

Plan for research studies¹

Title of the study	Reproducibility in AI research
Responsible people	Odd Cappelen and Martin Mølneå
Time period for the study	Autumn and master
Amount of resources in PM planned	1680 hours (20 hours per week during autumn, 40 hours per week during master)
Web address for the project (if any)	

¹ This plan is based on the 6Ps of research as described in Oates, Dr Briony J (2005). *Researching Information Systems and Computing*. SAGE Publications, enhanced based on Creswell, J. W. (2014). *Research design- Qualitative, quantitative, and mixed methods approaches (4th ed.)*. SAGE Publications.

Purpose

Our research project aims to investigate the state of reproducibility in Artificial Intelligence (AI) research today. Reproducibility is an important aspect of the scientific method, and is crucial to the development of good hypotheses, and the discarding of false leads. As stated by (Oates) *“if other researchers can’t repeat an experiment and get the same results as the original researchers, then they refute the hypothesis.”* In order for this reproduction effort to be possible, the original experiment must be documented well enough to be recreated. However, recent trends show increasing concern for a “replication crisis” within the academic society (Baker). Much of produced research code is not available, and in a recent study researchers estimated that only 32% of their code was available online (Stodden). As a result of poor documentation, there is uncertainty about the percentage of computational results that can be reproduced. According to (Donoho et al.) *“It’s impossible to verify most of the results that computational scientists present at conferences and in papers.”*

Studies into the state of documentation and reproduction within computational research areas have been performed. In 2010, (Mende) investigated issues in replication by attempting to reproduce two Defect Prediction Models studies, achieving one success and one failure. A similar study was performed by (Fokkens et al.) who attempted to reproduce two experiments in Natural Language Processing, without success. In 2016, (Collberg and Proebsting) attempted to estimate the repeatability of research articles by investigating the degree of code availability. Attempting to locate and build code from 402 papers, they only succeeded in 48.3% of the instances without contacting the original authors. More recently, (Gundersen and Kjensmo) attempted to quantify the state of reproducibility within AI research through a literature review. Surveying 400 papers, they estimated that on average only 25% of the “variables” required for replication were provided.

While these studies indicate that there are issues with reproduction within AI, there has not yet been a systematic attempt to reproduce a significant number of papers in one study. There have been attempts to reproduce results, but these have usually been limited to a few articles (Mende; Fokkens et al.). While both (Collberg and Proebsting) and (Gundersen and Kjensmo) surveyed a large number of papers, neither attempted to reproduce the results of the papers. In our project we intend to perform systematic reproduction attempts on a significant number of articles. For the project, we propose the following research questions:

RQ1: Are recent research results in AI documented well enough to be replicated or reproduced with reasonable effort?

RQ1.1: To what degree is it possible to replicate or reproduce the results presented in recent AI research, using only publicly available documentation and resources?

RQ1.2: What are the defining qualities separating easily reproducible articles from non-reproducible articles within the AI field?

RQ1.3: Is it possible to formulate common guides or procedures for the creation of easily reproducible articles within the AI field?

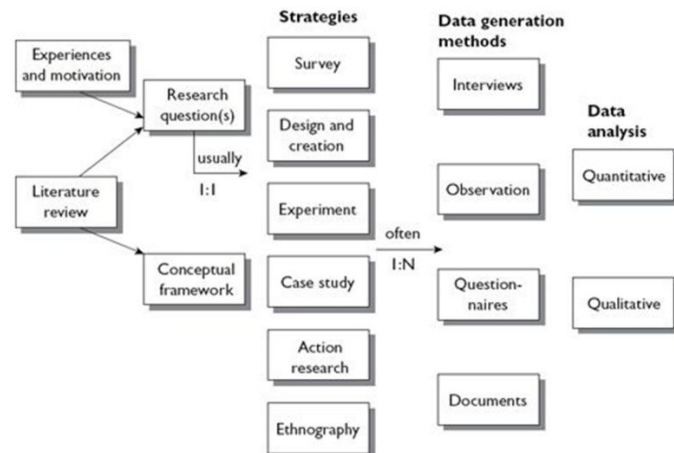
Contributions

Our chief aim is to contribute to research within the AI field by producing suggestions for improved methodology. Specifically, by providing suggestions for improved documentation practices aimed at increasing reproducibility of studies. We intend to perform a series of reproduction attempts,

document them systematically, and make the result publicly available for further discussion. Even if the guidelines we suggest are not widely employed, we hope to build upon the work already done in reproducibility research to further the discussion about how to make research reproducible. Our contributions will therefore add to the contributions of existing research such as (Collberg and Proebsting) and (Gundersen and Kjensmo).

Research Method

In our project we will attempt to reproduce a selection of studies and analyze the results. We aim to cover a significant number of research articles, and document each attempt in a systematic way. The goal of our study is to use these as a sample representing current AI research. The analysis of a sample with the goal of generalization is defining of the Survey strategy, which we intend to use.



For data generation we will employ a combination of Documents and Observations. In the first phase of the research we will attempt to reproduce a number of research articles (Documents). Each of these reproduction attempts will be documented, and each attempt will be used as an observation for the second part of the project, where we analyse the results. In this phase we will use Quantitative data analysis methods to try to understand which factors in an article influence the reproduction process.

Participants

The main participants of this project are the researchers. These include ourselves, Odd Cappelen and Martin Mølneå, master students in computer science, and Nicklas N., master student in Informatics, who will work with us on the reproduction attempts as part of his master thesis. Our supervisor for the project is Odd Erik Gundersen, associate professor at the Department of Computer Science at NTNU.

The majority of the work, including reproduction attempts and data analysis, will be performed by the three master students. The preparatory work will be divided between the researchers, with Odd and Martin creating a framework and method for the reproduction attempts, and Nicklas determining a suitable set of research articles to use in the reproduction attempts. Odd Erik will participate in and lead the writing of a research article summarizing the results of the project.

Research Paradigm

We classify our research as being within the positivism paradigm. In our project we will attempt to reproduce a series of research articles. Each reproduction attempt will be documented, and can be viewed as an observation or data point. The project uses a process of data collection, followed by interpretation and discussion of the data. The interpretation should be mostly objective, and based on quantitative data analysis.

Final Deliverables and Dissemination

The results of our project will be presented in three parts. The preparatory work and theoretical background will be presented in our project thesis. The main work, including reproduction attempts and data analysis, will be presented in our master thesis and the master thesis of Nicklas. The results will also be presented in a research article. The documentation we produce for each reproduction attempt will be made publicly available.

References

- Baker, Monya. "1,500 Scientists Lift the Lid on Reproducibility." *Nature*, no. 533, 26 May 2016, pp. 452–54.
- Collberg, Christian, and Todd A. Proebsting. "Repeatability in Computer Systems Research." *Communications of the ACM*, vol. 59, no. 3, 2016, pp. 62–69.
- Donoho, David L., et al. "Reproducible Research in Computational Harmonic Analysis." *Computing in Science & Engineering*, vol. 11, no. 1, 2009, pp. 8–18.
- Fokkens, F., et al. "Offspring from Reproduction Problems: What Replication Failure Teaches Us." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 1691–701.
- Gundersen, Odd Erik, and Sigbjørn Kjensmo. *State of the Art: Reproducibility in Artificial Intelligence*. 2017.
- Mende, Thilo. "Replication of Defect Prediction Studies." *Proceedings of the 6th International Conference on Predictive Models in Software Engineering - PROMISE '10*, 2010, doi:10.1145/1868328.1868336.
- Oates, Briony J. *Researching Information Systems and Computing*. SAGE, 2006.
- Stodden, Victoria. "The Scientific Method in Practice: Reproducibility in the Computational Sciences." *SSRN Electronic Journal*, 2010, doi:10.2139/ssrn.1550193.