https://royalsocietypublishing.org/doi/full/10.1098/rsta.2020.0210

The fundamental principles of reproducibility

Abstract
Reproducibility is a confused terminology. Fundamental view on reproducibility. Scientific method analysed and characterized. The literature on reproducibility and replication is surveyed. Machine learning used to exemplify approach. Reproducibility is defined and three different degrees of reproducibility as well as four types of reproducibility specified.

1. Introduction

"What is reproducibility?" Not clearly defined. Elusive concept. Many different definitions, with each capturing central but different properties.
Lack of definition → issue in automating the scientific endeavour. Next grand challenge for AI.

General definition that applies to all sciences. Computer science the right lens to see and try to understand reproducibility through. Represent the world and develop step-by-step processes of how to solve problems on a computer. CS developed many tools for analysing and understanding a problem before solving it. Examples are UML for visualizing, Petri nets for description of distributed systems and behaviour trees for modelling plan execution.

2. The scientific method and machine learning

Section outlines a process view of shared view of scientific method, provides example of machine learning. Acquiring knowledge about the world through observations to ensure our belief about the world are as certain as they can be. Done through testing of hypotheses. A hypothesis is statement that is either true or false, and is tested by conducting an experiment. Reproducbility is about confirming results of a past experiment by conducting a reproducibility experiment.
Process view of scientific method, 10 sub-processes are:
1)
Observe the world to form beliefs about it: We typically observe the world in an unsystematic and unstructured way, so that the observations have a sampling bias. Applying the scientific method helps reduce this bias. Exploratory and assessment studies are conducted to form these beliefs.
2)
Explain causes and effects by forming a scientific theory: The scientific theory underlying the example is that deep neural networks are models of the brain, although simple ones, and as such intelligence could emerge from them.
3)
Formulate a genuine test of the scientific theory as a hypothesis: The hypothesis that is tested in the example is that the performance of biological inspired deep convolutional neural networks is competitive with human performance on computer vision benchmark tasks.

4)
Design an experiment to test the hypothesis and document the experiment in a research protocol: Ideally, research protocols documenting the experiments should be written ahead of them being conducted. Experiment design has traditionally not been a very structured process in computer science. Often, making the research protocol and the experiment is an iterative process in which the experiment design, implementation and execution are done interchangeably. This could be considered HARKing or hypothesizing after results are known, which is not good practice.
5)
Implement the experiment so that it is ready to be conducted: The target of research in a machine learning experiment is often a system or an algorithm such as the biologically inspired deep neural network architecture of the example. In figure 1, this is called the target system, and it is a piece of code that is written by the researchers themselves that has some potential properties that they want to test. The experiment has many more components, many of which are also code written by the researchers. Often, the collected data requires pre-processing. In the example, the raw images are translated, scaled, rotated and distorted. Hyperparameters, random seeds, the learning rate, the number of epochs and so on are configured in the experiment set-up. Another piece of code could define the experiment workflow. Figure 1a in the example [11] illustrates the pre-processing workflow. Also, data must be gathered. In this case, the data are compiled from six different computer vision benchmarking datasets. Furthermore, the experiment must be run on some hardware, which in the example is a graphical processing unit (GPU). Most experiments rely on ancillary software to run. Ancillary software includes but is not limited to the operating system and software libraries that simplify the execution of the experiment.
6)
Conduct the experiment to produce outcomes: Conducting a machine learning experiment typically requires executing software on a computer without any input from the outside world except for the training and test data. The outcomes that are produced in the example are class labels for the images in the test data from the benchmarking datasets.
7)
Analyse the outcomes to make an analysis: The analysis typically consists of visualizations of the outcomes and metrics that are computed based on the outcomes. In the example, the analysis is to compute error rates and display them in tables.
8)
Interpret the analysis: The analysis has to be interpreted. This interpretation leads to a conclusion, and this conclusion is the result of the experiment. In the example, the analysis of the outcomes shows that computers have lower errors than humans on these tasks, which when interpreted leads to the conclusion that deep convolutional neural networks are competitive with humans on widely used computer vision benchmark tasks.
9)
Update beliefs according to the interpretation: Scientists update their beliefs based on trusted results and interpretations even if they are the exact opposite of previous beliefs. Surprising and counterintuitive results might not be trusted immediately, but they could spur new and different experiments to increase the trust. Although the analysis (low errors by deep learning methods on visual benchmarking tasks) has been reproduced many

times [16], the claim that deep learning achieves super-human performance is still debated [17].
10)
Observe the world systematically: To be a trusted source of knowledge, experiments must be designed to remove biases. As many datasets are biased [18], the bias can be reduced by conducting the experiment on several datasets, similar to what is done in the example.

3. A survey of definitions

Table 1 many definitions

4. On the difference between corroboration and reproducibility

Merriam-Webster's dictionary corroborate is to support with new evidence or make more certain and reproduce is to produce again or make a copy of. Reproducibility is related to experiments, while theories and hypotheses can only be corroborated.

Theory is a concept or a belief, no physical manifestations.

Conducting the same experiment over again will not add new evidence to the correctness of the hypothesis. It will only strengthen the belief in the result of the experiment.
Make us more certain that we draw the correct conlusion from experiment. When conducted several times, the same experiment can produce many different outcomes as evident by how reproducibility and repeatabililty are defined in measurement theory[25]. However, this is not only restricted to measurements. Running the same experiment on a computer does not necessarily produce the same outcomes as discussed by Nagarajan [41].
Differences in outcomes could be caused by using different hardware architectures, operating systems, compiler settings [42], intentional stochasticity in algorithms, random number generator seeds and hardware, such as conducting the experiment on a graphics processing unit (GPU) as discussed by Henderson et al. [43].
A result can be analysed in many different ways by using different charts, performance metrics or error analyses and these can reveal different information that allows for different interpretation, which means that different investigators might interpret the analysis in different ways and thus reach different conclusions [34].

So, to conclude, our beliefs in scientific theories and hypotheses are strengthened with new evidence. Reproducing the results of an experiment will increase our trust in the conclusions we draw from the experiment. The belief in the hypothesis will not be strengthened by reproducing an experiment. New and different experiments are needed for that.

5. Experiments as tasks

Proposes that tasks and problem-solving methods could be used as a tool for analysing reproducibility. Use the terms task and problem-solving method as specified by Öztürk et al. [49].
Tasks are characterized by their goal, input, output and a reference to the problem solving methods that complete them while problem-solving methods are characterized by their

input, output, the subtasks they decompose the parent task into and the control information specifies i.e. order of execution of tasks.

Classic experiment of comparing hand-written digits using MNIST datasets:
Collect data, Preprocessing, Ready to classify, but first select classification algorithm, algorithm trained, hyperparameters optimizied, tested on test data, analyse test errors, lowest error indicates best performance.
Detailed research protocol could specify all tasks that are to be competed as part of an experiment and how to complete them by describing the more detailed steps. Amount of details that are needed to describe everything to conduct a relatively complex experiment is huge, not necessarily possible in writing. For example, according to Tian et al. [51], who tried to reproduce the results achieved by AlphaGo [52] and Alpha Zero [45], several details were missing from the first paper that hindered them in achieving their goal. Sometimes a paper contains pointers to possible implementations of how a sub-task of the experiment could be implemented, but not necessarily exactly how it was implemented. Differences in implementation could lead to different results. For example, Henderson et al. [43] found that the result varied significantly for different implementations of the same algorithm. As a task can be achieved by different problem-solving methods, which problem-solving method exactly is selected to complete a task could affect the result. For example, the problem-solving method used for selecting which algorithms to compare could affect the result. Are the algorithms chosen based on whether they are the state-of-the-art or because they are easiest for the researchers conducting the study to implement? Also, the order that tasks are executed in might affect the result, so the control information of a problem-solving method is important as well. It is not obvious that the result would be the same if the order of the pre-processing sub-tasks identify centre of mass and translate to centre are switched in the MNIST example.

6. Reproducibility

In this section, reproducibility is defined. No distinction is made between reproducibility and replication, reflecting how Schmidt [24], Nosek & Lakens [33], Goodman et al. [34] and Gundersen & Kjensmo [36] did not distinguish between the two concepts. Reproducibility is defined as follows:

Definition.

“Reproducibility is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.”

Reprodcibility requires that another, independent team of investigators have to conduct the same experiment. Contrasted with repeatability, which is the ability of the same investigators to produce the same result when repeating an experiment. Also, an experiment is reproduced if the interpretation of the analysis leads to the same conclusions; the hypothesis that was supported by the original experiment is still supported after the independent investigators have conducted the same experiment. This means that results can be reproduced even when the outcome of the reproducibility experiment differs from the outcome of the original experiment, as long as the analysis can be interpreted in the same way and lead to the same conclusions. Similarly, the analysis can also be different as long as the interpretation of the analysis leads to the same conclusion. From

this it follows that there are three degrees of reproducibility:

Outcome reproducible:
The outcome of the reproducibility experiment is the same as the outcome produced by the original experiment. When the outcome is the same, the same analysis and interpretation can be made, which lead to the same result and hecne the hypothesis is supported by both experiments. The experiment is outcome reproducible.

Analysis reproducible:
The outcome of the reproducibility experiment does not have to be the same as the outcome produced by the original experiment, but as long as the same analysis can be made and it leads to the same interpretation, the experiment is analysis reproducible.

Interpretation reproducible:
Neither the outcome nor the analysis need to be the same as long as the interpretation of the analysis leads to the same conclusion. In this case, the experiment is interpretation reproducible.

If the outcome is not shared by the original investigators, independent researchers cannot compare their outcome with the outcome of the original experiment, and thus outcome reproducible is not achievable in such cases. In cases where researchers do not agree on the interpretation of an analysis, the results are not conclusive and the hypothesis is neither supported nor refuted. There is no new evidence to strengthen nor to weaken the belief that the hypothesis was formulated to test. New experiments must be conducted in order to decide whether the hypothesis should be supported or refuted.

Reproducibility experiment can only be conducted if some documentation of the original experiment is shared. The degree to which independent investigators can conduct the same experiment as the original investigators is dependent on how well the original investigators documented the experiment and how much of the documentation is shared with independent investigators.
Documentation is not restricted to textual descriptions. Gundersen & Kjensmo [36] distinguish between three types of documentation: text, code and data

Term degree refers to to the degree of closeness of the exact result that is reproduced while reproducibility type refers to which type of documentation is used for redoing the experiment.

Four reproducibility types:
R1 Description:
Only a textual description of the experiment is used as a reference for the reproducibility experiment. The text could describe the experimental procedure, the target system and its behaviour, the implementation of the target system for example in the form of pseudo code, the data collection procedure, the data, the outcome and the analysis and so on.

R2 Code:
Code and the textual description of the experiment are used as reference when conducting the reproducibility experiment. The code could cover the target system, the workflow, data pre-processing, experiment configurations, visualization and analyses.

R3 Data:
Data and the textual description of the experiment are used as reference for conducting the experiment. The data could include training, validation and test sets as well as the outcome produced in the experiment.

R4 Experiment:
The complete documentation of the experiment including data and code in addition to the textual description as shared by the original investigators are used as reference for the reproducibility experiment.

Figure 5 which documentation is required for the four reproducibility types.
The more documentation that is shared means a higher degree of transparency and the easy it gets for independent investigators to reproduce the results. However, the more variability in the conditions of the experiment, the more certain one can be that the conclusion is correct. An experiment executed with the same code on the same data validates the hypothesis for specific code executed on specific data while a new implementation executed on different data generalizes the result to be independent of both code and data. So, the less documentation that the original investigators share, the greater the generality of the results. See Gundersen et al. [53] for a discussion on the relationship between transparency and generality of results.

It is easier to reimplement a part of the experiment in order to introduce some difference, than having to implement the whole experiment.
If full transparency is provided, then one task, action or problem-solving method can be reimplemented or changed to investigate whether the conclusion is still valid.

Running the same experiment in the same environment is not considered to introduce any variability and is hence interpreted as repeatability. Also, other investigators executing a program that encodes the same experiment on the same hardware means for computational sciences, as mentioned before, only that someone else pushes a button. This has little if any value. However, reproducing an experiment by executing the same code and data on a different computer with different ancillary software introduces variability.

7. Conclusion

The literature on reproducibility agrees to a large degree that the same experiment is conducted as long as the same experimental method is followed. However, following the same method is not enough. An experiment is not reproduced unless the results are the same
Reproducibility requires variation; a reproducibility experiment requires both sameness and difference. The sameness must at least be the same experimental methodology.

The analysis resulted in a new definition of reproducibility and three degrees of reproducibility. Furthermore, four reproducibility types that are specified based on which types of documentation that are shared by the original investigators were proposed. The more documentation that is shared, the easier it is for independent researchers to reproduce the results. The easier it is to reproduce results the faster knowledge can be discovered. The understanding of reproducibility described here emphasizes that fast-paced and steady scientific progress requires transparency and openness.