP, FoSSaCS, ICALP, LICS, OOPSLA, POPL, or TACAS, to name only a few. If anyone, the program committee members should be considered experts in the respective field and furthermore they should more of less resemble the broad spectrum of the audience of the respective conference.

Experiments on program committees are not unprecedented. In 2014, the program chairs of NeurIPS, a top-tier conference in machine learning, conducted an experiment on their program committee members [3]: About 10% of the 1,678 submissions to NeurIPS 2014 were randomly selected to be reviewed by two independent program committees. A particularly striking outcome of that experiment was that, regarding which papers to accept and which not, the two program committees were only in agreement for about half of the papers. The experiment was repeated for NeurIPS 2021 [1].

Experiment Design Sketch. Our experiment on theory replicability could look roughly as follows: We randomly select a chunk of N papers from the list of *accepted papers* at conference ABC and ask each program committee member (who is willing to participate) to evaluate the replicability of all N papers. For each paper, questions for a questionnaire could be along the following lines:

1. How would you rate your expertise on the presented theoretical contribution?
2. How well did you understand the theoretical contribution?
   - If rather well, how many hours did you need to understand the material?
   - If not so well, how far did you get? (pages, percentage, etc.)
3. Do you feel confident that you could reformulate/replicate the theoretical contributions (at least the key results) in your own words?
   - How many hours would you need for the replication?
4. What level of expertise do you believe is required to perform such replication?

It is also conceivable to ask authors of accepted papers to create mini-quizzes about the key points of their theoretical contributions and then rate how well the program committee members actually understood the paper.

Obviously, also such a study should not be about praising or shaming the replicability of individual papers and the results would have to be appropriately anonymized. It should be possible, however, to draw conclusions like:

> *At least X% percent of the program committee members think they are able to replicate Y% of the papers accepted at ABC.*

5 Towards Replicable Theory

In more experimentally-driven research in our field, awareness for replicability is increasingly finding its way into the mainstream through *artifact evaluations*, e.g. at conferences like CAV, ESOP, OOPSLA, PLDI, POPL, or TACAS. For more theoretical contributions, replicability is certainly more difficult to evaluate and *"presentation"* is often already a (perhaps too secondary) evaluation criterion. But *theory replicability in the Smalian sense* is at least a more tangible — and perhaps more purposeful — criterion than *"good presentation"*. We believe that if theory replicability became a *core evaluation criterion* in the reviewing process, theoretical contributions would become more replicable on a broader scale, from which our community could only benefit. A first step, however, would be to find out whether and how much our field suffers from poor theory replicability.

References

1. Beygelzimer, A., Dauphin, Y., Percy, L., Wortman Vaughan, J.: The NeurIPS 2021 Consistency Experiment (2021), https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/, [Accessed online 24 February 2022]
2. Bordg, A.: A Replication Crisis in Mathematics? The Mathematical Intelligencer **43**(4), 48–52 (2021)
3. Cortes, C., Lawrence, N.D.: Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment. CoRR **abs/2109.09774** (2021)
4. Fidler, F., Wilcox, J.: Reproducibility of Scientific Results. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Summer 2021 edn. (2021)
5. Mochizuki, S.: A Panoramic Overview of Inter-universal Teichmüller Theory (2014)
6. Smale, S.: Finding a Horseshoe on the Beaches of Rio. The Mathematical Intelligencer **20**(1), 39–44 (1998)
7. Wikipedia: Inter-universal Teichmüller theory (2022), [Accessed online 23 February 2022]
8. Wikipedia: Replication crisis (2022), [Accessed online 24 February 2022]
9. Wikiquote: Albert Einstein (2022), [Accessed online 23 February 2022]