PRELIMINARY VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference

State of the Art: Reproducibility in Artificial Intelligence

Odd Erik Gundersen and Sigbjørn Kjensmo
Department of Computer Science
Norwegian University of Science and Technology

Abstract

Background: Research results in artificial intelligence (AI) are criticized for not being reproducible. **Objective:** To quantify the state of reproducibility of empirical AI research using six reproducibility metrics measuring three different degrees of reproducibility. **Hypotheses:** 1) AI research is not documented well enough to reproduce the reported results. 2) Documentation practices have improved over time. **Method:** The literature is reviewed and a set of variables that should be documented to enable reproducibility are grouped into three factors: Experiment, Data and Method. The metrics describe how well the factors have been documented for a paper. A total of 400 research papers from the conference series IJCAI and AAAI have been surveyed using the metrics. **Findings:** None of the papers document all of the variables. The metrics show that between 20% and 30% of the variables for each factor are documented. One of the metrics show statistically significant increase over time while the others show no change. **Interpretation:** The reproducibility scores decrease with increased documentation requirements. Improvement over time is found. **Conclusion:** Both hypotheses are supported.

Introduction

Although reproducibility is a cornerstone of science, a large amount of published research results cannot be reproduced. This is even the case for results published in the most prestigious journals; even the original researchers cannot reproduce their own results (Aarts *et al.* 2016; Begley and Ellis 2012; Begley and Ioannidis 2014; Prinz *et al.* 2011). (Goodman *et al.* 2016) presents data from Scopus that shows that the problem with reproducibility spans several scientific fields. According to (Donoho *et al.* 2009) *"it is impossible to verify most of the computational results presented at conferences and in papers today"*. This was confirmed by (Collberg and Proebsting 2016). Out of 402 experimental papers they were able to repeat 32.3% without communicating with the author, rising to 48.3% with communication. Papers by authors with industry affiliation showed a lower rate of reproducibility. They also found that some researchers are not willing to share code and data, while those that actually share, provide too little to repeat the experiment. Guidelines, best-practices and solutions to aid reproducibility point towards open data and open source code as requirements for reproducible research (Sandve *et al.* 2013; Stodden and Miguez 2014). The increased focus on reproducibility has resulted in an increased adoption of data and code sharing policies for journals (Stodden *et al.* 2013). Still, proposed solutions for facilitating reproducibility see little adoption due to low ease-of-use and the time required to retroactively fit an experiment to these solutions (Gent and Kotthoff 2014). (Braun and Ong 2014) argues that automation should be possible to a higher degree for machine learning, as everything needed is available on a computer. Despite of this, the percentage of research that is reproducible is not higher for machine learning and artificial intelligence (AI) research (Hunold and Träff 2013; Fokkens *et al.* 2013; Hunold 2015).

The scientific method is based on reproducibility; *"if other researchers can't repeat an experiment and get the same result as the original researchers, then they refute the hypothesis" (Oates 2006, p. 285)*. Hence, the inability to reproduce results affects the trustworthiness of science. To ensure high trustworthiness of AI and machine learning research measures must be taken to increase its reproducibility. However, before measures can be taken, the state of reproducibility in AI research must be documented. The state of reproducibility can only be documented if a proper framework is built.

Our *objective* is to quantify the state of reproducibility of empirical AI research, and our *main hypothesis* is that the documentation of AI research is not good enough to reproduce the reported results. We also investigate a *second hypothesis*, which is that documentation practices have improved during recent years. Two predictions were made, one for each hypothesis. The *first prediction* is that the current documentation practices at top AI conferences render most of the reported research results irreproducible, and the *second prediction* is that a larger portion of the reported research results are reproducible when comparing the latest installments of conferences to earlier installments. We surveyed research papers from the two top AI conference series, International Joint Conference on AI (IJCAI) and the Association for the Advancement of AI (AAAI) to test the hypotheses. Our *contributions* are twofold: i) an investigation of what reproducibility means for AI research and ii) a quantification of the state of reproducibility of AI research.

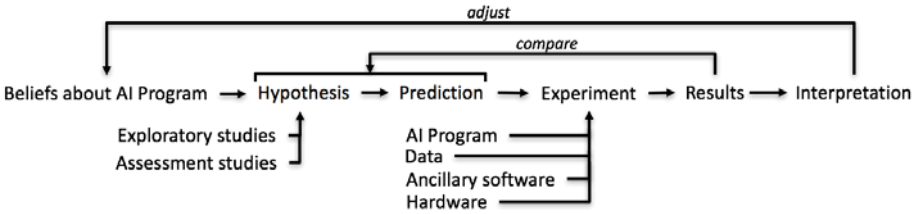

Figure 1: By comparing the results of an experiment to the hypotheses and predictions that are being made about the AI program, we interpret the results and adjust our beliefs about them.

Reproducing Results

We base the survey on a concise definition of reproducibility and three degrees of reproducibility. These definitions are based a review of the scientific method and the literature.

The Scientific Method in AI Research

Different strategies for researching information systems and computing exist (Oates 2006), and these include theory development and experiments among others. The scientific method and reproducibility is closely connected to experiments and empirical studies. We can distinguish between four different classes of empirical studies: 1) exploratory, 2) assessment, 3) manipulation and 4) observational studies (Cohen 1995). While exploratory and assessment studies are conducted to identify and suggest possible hypotheses, manipulation and observational studies test explicit and precise hypotheses. Although the scientific method is based on evaluating hypothesis, exploratory and assessment studies are not mandatory sub-processes of it. However, they may be conducted in order to formulate the hypotheses.

The targets of study in AI research are AI programs and their behavior (Cohen 1995). Changes to the AI program's structure, the task or the environment can affect the program's behavior. An *AI program* implements an abstract algorithm or system as a program that can be compiled and executed. Hence, the AI program is something distinct from the conceptual idea that it implements, which we will refer to as an *AI method*. Experiments should be formulated in such a way that it is clear whether they test hypotheses about the AI program or the AI method. Examples of tasks performed by AI methods include classification, planning, learning, decision making and ranking. The environment of the AI program is described by data. Typically, when performing AI experiments in supervised learning, the available data has to be divided into a training set, a validation set and a test set (Russell and Norvig 2009).

According to the scientific method and before performing an experiment, one should formulate one or more hypotheses about the AI program under investigation and make predictions about its behavior. The results of the experiments are interpreted by comparing their results to the hypotheses and the predictions. Beliefs about the AI program should be adjusted by this interpretation. The adjusted beliefs can be used to formulate new hypotheses, so that new experiments can be conducted. If executed honestly with earnest interpretations of the results, the scientific method updates our beliefs about an AI program so that they should converge towards objective truth. Figure 1 illustrates the scientific process of AI research as described here.

The Terminology of Reproducibility

While researchers in computer science agree that empirical results should be reproducible, what is meant by reproducibility is neither clearly defined nor agreed upon. (Stodden 2011) distinguishes between replication and reproduction. Replication is seen as re-running the experiment with code and data provided by the author, while reproduction is a broader term *"implying both replication and the regeneration of findings with at least some independence from the [original] code and/or data"*. (Drummond 2009) states that replication, as the weakest form of reproducibility, can only achieve checks for fraud. Due to the inconsistencies in the use of the terms replicability and reproducibility, (Goodman *et al.* 2016) proposes to extend reproducibility into:

Methods reproducibility: The ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results.

Results reproducibility: The production of corroborating results in a new study, having used the same experimental methods.

Inferential reproducibility: The drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study.

Replication, as used by (Drummond 2009) and (Stodden 2011), is in line with methods reproducibility as proposed by (Goodman *et al.* 2016) while reproducibility seems to entail both results reproducibility and inferential reproducibility. (Peng 2011) on the other hand suggests that reproducibility is on a spectrum from publication to full replication. This view neglects that results produced by AI methods can be reproduced using different data or different implementations. Results generated by using other implementations or other data can lead to new interpretations, which broadens the beliefs about the AI method, so that generalizations can be made. Despite the disagreements in terminology, there is a clear agreement on the fact that the reproducibility of research results is not one thing, but that empirical research can be assigned to some sort of spectrum, scale or ranking that is decided based on the level of documentation.

Reproducibility

We define reproducibility in the following way:

Definition. *Reproducibility in empirical AI research is the ability of an* independent research team *to produce the same* results *using the same* AI method *based on the* documentation *made by the original research team.*

Hence, reproducible research is empirical research that is documented in such detail by a research team that other researchers can produce the same results using the same AI method. According to (Sandve *et al.* 2013), a minimal requirement of reproducibility is that you should at least be able to reproduce the results yourself. We interpret this as repeatability and not reproducibility. Our view is that an important aspect of reproducibility is that the experiment is conducted independently. We will briefly discuss the three terms *AI method*, *results* and *independent research team* in this section. The next section is devoted to documentation.

An independent research team is one that conducts the experiment *by only using the documentation* made by the original research team.Enabling others to reproduce the same results is closely related to trust. Most importantly, other researchers can be expected to be more objective. They have no interest in inflating the performance of a method they have not developed themselves. More practically, they will not share the same preconceptions and implicit knowledge as the first team reporting the research. Also, other researchers will not share the exact same hardware running the exact same copies of software. All of this helps controlling for noise variables related to both the hardware and ancillary software as well as implicit knowledge and preconceptions.

The distinction between the AI program and the AI method is important. We must as far as possible remove any uncertainties to whether other effects than the AI method are responsible for the results. The concept of using an agent system for solving some problem is different from the specific implementation of the agent system. If the results are dependent on the implementation of the method, the hardware it is running on or the experiment setup, then the characteristics of the AI method do not cause the results.

The results are the output of the experiment, in other words, the dependent variables of the experiment (Cohen 1995), which typically are captured by performance measures. The result is the target of the investigation when reproducing an experiment; we want to ensure that the performance of the AI method is the same even if we change the implementation, the operating system or the hardware that is being used to conduct the experiment. As long as the results of the original and the reproduced experiments are the same, the original experiment is reproducible. What constitutes the same results depends on to which degree the results are reproduced.

Documenting for Reproducibility

In order to reproduce the results of an experiment, the documentation must include relevant information, which must be specified to a certain level of detail. What is relevant and how detailed the documentation must be are guided by whether it is possible to reproduce the results of the experiment using this information only. Hence, the color of the researcher's jacket is usually not relevant for reproducing the results. Which operating system is used on the machine when executing the experiment can very well be relevant though.

So what exactly is relevant information? The objective of (Claerbout and Karrenbach 1992; Buckheit and Donoho 1995) was to make it easy to rerun experiments and trace methods that produced the reported results. For (Claerbout and Karrenbach 1992), this meant sharing everything on a CD-ROM, so that anyone could read the research report and execute the experiments by pushing a button attached to every figure. (Buckheit and Donoho 1995) shared Wavelab, a Matlab package, that made all the code needed for reproducing their figures in one of their papers. (Goodman *et al.* 2016) highlights that *"reporting of all relevant aspects of scientific design, conduct, measurements, data and analysis"* is necessary for all three types of reproducibility. This is in line with the view of (Stodden 2011), which is that availability of the computational environment is necessary for computational reproducibility. (Peng 2011) argues that a paper alone is not enough, but that linked and executable code and data is the gold standard. We have grouped the documentation into three categories: 1) method, 2) data and 3) experiment.

Method: The method documentation includes the AI method that the AI program implements as well as the a motivation of why the method is used. As the implementation does not contain the motivation and intended behavior, sharing the implementation of the AI program is not enough. It is important to give a high-level description of the AI method that is being tested. This includes what the AI method intends to do, why it is needed and how it works. To decrease ambiguity, a description of how a method works should contain pseudo code and an explanation of the pseudo code containing descriptions of the parameters and sub-procedures. Sharing of the AI method is the objective of most research papers in AI. The problem that is investigated must be specified, the objective of the research must be clear and so must the research method being used.

Data: Sharing the data used in the experiment is getting simpler with open data repositories, such as the UCI Machine Learning Repository (Lichman 2013). Reproducing the results fully requires the procedure for how the data set has been divided into training, validation and test sets and which samples belong to the different sets. Sharing the validation set might not be necessary when all samples in the training set are used or might be hard when the method picks the samples randomly during the experiment. Data sets often change, so specifying the version is relevant. Finally in order to compare results, the actual output of the experiment, such as the classes or decisions made, are required.

Experiment: For others to reproduce the results of an experiment, the experiment and its setup must be shared. The experiment contains code as well as an experiment description. Proper experiment documentation must explain the purpose of the experiment. The hypotheses that are tested and the predictions about these must be documented, and so

must the results and the analysis. In order to rule out the possibilities that the results can be attributed to the hardware or ancillary software, the hardware and ancillary software used must be properly specified. The ancillary software includes, but is not restricted to, the operating system, programming environment and programming libraries used for implementing the experiment. Sharing the experiment code is not limited to open sourcing the AI program that is investigated, but sharing of the experiment setup with all independent variables, such as hyperparameters, as well as the scripts and environmental settings is required. The experiment setup consists of independent variables that control the experiment. These variables configure both the ancillary software and the AI program. Hyperparameters are independent variables that configure the AI method and examples include the number of leaves or depth of a tree and the learning rate. Documented code increases transparency.

In conclusion, there are different degrees to how well an empirical study in AI research can be documented. The degrees depend on whether the method, the data and the experiment are documented and how well they are documented. The gold standard is sharing all of the three groups of documentation through access to a running virtual machine in the cloud containing all the data, runnables, documentation and source code, as this includes the hardware and software stack as well and not only the software libraries used for running the experiments which was the case with the proposed solutions by (Claerbout and Karrenbach 1992; Buckheit and Donoho 1995). This is not necessarily practical, as it requires costly infrastructure that has a high maintenance cost. Another practical consideration is related to how long the infrastructure should and can be guaranteed to run and produce the same results.

Degrees of Reproducibility

We propose to distinguish between three different degrees of reproducibility, where an increased degree of reproducibility conveys an increased generality of the AI method. An increased generality means that the performance of the AI method documented in the experiment is not related to one specific implementation or the data used in the experiment; the AI method is more general than that. The three degrees of reproducibility are defined as follows:

R1: Experiment Reproducible The results of an experiment are experiment reproducible when the execution of the same implementation of an AI method produces the same results when executed on the same data.

R2: Data Reproducible The results of an experiment are data reproducible when an experiment is conducted that executes *an alternative implementation of the AI method* that produces the same results when executed on the same data.

R3: Method Reproducible The results of an experiment are method reproducible when the execution of *an alternative implementation of the AI method* produces the same results when executed on *different data*.

Results that are R1 reproducible require the same software and data used for conducting the experiment and a detailed description of the AI method and experiment. This is what is called fully reproducible by (Peng 2011) and method reproducibility by (Goodman *et al.* 2016). We call it experiment reproducible as everything required to run the experiment is needed to reproduce the results. The results when re-running the experiment should be *exactly the same*, as reported in the original experiment. Any differences can only be attributed to differences in hardware given that the ancillary software is the same.

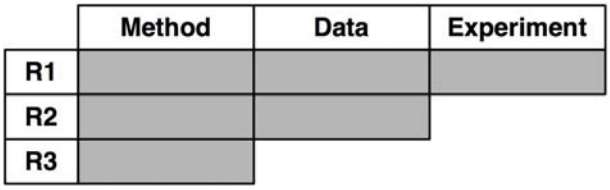

Figure 2: The three degrees of reproducibility are defined by which documentation is used to reproduce the results.

Results that are R2 reproducible require only the method description and the data in order to be reproduced. This removes any noise variables related to implementation and hardware. The belief that the result is being caused by the AI method is strengthened. Hence, the generality of the AI method is increased compared to an AI method that is R1 reproducible. As the results are achieved by running the AI method on the same data as the original experiment, there is still a possibility that the performance can only be achieved using the same data. The results that are produced, the performance, using a different implementation should be the same if not exactly the same. Differences in results can be attributed to different implementations and hardware, such as different ways of doing floating point arithmetic. However, differences in software and hardware could have significant impact on results because of rounding errors in floating point arithmetic (Hong *et al.* 2013).

Results that are R3 reproducible only requires the method documentation to be reproduced. If the results are reproduced, all noise variables related to implementation, hardware and data have been removed, and it is safe to assume that the results are caused by the AI method. As the results are produced by using a new implementation on a new data set, the AI method is generalized to other data and the implementation used in the original experiment. In order for a result to be R3 reproducible the results of the experiments must support the same hypotheses and thus support the same beliefs. The same interpretations cannot be made unless the results are statistically significant, so the analysis should be supported by statistical hypothesis testing with a p-value of 0.005 for claiming new discoveries (Johnson 2013; Benjamin *et al.* 2017).

When it comes to generality of the results the following is true: $R1 < R2 < R3$, which means that R1 reproducible results are less general than R2 reproducible results, which in turn are less general than R3 reproducible results. However, when it comes to the documentation required, the following is the case: $doc(R3) \subset doc(R)2 \subset doc(R1)$. The

documentation needed for R3 reproducibility is a subset of the documentation required for R2 reproducibility and the documentation required for R2 is a subset of the documentation required for R1 reproducibility. R3 reproducible is the most general reproducibility degree that also requires the least amount of information.

Current practice of publishers is not to require researchers to share data and implementation when publishing research papers. The current practice enables R3 reproducible results that have the least amount of transparency. For (Goodman *et al.* 2016), the goal of transparency is to ease evaluation of the weight of evidence from studies to facilitate future studies on actual knowledge gaps and cumulative knowledge, and reduce time spent exploring blind alleys from poorly reported research. This means that current practices enable other research teams to reproduce results at the highest reproducibility degree with the least effort of the original research team. The majority of effort in reproducing results, lays with the independent team, instead of the original team. Transparency does not only reduce the effort needed to reproduce the results, but it also builds trust in them. Hence, the results that are produced by current practices are the least trustworthy from a reproducibility point of view, because of the lack in transparency; the evidence showing that the results are valid is not published.

Research Method

We have conducted an observational experiment in form of a survey of research papers in order to generate quantitative data about the state of reproducibility of research results in AI. The research papers have been reviewed, and a set of variables have been manually registered. In order to compare results between papers and conferences, we propose six metrics for deciding whether research results are R1, R2, and R3 reproducible as well as to which degree they are.

Survey

In order to evaluate the two hypotheses, we have surveyed a total of 400 papers where 100 papers have been selected from each of the 2013 and 2016 installments of the conference IJCAI and from the 2014 and 2016 installments of the conference series AAAI. With an exception of 50 papers from IJCAI 2013, all the papers have been selected randomly to avoid any selection biases. Table 1 shows the number of accepted papers (the population size), the number of surveyed papers (sample size) and the margin of errors for a confidence level of 95% for the four conferences. We have computed the margin of error as half the width of the confidence interval, and for our study the margin of error is 4.29%. All the data and the code that has been used to calculate the reproducibility scores and generate the figures can be found on Github[1].

Factors and Variables

We have identified a set of variables that we believe are good indicators for reproducibility after reviewing the literature.

[1]https://github.com/aaai2018-paperid-62/aaai2018-paperid-62

Table 1: Population size, sample size (with number of empirical studies) and margin of error for a confidence level of 95% for the four conferences and total population.

Conference	Population size	Sample size	MoE
IJCAI 2013	413	100 (71)	8.54%
AAAI 2014	213	100 (85)	7.15%
IJCAI 2016	551	100 (84)	8.87%
AAAI 2016	549	100 (85)	8.87 %
Total	1726	400 (325)	4.30%

These variables have been grouped together into the three factors Method, Data and Experiment. For each surveyed paper, we have registered these variables. In addition, we have collected some extra variables, which have been grouped together in Miscellaneous. The following variables have been registered for the three factors:

Method: How well is the research method documented?

Problem (*): The problem the research seeks to solve.

Objective/Goal (*): The objective of the research.

Research method (*): The research method used.

Research questions (*): The research question asked.

Pseudo code: Method described using pseudo code.

Data: How well is the data set documented?

Training data: Is the training set shared?

Validation data: Is the validation set shared?

Test data: Is the test set shared?

Results: Are the results shared?

Experiment: How well is the implementation and the experiment documented?

Hypothesis (*): The hypothesis being investigated.

Prediction (*): Predictions related to the hypotheses.

Method source code: Is the method open sourced?

Hardware specifications: Hardware used.

Software dependencies: For method or experiment.

Experiment setup: Is the setup including hyperparameters described?

Experiment source code: Is the experiment code open sourced?

Miscellaneous: Different variables that describe the research.

Research type: Experimental (E) or theoretical (T).

Research outcome: Is the paper reporting a positive or a negative result (positive=1 and negative=0).

Affiliation: The affiliation of the authors. Academia (0), collaboration (1) or industry (2).

Contribution (*): Contribution of the research.

All variables were registered as true (1) or false (0) unless otherwise specified. When surveying the papers, we have looked for explicit mentions of the variables marked with an asterix (*) above. For example, when reviewing the variable

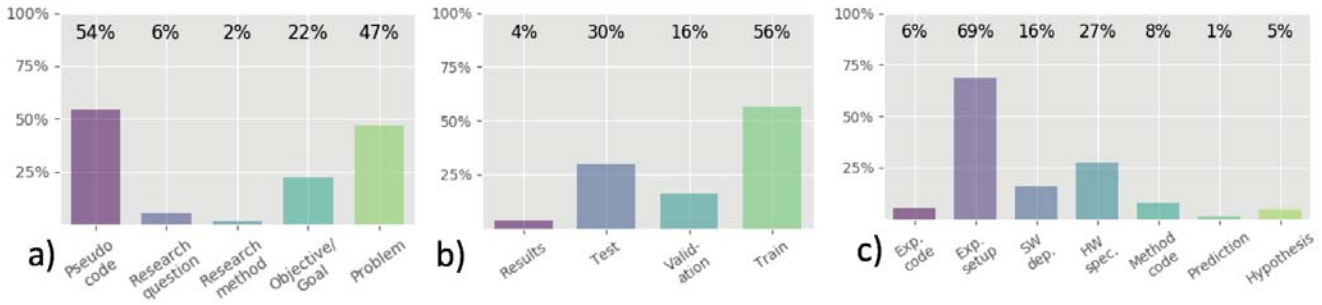

Figure 3: Percentage of papers documenting each variable for the three factors: a) Method, b) Data and c) Experiment.

Problem, we have looked for an explicit mention of the problem being solved, such as "To address this problem, we propose a novel navigation system ..." (De Weerdt *et al.* 2013). The decision to use explicit mentions of the terms, such as contribution, goal, hypothesis and so on, can be disputed. However, the reasons for looking for explicit mentions are both practical and idealistic. Practically, it is easier to review a substantial amount of papers if the criteria are clear and objective. If we did not follow this guideline, the registering of variables would lend itself to subjective assessment rather than objective, and the results could be disputed based on how we measured the variables. Our goal was to get results with a low margin of error, so that we could draw statistically valid conclusions. In order to survey enough papers, we had to reduce the time we used on each paper. Explicit mentions supported this. Idealistically, our attitude is that research documentation should be clear and concise. Explicit mentions of which problem is being solved, what the goal of doing the research is, which hypothesis is being tested and so on are required to remove ambiguity from the text. Less ambiguous documentation increases the reproducibility of the research results.

Quantifying Reproducibility

We have defined a set of six metrics to quantify whether an experiment e is R1, R2 or R3 reproducible and to which degree. The metrics measure how well the three factors method, data and experiment are documented. The three metrics $R1(e)$, $R2(e)$ and $R3(e)$ are boolean metrics that can be either true or false:

$$R1(e) = Method(e) \wedge Data(e) \wedge Exp(e), \quad (1)$$

$$R2(e) = Method(e) \wedge Data(e), \quad (2)$$

$$R3(e) = Method(e), \quad (3)$$

where $Method(e)$, $Data(e)$ and $Exp(e)$ is the conjunction of the truth values of the variables listed under the three factors Method, Data and Experiment in the section *Factors and Variables*. This means that for $Data(e)$ to be true for an experiment e, the training data set, the validation data set, the test data set and the results must be shared for e. Hence, $R1(e)$ is the most strict requirement while $R3$ is the most relaxed requirement when it comes to the documentation of an experiment e, as $R3(e)$ requires only variables of the factor Method to be true while $R1(e)$ requires all variables for all the three factors to be true.

The three metrics $R1(e)$, $R2(e)$ and $R3(e)$ are boolean metrics, so they will provide information on whether an experiment is R1, R2 or R3 reproducible in a strict sense. They will however not provide any information on to which degree experiments are reproducible, unless an experiment meets all the requirements. Therefore we suggest the three metrics $R1D(e)$, $R2D(e)$ and $R3D(e)$ for measuring to which degree the the results of an experiment e is:

$$R1D(e) = \frac{\delta_1 Method(e) + \delta_2 Data(e) + \delta_3 Exp(e)}{\delta_1 + \delta_2 + \delta_3} \quad (4)$$

$$R2D(e) = \frac{\delta_1 Method(e) + \delta_2 Data(e)}{\delta_1 + \delta_2}, \quad (5)$$

$$R3D(e) = Method(e), \quad (6)$$

where $Method(e)$, $Data(e)$ and $Exp(e)$ is the weighted sum of the truth values of the variables listed under the three factors Method, Data and Experiment. The weights of the factors are δ_1, δ_2 and δ_3 respectively. This means that the value for $Data(e)$ for experiment e is the summation of the truth values for whether the training, validation, and test data sets as well as the results are shared for e. It is of course also possible to give different weights to each variable of a factor. We use a uniform weight for all variables and factors for our survey, $\delta_i = 1$. For an experiment e_1 that has published the training data and test data, but not the validation set and the results $Data(e) = 0.5$. Note that some papers have no value for the training and validation sets if the experiment does not require either. For these papers, the δ_i weight is set to 0.

Results and Discussion

Figure 3 shows percentage of research papers that have documented the different variables for the three factors. None of the three factors are documented very well according to the survey. As can bee seen by analyzing the factor Method, an explicit description of the motivation behind research is not common. Figure 4 (b) shows this as well. None of the papers document all five variables, and most of them (90%) document two or less. This might be because it is assumed that

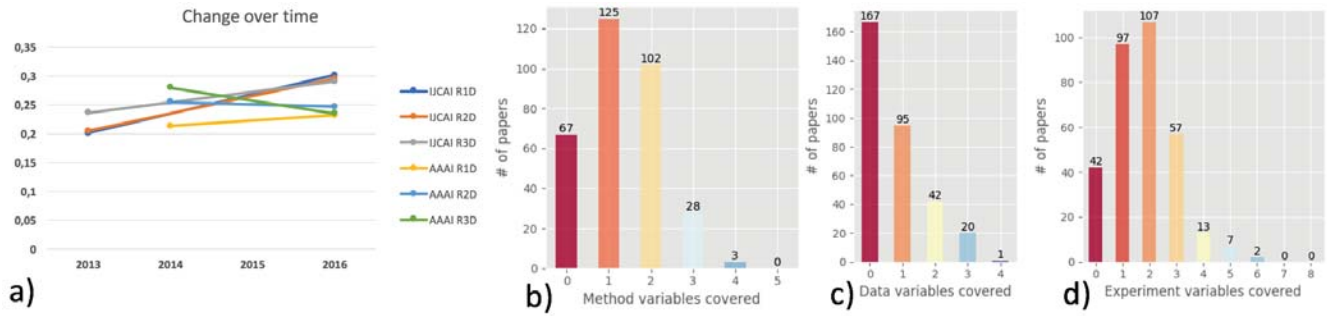

Figure 4: a) Change in the RXD metrics. b), c) and d) show the amount of variables registered for the three factors for all papers.

researchers in the domain are acquainted with the motivations and problems. Figure 3 (b) shows that few papers provide the results of the experiment, although, compared to the other two factors, an encouraging 49% of the research papers share data as seen from Figure 4 (c). The experiments are not documented well either as can be seen in figures 3 (c) and 4 (d). The variable Experiment setup is given a high score, which indicates that the experiment setup is documented to some degree. As we have not actually tried to reproduce the results, we have not ensured that the experiment setup is documented in enough detail to run the experiment.

The amount of empirical papers are shown in Table 1. For each conference, between 15% and 29% of the the randomly selected samples are not empirical. In total, 325 papers empirical and considered in the analysis. Table 2 presents the results of the RXD (R1D, R2D and R3D) metrics. All the RXD metrics vary between 0.20 and 0.30. This means that only between a fifth and a third of the variables required for reproducibility are documented. For all papers, R1D has the lowest score with 0.24, R2D has a score of 0.25 and R3D has a score of 0.26. The general trend is that R1D is lower than the R2D scores, which again are lower than the R3D scores. This is not surprising, as R1D has fewer variables than R2D, which has fewer variables than R3D. However, given the error there is little variation among the three reproducibility degrees.

The RX (R1, R2 and R3) scores were 0.00 for all papers. No paper had full score on all variables for the factor Method, and it is required for all the three RX metrics. The three RX metrics are very strict and are not very informative for a survey such as this. They might have a use though, as guidelines for reviewers of conferences and journal publications. The three RXD metrics do not have the same issue as the RX metrics, as they measure the degree of reproducibility between 0 and 1.

There is a clear increase in the RXD scores from IJCAI 2013 to IJCAI 2016, see figure 4 a). However, the trend is not as clear for AAAI as the R2D and R3D scores decrease. Table 3 shows the combined scores for the earlier years (2013 and 2014, 156 papers) and the combined scores for 2016 (169 papers). The results show that there is a slight, but statistically significant increase for R1D. The increase for R2D is not statistically significant, and there is no change for R3D. This means that only the experiment documentation has improved with time, and that there is no such evidence for the documentation of methods and data.

Table 2: The 95% confidence interval for the mean R1D, R2D and R3D scores where $\varepsilon = 1.96\sigma_{\bar{x}}$ and $\sigma_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{N}}$.

Conference	$R1D \pm \varepsilon$	$R2D \pm \varepsilon$	$R3D \pm \varepsilon$
IJCAI 2013	0.20 ± 0.02	0.20 ± 0.03	0.24 ± 0.04
AAAI 2014	0.21 ± 0.02	0.26 ± 0.03	0.28 ± 0.04
IJCAI 2016	0.30 ± 0.03	0.30 ± 0.04	0.29 ± 0.04
AAAI 2016	0.23 ± 0.02	0.25 ± 0.04	0.24 ± 0.04
Total	0.24 ± 0.01	0.25 ± 0.02	0.26 ± 0.02

Table 3: The 95% confidence interval for the mean R1D, R2D and R3D scores when combining the papers from all four installments of IJCAI and AAAI into two groups according to the years they were published. One group contains all papers from 2013 and 2014 and the other group contains all the papers from 2016.

Years	$R1D \pm \varepsilon$	$R2D \pm \varepsilon$	$R3D \pm \varepsilon$
2013/2014	0.21 ± 0.02	0.23 ± 0.02	0.26 ± 0.03
2016	0.27 ± 0.02	0.27 ± 0.03	0.26 ± 0.03

Conclusion

The survey confirms our prediction that the current documentation practices at top AI conferences render most of the reported research results irreproducible, as the R1, R2 and R3 reproducibility metrics show that no papers are fully reproducible. Only 24% of the variables required for R1D reproducibility, 25% of the variables required for R2D reproducibility and 26% of the variables required for R3D reproducibility are documented. When investigating whether there is change over time, we see improvement, which then confirms our second hypothesis. No improvement is indicated by the R1, R2, R3, R2D and R3D metrics. There is however a statistically significant improvement in the R1D metric. Hence, overall there is an improvement.

Acknowledgments

This work has been carried out at the Telenor-NTNU AI Lab, Norwegian University of Science and Technology, Trondheim, Norway.

References

Alexander A. Aarts, Christopher J. Anderson, Joanna Anderson, Marcel A. L. M. van Assen, Peter R. Attridge, Angela S. Attwood, Jordan Axt, Molly Babel, Štěpán Bahník, Erica Baranski, and et al. Reproducibility project: Psychology, Dec 2016.

C. Glenn Begley and Lee M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, mar 2012.

C. G. Begley and J. P. A. Ioannidis. Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, 116(1):116–126, dec 2014.

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, Eric-Jan Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature Human Behaviour*, 2017.

Mikio L Braun and Cheng Soon Ong. Open science in machine learning. *Implementing Reproducible Research*, page 343, 2014.

Jonathan B. Buckheit and David L. Donoho. Wavelab and reproducible research. Technical report, Standford, CA, 1995.

Jon F. Claerbout and Martin Karrenbach. Electronic documents give reproducible research a new meaning. In *Proceedings of the 62nd Annual International Meeting of the Society of Exploration Geophysics*, New Orleans, USA, 1992. 25 to 29 October 1992.

Paul R Cohen. *Empirical methods for artificial intelligence*, volume 139. MIT press Cambridge, MA, 1995.

Christian Collberg and Todd A. Proebsting. Repeatability in computer systems research. *Communications of the ACM*, 59(3):62–69, February 2016.

Mathijs M De Weerdt, Enrico H Gerding, Sebastian Stein, Valentin Robu, and Nicholas R Jennings. Intention-aware routing to minimise delays at electric vehicle charging stations. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 83–89. AAAI Press, 2013.

David L Donoho, Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. Reproducible research in computational harmonic analysis. *Computing in Science & Engineering*, 11(1), 2009.

Chris Drummond. Replicability is not reproducibility: nor is it good science. *International Conference on Machine Learning*, June 2009.

Antske Fokkens, Marieke Van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1691–1701. Association for Computational Linguistics (ACL), 2013.

Ian P Gent and Lars Kotthoff. Recomputation. org: Experiences of its first year and lessons learned. In *Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on*, pages 968–973. IEEE, 2014.

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12, jun 2016.

Song-You Hong, Myung-Seo Koo, Jihyeon Jang, Jung-Eun Esther Kim, Hoon Park, Min-Su Joh, Ji-Hoon Kang, and Tae-Jin Oh. An evaluation of the software system dependency of a global atmospheric model. *Monthly Weather Review*, 141(11):4165–4172, 2013.

Sascha Hunold and Jesper Larsson Träff. On the state and importance of reproducible experimental research in parallel computing. *CoRR, abs/1308.3648*, 2013.

Sascha Hunold. A survey on reproducibility in parallel computing. *CoRR, abs/1511.04217*, 2015.

Valen E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317, 2013.

M. Lichman. UCI machine learning repository, 2013.

Briony J Oates. *Researching Information Systems and Computing*. SAGE Publications Ltd, 2006.

Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–712, aug 2011.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A modern approach*. Prentice-Hall, 2009.

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10):e1003285, oct 2013.

Victoria C. Stodden and Sheila Miguez. Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *Journal of Open Research Software*, 2(1):e21, jul 2014.

Victoria C. Stodden, Peixuan Guo, and Zhaokun Ma. Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE*, 8(6):e67111, jun 2013.

Victoria C. Stodden. Trust your science? Open your data and code. *Amstat News*, pages 21–22, 2011.