Received: 2 July 2024 | Revised: 19 December 2024 | Accepted: 25 February 2025

DOI: 10.1002/aaai.70002

ARTICLE

# Reproducibility in machine-learning-based research: Overview, barriers, and drivers

**Harald Semmelrock**[1] | **Tony Ross-Hellauer**[2] | **Simone Kopeinik**[2] | **Dieter Theiler**[2] | **Armin Haberl**[2] | **Stefan Thalmann**[3] | **Dominik Kowald**[1,2]

[1]Graz University of Technology, Graz, Austria

[2]Know Center Research GmbH, Graz, Austria

[3]University of Graz, Graz, Austria

**Correspondence**
Dominik Kowald, Graz University of Technology, Graz, Austria.
Email: dkowald@know-center.at

**Funding information**
Horizon Europe project TIER2, Grant/Award Number: GA: 101094817; FFG COMET program

**Abstract**

Many research fields are currently reckoning with issues of poor levels of reproducibility. Some label it a "crisis," and research employing or building machine learning (ML) models is no exception. Issues including lack of transparency, data or code, poor adherence to standards, and the sensitivity of ML training conditions mean that many papers are not even reproducible in principle. Where they are, though, reproducibility experiments have found worryingly low degrees of similarity with original results. Despite previous appeals from ML researchers on this topic and various initiatives from conference reproducibility tracks to the ACM's new Emerging Interest Group on Reproducibility and Replicability, we contend that the general community continues to take this issue too lightly. Poor reproducibility threatens trust in and integrity of research results. Therefore, in this article, we lay out a new perspective on the key barriers and drivers (both procedural and technical) to increased reproducibility at various levels (methods, code, data, and experiments). We then map the drivers to the barriers to give concrete advice for strategies for researchers to mitigate reproducibility issues in their own work, to lay out key areas where further research is needed in specific areas, and to further ignite discussion on the threat presented by these urgent issues.

## INTRODUCTION

Trustworthy AI requires reproducibility (Kowald et al. 2024). Unreliable results risk hindering scientific progress by wasting resources, reducing trust, slowing discovery, and undermining the foundation for future research (Gundersen and Kjensmo 2018; Munafò et al. 2017). However, many scientific fields currently face crucial questions over the reproducibility of research findings (Baker 2016). Concerns of a "reproducibility crisis" have been most prominently raised in biomedical (Errington et al. 2021; Freedman, Cockburn, and Simcoe 2015; Ioannidis 2005; Iqbal et al. 2016) and social (Collaboration 2015; Hardwicke et al. 2020; Nosek et al. 2022) sciences, but research employing artificial intelligence (AI) in general, and machine learning (ML) in particular, is also under scrutiny Hutson (2018a). ML is becoming ever more deeply integrated into research methods, not just in computer science but across disciplines (Dwivedi et al. 2021; Ooi et al. 2023). Indeed, recipients of the 2024 Nobel Prizes for both chemistry and physics included ML researchers. Hence, issues regarding the reproducibility of ML raise urgent concerns about the reliability and validity of findings not only for computer scientists but for large swathes of cutting-edge scientific research across disciplines.

The causes of poor reproducibility can be technical, methodological, or cultural. Viewed from a high level, some causes, such as lack of sharing data and code, lack of or poor adherence to standards, suboptimal research design, or poor incentives, may be seen as common to many domains. Apart from the common challenges faced by other disciplines, the use of ML introduces unique obstacles for reproducibility, including sensitivity to ML training conditions, sources of randomness (Raste et al. 2022), inherent nondeterminism, costs (economic and environmental) of computational resources, and the increasing use of Automated-ML (AutoML) tools (Haberl and Thalmann 2025; Koenigstorfer et al. 2024). Among the methodological and cultural aspects, specificities of ML research, like "data leakage," as well as ML-specific issues regarding unobserved bias, lack of transparency, selective reporting of findings, and publishing cultures, each play a role as well. Indeed, this cultural aspect must not be underestimated. The culture of "publish or perish" pervades academia, pushing researchers to publish as many papers in the highest-ranked or most prestigious journals or conferences as possible (Pontika et al. 2022). In turn, this culture distorts incentives towards corner-cutting, giving rise to so-called "questionable research practices" and "design, analytic, or reporting practices that have been questioned because of the potential for the practice to be employed with the purpose of presenting biased evidence in favor of an assertion" (Banks et al. 2016).

This paper aims to provide a detailed overview of reproducibility and its associated barriers and drivers in ML. This is urgently needed since, despite previous appeals from ML researchers on this topic, various initiatives from conference reproducibility tracks to the ACM's new Emerging Interest Group on Reproducibility and Replicability, and an expanding literature on the topic (Albertoni et al. 2023; Gundersen 2021; Gundersen, Coakley, et al. 2023; Gundersen, Shamsaliei, and Isdahl 2022; Heil et al. 2021; Kapoor and Narayanan 2023; McDermott et al. 2021; Semmelrock et al. 2023), we contend that the general community continues to take this issue too lightly. In addition, despite the growing literature, no such comprehensive overview exists. For example, in Gundersen, Coakley, et al. (2023), the authors identify and categorize sources of irreproducibility in ML and how these sources affect conclusions drawn from ML experiments. However, this study does not investigate the drivers to address these sources of irreproducibility. Thus, our paper provides a contextual categorization of the barriers and drivers to the four types of ML reproducibility (description, code, data, and experiment) proposed by Gundersen (2021), with specific reference to research in both computer science and biomedical fields. We also propose a drivers–barriers matrix to summarize and visualize the results of the discussion. Such an analysis stands to clarify the current state regarding ML reproducibility, to give concrete advice for strategies for researchers to mitigate reproducibility issues in their own work, to lay out key areas where further research is needed in specific areas, and to further ignite discussion on the threat presented by these urgent issues. The paper is structured as follows: In "Defining Reproducibility" section, we clarify terms and working definitions. We then analyze the barriers to increased reproducibility of ML-driven research ("Barriers in ML Reproducibility" section), and next, the drivers that support ML reproducibility, including different tools, practices, and interventions ("Drivers for ML Reproducibility" section). Here, we also provide a comparison of the strengths and potential limitations of these drivers. Finally, we map the barriers to the drivers to help determine the feasibility of various options for enhancing ML reproducibility ("Mapping Drivers to Barriers" section). We close the paper with a conclusion and an outlook into our future research in "Conclusion" section.

## DEFINING REPRODUCIBILITY

The concept of reproducibility can have different interpretations across various research fields and even within the same field (Fidler and Wilcox 2021). To avoid confusion, we first specify our terms, broadly defining reproducibility and then further categorizing it into various types and degrees. The first distinction comes from Goodman, Fanelli, and Ioannidis (2016), who specify a fundamental division between whether we (i) mean reproducible in principle (termed "methods" reproducibility) due to sufficient description/sharing of methodologies, materials, and so forth, or (ii) whether results/conclusions actually prove to be reproducible when experiments or analyses are redone. In the second category, they distinguish "results" and "inferential" reproducibility, depending on whether the analyses or inferences to broader conclusions are reproduced.

Within ML research, widely accepted definitions that build further on these key distinctions have been proposed by Gundersen (2021) and Gundersen, Shamsaliei, and Isdahl (2022). We follow and build upon these latter definitions, and hence, here outline them at some length. Gundersen, Coakley, et al. (2023) define reproducibility in general as "the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators." Relating to point (ii) of Goodman et al.'s schema, Gundersen (2021) and Gundersen, Shamsaliei, and Isdahl (2022) further distinguish the targets of reproducibility, that is, how closely an experiment can be reproduced:

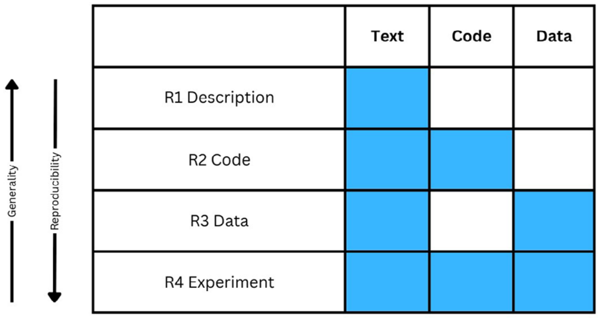
– **Outcome reproducibility** requires the reproduced experiment to have the same or adequately similar outcome as the original experiment. Due to this, the same analysis and interpretation follow, and the hypothesis is either supported or rejected by both experiments.

– **Analysis reproducibility** does not require the reproduced experiment to have the same/similar outcome; however, if the same/similar analysis and, therefore, also interpretation can be made, an experiment is analysis reproducible.

– **Interpretation reproducibility** does not require the reproduced experiment to have the same/similar outcome nor analysis but requires the interpretation to be the same as the original one.

This categorization aims to overcome the problem of ambiguity when making specific claims about the reproducibility of an experiment. Often in literature, authors write about reproducing the same "results" of an experiment. It is not apparent, however, in which cases they mean to achieve the same computational outcome, that is, outputs of the algorithms, or whether they mean to reach the same analysis or interpretation. Therefore, achieving interpretation reproducibility is a more general and often less stringent requirement than achieving outcome reproducibility. This categorization is not specific to ML, but is generally applicable to any research field that conducts data analysis and interpretation.

In addition, relating to point (i) of the Goodman et al. schema ("methods" reproducibility), Gundersen (2021) specifies reproducibility types to which methods can be made transparent through description or sharing. These four types are defined as *R1 Description*, *R2 Code*, *R3 Data*, and *R4 Experiment*. The lower the level of reproducibility, the less shared information is shared, making the study more difficult to reproduce. As an example, in general, all published research experiments are accompanied by a textual description of the experiment. If this textual description is the only information shared by the authors, the research is categorized as *R1 Description*, which, according to this scheme, is the minimal kind of reproducibility. In contrast, if all three building blocks, that is, text, code, and data, are shared, the experiment can be categorized as *R4 Experiment*, the most expansive kind of reproducibility. Furthermore, the distinction between *R2 Code* and *R3 Data* is defined by whether the textual description is accompanied by either code or data, respectively. Also, the different types of ML reproducibility exhibit an interplay between generalizability and reproducibility. *R1 Description* leads to strong generalizability but to weak reproducibility, while *R4 Experiment* leads to stronger reproducibility but weaker generalizability. What this means in sum is that rerunning the same code on the same data using the same description will make it likelier to obtain the same results, but those results might still be wrong due to errors or biases in any of those elements. On the other hand, building code from scratch and using alternative datasets for analysis will show that the techniques give similar results across contexts and, hence, higher levels of confidence in the generalizability of findings. These relationships are illustrated in Figure 1. In what follows, unless stated otherwise, we use the definitions proposed by Gundersen and colleagues (Gundersen 2021).



| | Text | Code | Data |
|---|---|---|---|
| R1 Description | 🟦 | | |
| R2 Code | 🟦 | 🟦 | |
| R3 Data | 🟦 | | 🟦 |
| R4 Experiment | 🟦 | 🟦 | 🟦 |

**FIGURE 1** Types of reproducibility. Adapted from Gundersen (2021).

## BARRIERS IN ML REPRODUCIBILITY

Next, we discuss nine barriers to ML reproducibility, categorized into the four types of reproducibility mentioned beforehand. Where applicable, we give examples within the research fields of biomedical science and computer science.

## R1 description

### Completeness and quality of reporting

Research often lacks reproducibility due to missing or vague methodological details. Mainly, there are three issues in this regard, which often hinder the reproduction of study results Pineau et al. (2021):

1. The ML model or training procedure is either incorrectly specified or under-specified. Reports should give clear details on all steps of the procedure, even if data and code are not shared. This includes details about

which ML models are used, as well as details on the training data and data preprocessing.
2. The evaluation metrics used to report results are not properly specified. There are many metrics which can be used to evaluate ML models, for example, accuracy, receiver operator curve (ROC), or mean squared error (MSE). It is important to define these metrics and also explain why they were used.
3. Often results are selectively reported, for example, researchers may only provide results for the best test run out of many test runs, instead of properly assessing and reporting average values and variances (Belz et al. 2021).

Generally, it is important for studies to use a robust methodology and provide detailed reports so that other researchers can verify results and understand how analyses were conducted. While ML models have proven highly effective in biomedical fields, studies often fall short in providing comprehensive and high-quality reporting. For example, in studies on predicting cardiometabolic risk from dietary patterns (Panaretos et al. 2018) or supporting the clinical management of diabetes (Rahimi et al. 2022), ML models were observed to be very promising. While this further enhances the promise of applying ML models for various clinical prediction tasks, there is a clear need for thorough reporting and validation of these models to allow for their integration into routine clinical care (Rahimi et al. 2022). This is also true for the application of ML models for cancer imaging, where ML models often surpass radiologists in performance, but publications on these models lack the documentation details needed to reproduce the results (Provenzano et al. 2021).

### Spin practices and publication bias

Another issue commonly observed in ML-based research that negatively affects reproducibility is "spin." It refers to the misuse of language to "intentionally or unintentionally affect the interpretation of study findings." It is also understood as an inconsistency between the study results and the conclusions, in the sense that results are over-generalized or the claimed conclusions are not supported by the scientific method. This has been shown to impact both the interpretations and decision-making by readers (Navarro et al. 2023). In ML-based biomedical research, the most common practice of spin includes recommending models for various applications without providing external validation in the same study. More concretely, the recommendation to use a model either in a clinical setting or for a different population is only validated in approximately 15% of the cases. Other observed instances of spin are invalid comparisons of results to previous studies and the use of leading words and strong statements to make the results sound more significant (Navarro et al. 2023). The prevalence of spin can perhaps be attributed in large part to the academic culture of "publish or perish" and its associated reward systems. Valuing and rewarding perceived novelty and potential impact over basic rigor and responsible reporting can lead researchers to inflate claims in hopes of acceptance in the most prestigious venues. It can also skew the literature in other ways, leading to so-called "publication bias" (Saidi, Dasarathy, and Berisha 2024). Here, in addition to spin and the aforementioned selective reporting of (usually positive) results, the role of the peer review system is also in question, given known biases on the part of reviewers that can lead to preferential treatment for researchers from specific regions, institutions, or demographics, or for certain types of research (Lee et al. 2013). Other kinds of bias, such as "complexity bias" (tendency to prefer complicated over simple results and explanations), are also known to influence acceptance decisions (Trosten 2023).

Finally, we note how a core aspect of computer science culture may exacerbate these issues, namely the importance of conference rankings. Within computer science, the prime mode of publication is within conference proceedings, with conferences ranked A* to C by bodies such as ICORE.[1] To date, surprisingly little has been said regarding the analogous nature of such conference rankings to other metrics like the Journal Impact Factor, where a rich literature exists critiquing its worth as an indicator of quality or impact for individual pieces of work (Lariviere and Sugimoto 2019). Although elaboration at length on this issue is outside the scope of this article, we suggest this as an underexplored topic for future research. Such research can build on a rich evidence-base exploring downstream ill-effects of badly designed or misused metrics, including distortion of incentives (Smaldino and McElreath 2016), inviting manipulation or gaming (Biagioli and Lippman 2020), goal displacement and task reduction (Rijcke et al. 2016), and influencing core academic values (Burrows 2012).

## R2 code

### Limited access to code

Published ML research is often not accompanied by available data and code. Only one-third of researchers share data, and even fewer share their source code (Hutson 2018). This can be attributed to several factors, such as the increasing pressure on researchers to publish quickly, which often leaves insufficient time to refine code and decreases the willingness to share it. Additionally, concerns about intellectual property may further discourage researchers from releasing their code. According to Gundersen and Kjensmo (2018), sharing ML code to facilitate reproducibility requires publishing seven pieces of information: hypothesis, prediction method, source code, hardware specifications, software dependencies, experiment setup, and experiment source code. Unfortunately, current research rarely meets these requirements, leading to reproducibility issues due to different software versions, hyperparameter settings, or hardware differences (Hong 2013; Belz 2021). For example, research in recommender systems has struggled with the lack of shared code, significantly contributing to a lack of reproducibility. Even when code is shared, it is often incomplete, poorly documented, or limited to pseudocode or skeletal implementations rather than fully executable code (Cremonesi and Jannach 2021). To address this, shared code should encompass comprehensive documentation, including scripts for data preprocessing, hyperparameter tuning, management of random seeds, and implementations for comparisons against baseline models.

## R3 data

### Limited access to data

The main reproducibility barrier associated with *R3 Data* is that data is simply not shared or made publicly available most of the time (Hutson 2018b). A review in biomedical research, specifically radiomics, investigated 257 recent ML publications and found that only 16 of them shared data or used publicly accessible datasets (D'Antonoli et al. 2024). This could be due to privacy concerns or a lack of incentives and motivation. Moreover, many benchmark and training datasets encounter challenges related to copyright, licensing, and longevity. These datasets may also raise ethical concerns, such as the unintentional inclusion of privacy-sensitive or harmful content, making it difficult to share the data for ML model training (Paullada et al. 2021). Similar to the sharing of source code, sharing only the datasets is insufficient without sufficiently detailed levels of documentation. For proper use, it is also important to share specific splits, that is, the training dataset, validation dataset, and test dataset (Gundersen and Kjensmo 2018). Furthermore, data sharing needs to be accompanied by documentation specifying details about the provenance and preprocessing of data. Significant recent initiatives aiming at improvement here are as follows: *Croissant*, a unified format for ML datasets that integrates metadata, resource descriptions, data structure, and default ML semantics in a single file[2]; and *MLCommons*, which is working towards open benchmarks and public data[3]. In addition, RO-Crate is a standard for packaging data and other research objects together with data to enable reuse and reproducibility.[4]

Standardizing and mainstreaming these practices is essential for validation and checking of methods. We next discuss two common methodological errors related to data, data leakage, and bias, and their impact on reproducibility.

### Data leakage

In practice, methodological issues such as data leakage (also referred to as target leakage) often hinder the reproducibility of ML-based research (Kapoor and Narayanan 2023). This is due to the growing number of nonexperts employing ML across different research fields (Gibney 2022), which is fueled by the ease of application of ML libraries and no-code off-the-shelf AI tools. In essence, data leakage happens when data on which the ML model should not be trained leaks into the training process. Data leakage can be categorized into three subcategories (Kapoor and Narayanan 2023):

1. No clean train/test split. Here, four variants are possible: (1) training data and test data are not split at all, (2) test data are also used to select the best features from the training data (feature selection), (3) test data are also used for imputation of missing data during preprocessing, and (4) duplicates occur in the training and test data.
2. Use of nonlegitimate data. For example, when the use of antihypertensive drugs is used as a feature to predict hypertension. This data is nonlegitimate, since it would not be available in a real-world scenario (people are prescribed those drugs because of a hypertension diagnosis) and would be useless for predicting hypertension in undiagnosed patients.
3. Test set is not drawn from the distribution of scientific interest. There are three possible variants as follows: (i) temporal leakage, which is problematic for ML models that attempt to predict future outcomes, that is, when some training samples have a later timestamp than samples available in the test set; (ii) the training and test data are not independent of each other, for example, there should not be samples in the training and test data that are drawn from the same person; and (iii) the test set is not chosen selectively, for instance, if the model is solely evaluated on data, on which it performs well.

### Bias

Bias in ML refers to the error introduced by approximating a real-world problem, which may be complex, by a simplified model, often leading to systematic deviations from the real world. Furthermore, biases can arise when the model contains imbalances or reflects existing societal biases (Mehrabi et al. 2021). ML models, which are subject to bias, are prone to generalization issues, and are therefore potentially problematic for ML reproducibility. There are eight kinds of bias that can arise during the data handling phase of ML development (Rouzrokh et al. 2022): (i) selection bias—using data not being representative of the target group; (ii) exclusion bias—excluding particular data samples based on the belief that they are unimportant; (iii) measurement bias—favoring certain measurement results; (iv) recall bias—labeling similar data samples differently; (v) survey bias—introducing data issues stemming from data collection surveys; (vi) confirmation bias—favoring information, which confirms previous beliefs; (vii) prejudice bias—including human-related prejudices in training data; and (viii) algorithmic bias—replicating or amplifying biases by the inner workings of the algorithms. Bias, such as selection bias, often leads to the issue of validity shrinkage in biomedical science research (Ivanescu et al. 2016). For example, in obesity and nutritional research, ML is used to predict obesity, heart rate, or the risk of a heart attack based on data from an individual. Here, validity shrinkage refers to the issue that a predictive model trained on a subset of data will most probably not perform well on new samples. The difference between predictive performance on known data and new data, however, is most often not accounted for in nutritional science, and therefore also leads to performance claims that cannot be reproduced (Ivanescu et al. 2016).

## R4 experiment

### Inherent nondeterminism

Inherent nondeterminism in ML models means that results can vary between test runs, even with identical code, data, and hyperparameters. This variation arises from sources of randomness during training, such as, for example, random parameter initialization, stochastic optimization, and random data subsampling (e.g., in k-fold cross-validation) and the complex interactions between them, which are fundamental characteristics of most ML workflows (Leventi-Peetz and Östreich 2022; Raste et al. 2022). Neural networks are especially known for their inherent nondeterminism, leading to varied computational outcomes in multiple reruns due to increased sources of randomness during training (Ahmed and Lofstead 2022). In some cases, inherent nondeterminism can cause such large variations that reruns not only yield slightly different outcomes, but also lead to significant fluctuations in the performance of an ML model (Ahmed and Lofstead 2022) or to varying conclusions in ML model comparisons (Gundersen, Shamsaliei, et al. 2023). This issue is exacerbated when other sources of variation, such as different hyperparameters, are introduced to the ML model. In such cases, the impact can be magnified, and it is often observed that minor changes in hyperparameters can result in significant performance loss (Belz et al. 2021). Reviews of reproducibility in both NLP research (Belz et al. 2021) and biomedical research (Ahmed, Tchoua, and Lofstead 2022) highlight these core issues with nondeterminism that are exemplary for ML research because they reflect challenges that are prevalent across ML-based research and serve as representative examples of the broader reproducibility crisis faced by the ML community. Simply rerunning the original code of an experiment during a reproduction leads to large variances of results and different computational outcomes on each run. Reinforcement learning, a subfield of ML, is particularly susceptible to these reproducibility issues, partially because of additional sources of nondeterminism, such as the reinforcement learning environment or policy (Nagarajan, Warnell, and Stone 2018).

### Environmental differences

Various studies have demonstrated that hardware differences, such as different GPUs or CPUs, and compiler settings can lead to different computational outcomes (Hong et al. 2013). Additionally, a comparison between the same ML algorithm with fixed random seeds executed using PyTorch[5] and TensorFlow[6] resulted in different performances (Pouchard, Lin, and Van Dam 2020). Furthermore, even different versions of the same framework can lead to different performance results (Shahriari, Ramler, and Fischer 2022). A comparison of the results of experiments performed on different hosted ML platforms also found that out-of-the-box reproducibility is not guaranteed there (Gundersen, Shamsaliei, and Isdahl 2022). Another important factor is the use of GPUs, which can increase randomness compared to the use of CPUs. This is due to parallel optimization and the use of optimizers in ML frameworks, such as PyTorch and TensorFlow. As a result, some researchers have resorted to solely using CPUs for executing their experiments. However, this comes at the expense of runtime-efficiency (Alahmari et al. 2020).

### Limited access to computational resources

The barrier to ML reproducibility posed by limited access to computational resources has recently become evident in the case of transformer-based Large Language Models (LLMs) (Beam, Manrai, and Ghassemi 2020). These transformer architectures need a vast amount of data and computational resources, to which most researchers have limited access. Estimates have calculated the costs to reproduce one model to be around $1 million to $3.2 million (Beam, Manrai, and Ghassemi 2020). Another study found that the needed computational resources are one of the most significant factors impacting reproducibility (Raff 2019). Especially ML models, which require computational clusters for training and optimization, are notably hard to reproduce.

## DRIVERS FOR ML REPRODUCIBILITY

In this section, we discuss drivers for ML reproducibility, which we subdivide into (i) technology-based drivers, (ii) procedural drivers, and (iii) drivers related to awareness and education. For every driver, we also provide case studies or examples from the literature illustrating the effectiveness of the driver for ML reproducibility.

## Technology-based drivers

### Hosting services

Utilizing hosting services offers an efficient way to share code, data, and ML model parameter settings, thus supporting the reproducibility of ML-driven research (Tatman, VanderPlas, and Dane 2018). Examples of hosting services include the runtime environments of ML platforms. If the original author runs the ML experiment in such a runtime environment, for example, Kaggle Notebooks[7], Google Colab[8], or CodaLab[9], researchers attempting to reproduce the results should be able to execute the experiment within the same environment. The main advantage of using a hosting service is that the provider takes care of the logistics of code hosting and distribution. However, the main drawbacks are the limits on data size and computational resources. Since these hosting services are run in the cloud, there are restrictions on how many resources a single user can utilize. The limit on resources varies between different hosting services and is limited by users' available funds and sometimes subscription levels. Because of these limits, hosting services may not be suitable for all research purposes, especially considering the compute-intensive nature of novel ML models, such as LLMs.

As said above, the degree to which such services offer out-of-the-box reproducibility remains highly questionable (Gundersen, Shamsaliei, and Isdahl 2022). Nevertheless, some hosting services *have* effectively been used to create end-to-end reproducible AI pipelines, especially in conjunction with standardized datasets such as the National Cancer Institute Imaging Data Commons (Fedorov et al. 2023). The effectiveness of hosting services for ML reproducibility has been shown for research in radiology (Bontempi et al. 2024) and similar reproducibility experiments have been successfully conducted in pathology research (Schacherer et al. 2023). As also noted in these experiments, it is important that results are reported with a quantification of the variance across different test runs, since effects of randomness could still be prevalent.

### Virtualization

Reproducing the environment and setup of any ML experiment requires the consideration of existing dependencies and software versions, and is usually a complex task itself. Virtualization can simplify this process by bundling the essential components of ML models and experiments, such as the dependencies and code, into a single package for sharing with other researchers. Thus, if the authors of a paper build the experiment in a virtual environment, issues associated with setup reproduction can be greatly reduced. However, the adoption of virtualization by researchers depends on its user-friendliness and the effort of integration into their current workflows (Boettiger 2015). Concerns about virtualization include its limitations in allowing researchers to build upon them in a scalable manner. Traditional virtual machines (VMs) emulate an entire operating system for setting up and running experiments. The use of containerization software like Docker[10] has become more popular in recent years. Containers are more lightweight and flexible than VMs, making it easier to adapt environments for follow-up studies (Boettiger 2015). There are also designated platforms for computational research, such as Code Ocean[11], that offer virtualization via so-called reproducible capsules. Their focus, in particular, is to simplify the virtualization process and allow researchers to focus on the research itself rather than the standardization of environments (Clyburne-Sherin, Fei, and Green 2019). Additionally, there are many other tools, such as ReproZip (Chirigati et al. 2016), one of the recommended tools by the SIGMOD Reproducibility Availability and Reproducibility Initiative[12] to streamline reproducibility, and DetTrace (Navarro Leija et al. 2020), which aims to ensure completely deterministic computations.

The use of containers is rapidly gaining in popularity across many research fields, for example, neuroscience and genomics (Moreau, Wiebels, and Boettiger 2023). Notably, the platform Code Ocean has been integrated into the peer reviewing process by Nature journals to support the submission process of experiments (Editorial 2022). This widespread adoption highlights the suitability of containers to enhance experiment sharing and improve reproducibility. Furthermore, a case study has compared 10 different containerization-based approaches for reproducibility (Choi et al. 2023). The strengths and weaknesses of each approach were analyzed, with results demonstrating the suitability of containers to enhance reproducibility by encapsulating the computational environment and to decrease the effort for publishing reproducible ML-based experiments.

## Managing sources of randomness

Many different sources of randomness during ML training lead to the ifor example, via random number seeds, deterministic algorithms, or other methods, could therefore greatly increase reproducibility. Fixed random number seeds should be used and published to make ML experiments more reproducible and control a number of sources of inherent nondeterminism. A seed is a first value used to initialize the pseudo-random number generator. When the same seed is used, the sequence of pseudo-random numbers generated is deterministic, meaning it will be the same every time the code is run. Experiments have shown that fixing random seeds can effectively ensure reproducible results when algorithms are not being executed in parallel on GPUs (Ahmed and Lofstead 2022). Additionally, one case study has shown that achieving reproducibility for GPU-trained neural networks (Chen et al. 2022) is possible through a method known as patching. Patching aims to replace nondeterministic operations with deterministic ones. In the case study, a systematic patching approach was successful in achieving reproducible image classification results for six different neural network. However, this process also leads to higher computational costs and a time overhead, which was also analyzed in the study. Additionally, ML models should be benchmarked and evaluated with multiple random number seeds, such that the variance can be reported and inform about the true performance of an ML model (Raste et al. 2022). Similarly, the use of uncertainty-aware quantification metrics to evaluate ML models can also help increase reproducibility (Pouchard, Lin, and Van Dam 2020).

Additionally, to counteract inherent nondeterminism in reinforcement learning and achieve reproducible evaluations, there exist frameworks, such as Gym-Ignition (Brockman et al. 2016; Ferigo et al. 2020), rl_reach (Aumjaud et al. 2021), or MinAtar (Young and Tian 2019), which act as standardized benchmarking environments. Within them, different algorithms designed for the same task can be evaluated and compared against each other in a common environment. Some frameworks counteract the effects of inherent nondeterminism by automatically controlling random seeds and evaluating algorithms over a number of runs. Finally, it is an ongoing field of research to implement fully deterministic reinforcement learning algorithms (Nagarajan, Warnell, and Stone 2018) and make use of them within such frameworks (Tassa et al. 2018).

## Privacy-preserving technologies

Privacy-preserving technologies support reproducibility, as they enable the collaborative training of ML models without sharing private or sensitive data. The main benefit of this is that ML models can make use of larger and more diverse data, thus helping to decrease bias and leading to more reproducible ML models. The main aim of Privacy-Preserving Machine Learning (PPML) is to facilitate the use of privacy-sensitive data to create better ML models, and, to allow data owners to collaboratively train ML models on private data. In that regard, PPML has several requirements. First, protecting the confidentiality of the training data. Second, preventing the leakage of sensitive information from ML model parameters and outputs, that is, to hinder the re-identification of individuals. Third, achieving the listed security and privacy aspects while still preserving the utility of the ML model (Xu, Baracaldo, and Joshi 2021). To achieve this, a number of different techniques are being used and developed, mainly Differential Privacy (DP), Homomorphic Encryption (HE), Secure Multi-Party Computation (SMPC), and Federated Learning (FL) (De Cristofaro 2021). These techniques are implemented in software libraries such as TensorFlow Privacy, PySyft, ML Privacy Meter, CryptFlow, or Crypten (Aslanyan and Vasilikos 2020). Furthermore, data can be made anonymous by removing identifiable personal information. However, if too much data are removed, the ML models may perform poorly. If not enough data are removed, it may still be possible to re-identify individuals by combining many different nonunique features (Xu, Baracaldo, and Joshi 2021).

An alternative approach to PPML is to generate synthetic data that captures the same information as the original data. A robust technique for creating such datasets can produce readily available datasets of nearly any size, as demonstrated in biomedical fields. This approach has led to the development of Synthea, a software package designed to generate synthetic patient data and electronic health care records (Walonoski et al. 2018). It is, however, important to mention that there is still a gap in efficiency between theoretical advancements and real-world applications when using PPML techniques. To this end, one study conducting reproductions of 26 state-of-the-art applications of PPML has highlighted the challenges of balancing computational efficiency, privacy guarantees, and model utility, while emphasizing the need for improved reproducibility, open-source availability, and practical scalability Khan et al. (2024).

### Tools and platforms

There are many tools and platforms that assist in the implementation and management of ML models and ML-based applications. A recent study has evaluated 19 ML tools to gain insights into their concepts constituting reproducibility support (Quaranta, Calefato, and Lanubile 2022). As a result, five main pillars of ML reproducibility in tools and platforms were identified: (i) code versioning, (ii) data access, (iii) data versioning, (iv) experiment logging, and (v) pipeline creation. Most of these pillars are associated with managing and keeping track of different artifacts created during phases of the ML lifecycle (i.e., design, development, and deployment) as for instance, datasets, labels, code, logs, environment dependencies, random number seeds, or hyperparameters (Schlegel and Sattler 2023). Each of these artifacts influences the final results of the ML model. Consequently, most tools aim to collect, store, and manage these artifacts, ensuring researchers can access and use them during reproduction attempts. Notable are also various tools and platforms for experiment tracking (Schlegel and Sattler 2023), such as:

– **DVC**[13]: A version control system for ML projects with a command-line interface similar to Git[14]. It integrates with Git, supports cloud storage, and handles large versioning of datasets. DVC ensures full code and data provenance by enabling experiment tracking.

– **MLflow**[15]: An open-source tool for supporting ML experiment tracking, ML model deployment, and centralized model storage. Additionally, it provides an easy-to-use Web dashboard.

– **RO-Crate**[16]: A specification, implemented by a number of tools, aimed at aggregating and describing research data and metadata Soiland-Reyes et al. (2022). Although not specifically designed for ML, RO-Crate can aggregate and represent any resource, making it applicable for managing ML artifacts as well.

– **dToolAI** (Hartley and Olsson 2020): Collects and packages ML models together with supplemental information, such as hyperparameter settings, appropriate metadata, and persistent URIs for model training data. In contrast to the other tools, dToolAI is specifically tailored towards Deep Learning models.

AutoML platforms, such as H2O Driverless AI[17], Google Cloud AutoML[18], DataRobot[19], are a novel subcategory of ML tools that aim to aid with every aspect of the ML lifecycle, from data aggregation to model deployment. Thus, AutoML tools could facilitate more standardized ML models and also take care of tasks like hyperparameter optimization. It is, however, questionable how practical these tools are for reproducible ML research, since they often hide ML model optimization procedures. Recent assessments of the reproducibility of AutoML tools also came to the conclusion that current platforms cannot provide out-of-the-box-reproducibility (Gundersen, Shamsaliei, and Isdahl 2022; Pletzl et al. 2024). In a qualitative analysis of the reproduction experiments, the latter study did identify areas in which such tools can be enablers for reproducibility, for example, due to their automatic documentation capabilities. However, the authors also identified aspects that need to be addressed, such as the need for simplified tool user interfaces—as many participants were overwhelmed by tool complexity and could not make use of the documentation capabilities—and more built-in reproducibility capabilities, which support the sharing of code and data (Pletzl et al. 2024). Furthermore, some AutoML tools, such as H2O Driverless AI, aim to address problems such as model overfitting. In the case of data leakage, this is done by checking for a strong correlation between a feature and the target and then taking action, for example, warning the user or automatically handling it. This is, however, a very simple solution to the problem and does not address the more complex cases of data leakage that are often present in research, for example, temporal leakage.

## Procedural drivers

### Standardized datasets and evaluation

Due to a lack of shared datasets, many researchers in ML-driven research—most notably in biomedical fields—have to use individually acquired data (McDermott et al. 2021). The collection of such data is a time-consuming task and bears a significant risk of causing reproducibility issues, for example, bias or data leakage. Often, the number of individual participants represented within datasets is not very large and, thus, findings might suffer poor generalizability. Creating shared and standardized datasets can, therefore, (i) save researchers time in acquiring new data, (ii) facilitate the collaborative and independent maintenance and verification of data to minimize methodological errors, and (iii) support transferability and generalizability through the use of multi-institutional data (McDermott et al. 2021). In addition to the standardization of datasets, data cards (Pushkarna, Zaldivar, and Kjartansson 2022) provide a consistent and comparable framework for reporting essential aspects of ML datasets. This includes information, for example, about access restrictions, risks, and limitations associated with the usage of the dataset, or any preprocessing steps, amongst many other contents, which are needed for reproducible ML development.

Another issue is the lack of standardized evaluation methods, which leads to reported performances of ML models often being overly optimistic (Cremonesi and Jannach 2021). To ensure the statistical significance of ML model evaluations, it is crucial to report performance as an aggregate of results obtained from multiple random runs, and with different random number seeds (Colas, Sigaud, and Oudeyer 2018). Furthermore, ML models should, if possible, be tested and evaluated on multiple different datasets (Dror et al. 2017). This underscores the need for standardized evaluation methods, which can be supported by checklists or tools to prevent errors in this critical aspect of ML research. For this, similarly to data cards, model cards (Mitchell et al. 2019) are aimed to standardize the evaluation and reporting of the performance of ML models for a variety of use cases. Model cards should inform users about the possible applications of the ML model and its limitations. In 2020, Google introduced the Model Card Toolkit for the creation of model cards[20]. In reinforcement learning, the creation of standardized evaluation pipelines is continually being researched to enable reproducible benchmarking of different reinforcement learning algorithms (Khetarpal et al. 2018).

The National Cancer Institute Imaging Data Commons is an established example of standardized datasets in biomedical research (Fedorov et al. 2023). As a cloud-based repository, it contains a collection of cancer imaging data and has been used successfully in reproduction experiments in combination with hosting services. Other notable examples include the MIMIC (Johnson et al. 2023) database for electronic health records, or OGB (Hu et al. 2020) for applying ML on graph data. Efforts are also being invested into increasing the reproducibility of language models, for example, with the Holistic Evaluation of Language Models (HELM) (Liang et al. 2022), which offers a broad, scenario-diverse, and multimetric benchmarking suite for language models. As demonstrated by the authors, using HELM, new language models can be evaluated in a more comprehensive way. Furthermore, the Language Model Evaluation Harness (lm-eval) toolkit (Biderman et al. 2024) is a framework designed for language model evaluations and concerned with reproducibility aspects. This tool has already been used by other researchers for more reproducible language model evaluations (Faysse et al. 2024; Kweon et al. 2024). However, it is important to recognize how irreproducible most major models currently are. There exists a live-tracker of model openness[21], which has reported that many projects, even those claiming to be open source, "inherit undocumented data of dubious legality," that few projects share data or model or human reinforcement learning (RLHF) weights, and that "careful scientific documentation is exceedingly rare" (Liesenfeld, Lopez, and Dingemanse 2023).

## Guidelines and checklists

There are many guidelines and checklists that outline best practices for increasing the reproducibility of ML. The guidelines are often aimed at specific parts of the ML workflow. For example, the FAIR principles[22] aim to improve the management and stewardship of scientific data by making scientific data findable, accessible, interoperable, and re-usable. Other guidelines promote the transparency and openness of scientific reporting in general, such as the TOP guidelines[23], which target journals. Similarly, checklists provide a simple framework for ensuring certain criteria are met. Checklists have been applied effectively in the past, for example, in safety-critical systems, where they were used as early as in 1935 to complete preflight checks in Boeing airplanes. A promising example is the ML checklist proposed in Pineau et al. (2021), which has been suggested as best practice by researchers of different fields, for example, in chemistry (Artrith et al. 2021). The checklist requests information about (i) the models and algorithms being used, (ii) theoretical claims in the research article, (iii) data, (iv) code, and (v) the ML experiment(s). However, one drawback of reproducibility checklists when used for academic conferences and journals is the additional workload they impose on already overburdened reviewers. To mitigate this, one suggestion is to leverage LLMs to assist the review process (Liu and Shah 2023).

Finally, numerous guidelines and checklists for ML reproducibility have been recommended in various research fields (Artrith et al. 2021). Especially in biomedical fields, there has been a considerable adoption of guidelines and checklists, such as the TRIPOD statement (Collins et al. 2015), the CLAIM checklist (Mongan, Moy, and Kahn 2020), the ROBUST-ML checklist (Al-Zaiti et al. 2022), or PROBAST (Wolff et al. 2019). A systematic review in biomedical research has shown that the use of checklists is linked to increased reporting quality (Han et al. 2017). The review examined 943 articles over 2 years and found that mandatory checklists increased the inclusion of the main methodological information needed to reproduce the experiments by 65%.

## Model cards and model info sheets

Model cards (Mitchell et al. 2019) are documentation sheets that provide information about ML models, including their intended use, potential limitations, and ethical considerations. They aim to enhance transparency in AI, by detailing aspects such as data used for training, performance metrics, evaluation methodologies, and possible biases. Model cards help users to understand and evaluate ML models more comprehensively, such that they are not deployed in unsuited contexts, and thus to increase reproducibility. Similarly, model info sheets also provide documentation about ML models, but are specifically designed for the detection and prevention of data leakage in ML models (Kapoor and Narayanan 2023). Model info sheets are published alongside research to enable other researchers to quickly verify the validity of the data used to train ML models. They require authors to answer detailed questions about the data and corresponding train/test splits, targeting various types of data leakage (Kapoor and Narayanan 2023).

An empirical study investigating 12 papers, making use of ML methods for prediction, found that a third were subject to some type of data leakage (Kapoor and Narayanan 2023), and that in all those cases, leakage errors could have been prevented by the use of model info sheets. Despite that, model info sheets have two main drawbacks: first, verifying the correctness of info sheets only works after reproducing the results; second, completing these sheets requires a certain level of expertise in ML. In general, model cards and model info sheets represent a promising, low-effort driver for ML reproducibility. They are especially useful in handling some of the methodological issues associated with ML models that could arise (Kapoor and Narayanan 2023).

## Awareness and education

Awareness of reproducibility issues and available training/education to support reproducibility can be a powerful driver for ML reproducibility (Wiggins and Christopherson 2019).

### Publication policies and initiatives

To enhance awareness and establish a minimum of reproducibility standards, the policies of scientific journals are considered an influencing factor. A number of journals already mandate data and/or code availability for publication (Hardwicke et al. 2018; Peng 2015; Pineau et al. 2021). However, to address issues such as result manipulation, more extensive journal participation is needed to, for instance, introduce preregistration where researchers register their research intentions for future publication. This approach ensures credibility by separating the research plan from experimental outcomes (Nosek et al. 2019; Strømland 2019), thereby reducing spin practices, HARKing, and p-hacking (Gundersen, Shamsaliei, and Isdahl 2022). The ACM TORS (Transactions on Recommender Systems) journal exemplifies this by allowing preregistration and publishing "reproducibility papers" dedicated to reproduction studies and enhancing reproducibility tools. Apart from that, various initiatives have been launched to raise awareness of reproducibility issues. A few examples are the following:

– The ReScience journal publishes peer-reviewed papers discussing attempts to reproduce original publications. These reproductions are published on GitHub[24] and available to other researchers (Rougier et al. 2017).

– PapersWithCode.com[25] is a resource for (i) ML papers, accompanied by the code; (ii) datasets; and (iii) ML methods. The ML papers include a link to a repository, which features the code and other artifacts for reproducing the results.

– Reproducibility challenges, where several researchers try to reproduce many recent publications in parallel, are being held frequently. These challenges allow for an analysis of the success rate of reproduction and can be used to evaluate progress over multiple years (Pineau et al. 2021). Additionally, conferences such as the European Conference on Information Retrieval (ECIR) provide special reproducibility tracks, in which researchers are encouraged to reproduce existing papers and build upon their results (e.g., Kowald and Lacic (2022); Kowald, Schedl, and Lex (2020); Muellner, Kowald, and Lex (2021)).

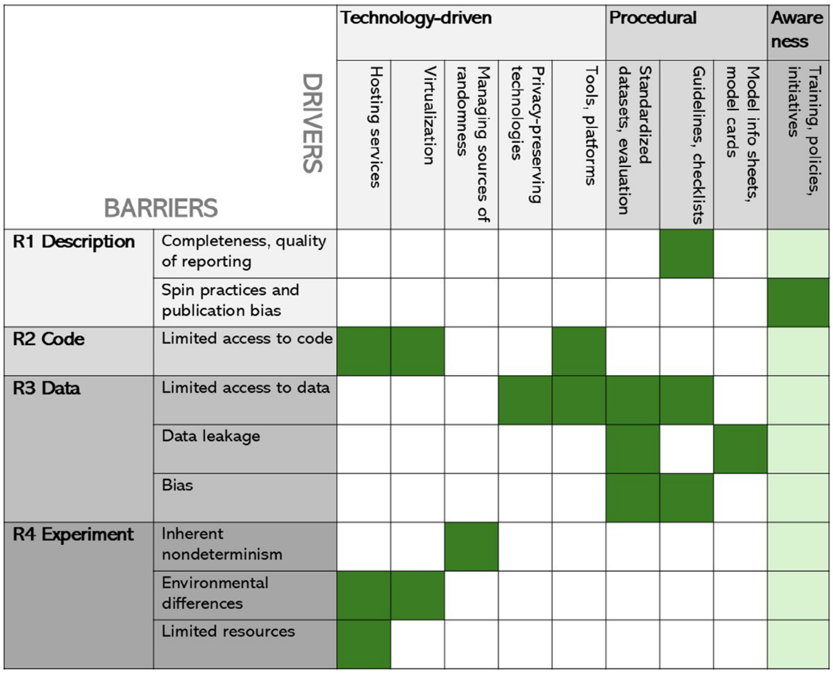
– The ACM has convened a new emerging interest group on reproducibility[26]. The main goals are to (i) contribute to the development of reproducibility standards, practices, and policies; (ii) promote the development and evaluation of tools and methodologies; and (iii) encourage best practices.

– ReproducedPapers.org is another online repository fostering reproductions. It further focuses on education by incorporating a reproduction project into a Master's level ML course at TU Delft (Yildiz et al. 2021).

As also indicated by research in information retrieval and recommender systems, increased awareness and education in the form of publication policies and initiatives can address reproducibility issues by emphasizing robust experimental practices, methodological rigor, and the development of shared resources among the different actors identified, that is, students, educators, scholars, practitioners, and decision-makers (Bauer et al. 2023).

## MAPPING DRIVERS TO BARRIERS

In this section, we map the drivers of reproducibility to the barriers in the form of a drivers–barriers matrix. This will be based on the definition and categorization of



| BARRIERS / DRIVERS | | Technology-driven | | | | | Procedural | | | Aware ness |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Hosting services | Virtualization | Managing sources of randomness | Privacy-preserving technologies | Tools, platforms | Standardized datasets, evaluation | Guidelines, checklists | Model info sheets, model cards | Training, policies, initiatives |
| R1 Description | Completeness, quality of reporting | | | | | | | 🟩 | | ░ |
| | Spin practices and publication bias | | | | | | | | | 🟩 |
| R2 Code | Limited access to code | 🟩 | 🟩 | | | 🟩 | | | | ░ |
| R3 Data | Limited access to data | | | | 🟩 | 🟩 | 🟩 | 🟩 | | ░ |
| | Data leakage | | | | | | 🟩 | | 🟩 | ░ |
| | Bias | | | | | | 🟩 | 🟩 | | ░ |
| R4 Experiment | Inherent nondeterminism | | | 🟩 | | | | | | ░ |
| | Environmental differences | 🟩 | 🟩 | | | | | | | ░ |
| | Limited resources | 🟩 | | | | | | | | ░ |

**FIGURE 2** Drivers–barriers matrix. We map the nine drivers to the nine barriers identified in this paper. The colored boxes show that a specific driver is applicable to a specific barrier. We argue that drivers related to awareness and education are, in general, applicable to address all barriers.

reproducibility as a foundation ("Defining Reproducibility" section), and the identification of the major barriers ("Barriers in ML Reproducibility" section) and drivers ("Drivers for ML Reproducibility" section) of ML reproducibility. The resulting drivers–barriers matrix is depicted in Figure 2 and categorizes the barriers into the four different types of ML reproducibility, that is, *R1 Description*, *R2 Code*, *R3 Data*, *R4 Experiment* (Gundersen 2021), and drivers into technology-driven drivers, procedural drivers, and drivers related to awareness and education.

Our drivers–barriers matrix shows that there are often multiple drivers for the same barrier. Consequently, there are also several possible solutions for a barrier or different aspects of a barrier. The mapping allows us to quickly assess which drivers address the different barriers and which barriers have a higher or lower number of drivers associated with them. It underlines the need for context-dependent approaches instead of "one-size-fits-all" solutions, as the proper selection of a suitable driver depends on the specific conditions and existing barriers relevant to any ML application. We describe intersections between drivers and barriers in more detail in the following and close this section with an overview of strengths and potential limitations of the identified drivers.

**R1 description**. *Completeness and quality of reporting*, as well as *spin practices and publication bias*, present the major barriers associated with *R1 Description*. These are characterized as missing information in reports and overinflated results that hinder reproducibility.

The major drivers for *completeness and report quality* are *guidelines and checklists*. Guidelines provide best practices to adopt in order to achieve reproducible ML research. Furthermore, many checklists exist that comprehensively state the different pitfalls and provide information on how they can be avoided. Researchers can use them to ensure their research meets the desired standards. Furthermore, some of these checklists and guidelines are enforced by journals, such that research will only be published if certain criteria are met. In comparison, *spin practices* are not as easily identifiable. In this case, the discussion within the research community centers around removing the incentives for inflating research results. A particularly effective driver for this is preregistration as an example for *publication policies and initiatives*, where researchers submit research objectives and methods for review before conducting the research. If accepted, the research will be published regardless of the outcome (i.e., whether results are positive, negative, or null), thereby minimizing spin practices.

**R2 code**. Code sharing is essential to reproducibility, which makes *limited access to code* a significant barrier in the field. However, it is often neglected as the process is not trivial. To make shared code useful to the scientific community, it is necessary to share, in addition to the source code, the information about the entire software setup and dependencies, including software versions and hardware configurations. To assist with this, researchers may consider running their code in *hosting services* or *virtualization* environments, which we identified as drivers for code sharing. Both have similar advantages, that is, they can easily be shared and made public for other researchers to use. As a consequence, it will give reproducers immediate access to code, including the complete configuration setup, such as dependencies and versions. Hosting services are a quicker and easier way of achieving this; however, they may be subject to different resource limits. *Virtualization* (e.g., VMs or containers) is more difficult to set up but offers more flexibility and is not externally (e.g., by a provider) restricted in capabilities and resources. Furthermore, *tools and platforms* can be drivers for reproducibility. A lot of ML tools provide capabilities for code versioning or other features, which are key to reproducibility. One example is dToolAI (Hartley and Olsson 2020), which automatically logs the supplemental information of the code, that is, metadata, hyperparameters, and more, which are essential for ML reproducibility.

**R3 data**. Data-related barriers are a severe obstacle to ML reproducibility due to the research fields' data-driven nature, where *limited access to data* forms a major challenge. Privacy concerns are among the crucial arguments that cause hesitation in sharing data. The need for data privacy is evident, especially in biomedical fields, which deal with patients' electronic health records. Nevertheless, it increases reproducibility issues in ML-based science and, thus, delays technological progress within these domains. However, there are several approaches that aim to meet the requirements of sharing sensitive data: *Privacy-preserving technologies* allow reproducers to train ML models on private data without actually possessing the data. This way, reproduction becomes possible without violating potential privacy regulations. Other than that, the use of *standardized datasets and evaluation* can support issues in regard to dataset meta-information, including the specification of train-test splits and data provenance. Once again, *tools and platforms* can assist with data versioning, and numerous *guidelines and checklists* have been proposed to address the provenance of data. These guidelines and checklists are designed to help researchers to avoid common pitfalls. Current initiatives are supported by journals that more frequently require data to be shared as part of a publication.

Concerning methodological errors associated with the data, *data leakage* is a major issue, which can, for instance, be mitigated using *standardized datasets and evaluation*. Other drivers to solve data leakage are *model info sheets and model cards*, which are provided as supplemental information to a published dataset. Even though there are some limitations to *model info sheets*, they are capable of detecting all types of data leakage. *Bias* is another methodological error, leading to irreproducible results. This is because the biased data usually do not generalize well to problems outside the experimental setup of a specific ML study. *Bias* has been an important source of concern, for example, in biomedical fields. Effects thereof can again be minimized using *standardized datasets and evaluation* or specific *guidelines and checklists*, for example, ROBUST-ML (Al-Zaiti et al. 2022).

**R4 experiment**. If an ML experiment is shared entirely and code and data are available, that is, reproducibility type *R4 Experiment*, there are still three barriers, which can lead to irreproducible results. *Inherent nondeterminism* arises from the different sources of randomness in ML, and makes it difficult to achieve repeatable results, even on the same machine. There are, however, methods to *manage the sources of randomness*, such as fixed random seeds and deterministic implementations, while comprehensively mitigating all sources of randomness is still a very challenging endeavor.

Another barrier is described as *environmental differences*, which has two main issues associated with it, that is, software differences and hardware differences. Both types of differences can be avoided by using either *hosting services* or *virtualization*; constraints can be assumed to be similar to the barrier of *limited access to code*. *Limited access to computational resources* constitutes another barrier to ML reproducibility identified in this work. The issue is particularly problematic for research using LLMs because of their need for extensive computational resources in training and reproduction. *Hosting services* offer a solution, providing access to pretrained models and allowing researchers to directly access and run respective models on-site. Finally, Table 1 gives an overview of strengths and potential limitations of the identified drivers. As we can see, the choice of using a particular driver strongly depends on the given use and to what extent potential limitations are applicable for the use case.

## CONCLUSION

In this paper, we examined the barriers and drivers associated with the four types of ML reproducibility as outlined by Gundersen et al. (description, data, code, and experiment) (Gundersen 2021), specifically in the cases of computer science and biomedical research. We synthesized our findings into a drivers–barriers matrix to summarize and illustrate which drivers are feasible solutions to the various barriers. We observe that the barriers to ML reproducibility can be addressed through three kinds of drivers: technology-driven solutions, procedural improvements, and enhanced awareness and education. It is important to highlight that, in theory, awareness and education can complement the other drivers and serve as a foundational basis for overcoming reproducibility-related challenges.

One of the main issues hindering reproducibility in research appears to be rooted in the cultural aspects of research communities. As argued by Bauer et al. (2023) and Chiarelli, Loffreda, and Johnson (2021), the current incentives for conducting reproducible research are limited, and open research is often regarded as an unrewarded

**TABLE 1** Comparison of strengths and weaknesses of our identified drivers.

| Driver | Strengths | Potential Limitations |
|---|---|---|
| Hosting services | Facilitates sharing of models, code, and datasets; increases accessibility. | Out-of-the-box reproducibility is not yet provided and there are limits to the available compute. |
| Virtualization | Enables environment replication (e.g., Docker, virtual machines); resolves dependency issues. | Requires technical expertise; may introduce overhead for simple experiments. |
| Managing sources of randomness | Critical for deterministic outcomes; has the potential to reduce and even eliminate variances across multiple runs. | Can be hard to implement consistently across frameworks; only leads to point estimates of performance. |
| Privacy-preserving technologies | Expands access to sensitive datasets without compromising privacy. | Still an emerging field; performance trade-offs can make widespread adoption slower. |
| Tools and platforms | Can streamline reproducibility practices and automatically acquire reproducibility artifacts. | Out-of-the-box reproducibility is not yet provided, and fragmentation of tools can lead to siloed solutions rather than unified workflows. |
| Standardized datasets and evaluation | Provides consistency and comparability for results across studies. | May not generalize well to niche or domain-specific problems and can be subject to privacy concerns. |
| Guidelines, checklists | Promotes best practices through structured processes (e.g., reproducibility checklists). | Compliance can be time-consuming and may not be enforced consistently. |
| Model info sheets and model cards | Improves transparency around model design and intended use. | Adoption is still limited; requires effort to standardize and maintain across the community. |
| Publication policies, initiatives | Drive cultural change by incentivizing openness (e.g., benchmarks, competitions). | Impact depends on community participation and is a slow process in general. |

additional effort. Consequently, there is a lack of training and insufficient funding to cover the additional time and resources required by researchers. Notably, the lack of funding also impacts the ability to perform quality checks during and after the publication process. Therefore, we strongly believe that the way forward towards ML reproducibility is rooted in better education and more awareness of this topic among all involved stakeholders, for example, students, educators, researchers, publishers, and policymakers. This, combined with the other tools and drivers described in this paper, could lead to more reproducible ML pipelines and, with this, more robust findings.

From a more technical perspective, the rise of AutoML tools for ML development and ML tasks performed by domain experts, (potentially) not having in-depth computer or data science knowledge, could pose another barrier to reproducibility (Haberl and Thalmann 2025). We thus believe that future work should address the increasing use of AutoML tools for AI development in research (Haberl and Thalmann 2025) among noncomputer or data science experts. While these easy-to-use tools can standardize ML workflows by default and include documentation features, domain experts often lack the necessary expertise to recognize potential problems associated with ML, such as biased or imbalanced data. Thus, research on reproducibility should emphasize this challenge and aim to establish standards and guidelines for the use of No and Low Code ML tools in research, as well as the training required for their responsible application.

In summary, we hope that our paper provides practical guidance and orientation for researchers employing ML and clarifies the current state of play. Of course, in such a dynamic and fast-paced research area, this discussion opens up a series of further questions and avenues for exploration. We recommend further investigation of the various issues and potential solutions laid out here. We would also encourage further investigation into the potential role of platforms (Gundersen, Shamsaliei, and Isdahl 2022) or foundation models (Hosseini et al. 2024) in further exacerbating or alleviating these challenges.

## ACKNOWLEDGMENTS

This research is supported by the Horizon Europe project TIER2 (GA: 101094817), and the FFG COMET program.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

*Harald Semmelrock* https://orcid.org/0009-0000-1759-1570
*Tony Ross-Hellauer* https://orcid.org/0000-0003-4470-7027
*Simone Kopeinik* https://orcid.org/0000-0002-6440-7286
*Dieter Theiler* https://orcid.org/0000-0001-5340-8374
*Armin Haberl* https://orcid.org/0009-0000-5356-5126
*Stefan Thalmann* https://orcid.org/0000-0001-6529-7958
*Dominik Kowald* https://orcid.org/0000-0003-3230-6234

## ENDNOTES

[1] https://portal.core.edu.au/conf-ranks/
[2] https://github.com/mlcommons/croissant
[3] https://github.com/mlcommons
[4] https://www.researchobject.org/ro-crate/
[5] https://pytorch.org/
[6] https://www.tensorflow.org/
[7] https://www.kaggle.com/code
[8] https://colab.google/
[9] https://codalab.org/
[10] https://www.docker.com/
[11] https://codeocean.com/
[12] https://reproducibility.sigmod.org/
[13] https://dvc.org/
[14] https://git-scm.com/
[15] https://mlflow.org/
[16] https://www.researchobject.org/ro-crate/
[17] https://h2o.ai
[18] https://cloud.google.com/automl
[19] https://www.datarobot.com/platform
[20] https://research.google/blog/introducing-the-model-card-toolkit-for-easier-model-transparency-reporting/
[21] https://opening-up-chatgpt.github.io/
[22] https://www.go-fair.org/fair-principles/
[23] https://www.cos.io/initiatives/top-guidelines
[24] https://github.com/
[25] https://paperswithcode.com/
[26] https://reproducibility.acm.org/

## REFERENCES

Ahmed, H., and J. Lofstead. 2022. "Managing Randomness to Enable Reproducible Machine Learning." In Proceedings of the 5th International Workshop on Practical Reproducible Evaluation of Computer Systems, 15–20. Minneapolis, MN: ACM. ISBN 978-1-4503-9313-3.

Ahmed, H., R. Tchoua, and J. Lofstead. 2022. "Measuring Reproduciblity of Machine Learning Methods for Medical Diagnosis." In 2022 Fourth International Conference on Transdisciplinary AI (TransAI), 9–16.

Al-Zaiti, S. S., A. A. Alghwiri, X. Hu, G. Clermont, A. Peace, P. Macfarlane, and R. Bond. 2022. "A Clinician's Guide to Understanding and Critically Appraising Machine Learning Studies: A Checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML)." *European Heart Journal - Digital Health* 3(2): 125–40.

Alahmari, S. S., D. B. Goldgof, P.R. Mouton, and L. O. Hall. 2020. "Challenges for the Repeatability of Deep Learning Models." *IEEE Access* 8: 211860–68.

Albertoni, R., S. Colantonio, P. Skrzypczyński, and J. Stefanowski. 2023. "Reproducibility of Machine Learning: Terminology, Recommendations and Open Issues." *arXiv preprint arXiv:2302.12691*.

Navarro, C. L. A., J. A. Damen, T. Takada, S. W. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. Moons, and L. Hooft. 2023. "Systematic Review Finds "Spin" Practices and Poor Reporting Standards in Studies on Machine Learning-Based Prediction Models." *Journal of Clinical Epidemiology* 158: 99–110.

Artrith, N., K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, and A. Walsh. 2021. "Best Practices in Machine Learning for Chemistry." *Nature Chemistry* 13(6): 505–8.

Aslanyan, Z., and P. Vasilikos. 2020. *Privacy-Preserving Machine Learning*, The Alexandra Institute, Aarhus, NL.

Aumjaud, P., D. McAuliffe, F. J. R. Lera, and P. Cardiff. 2021. "rl_reach: Reproducible Reinforcement Learning Experiments for Robotic Reaching Tasks." *Software Impacts* 8: 100061.

Baker, M. 2016. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533(7604): 452–54.

Banks, G. C., S. G. Rogelberg, H. M. Woznyj, R. S. Landis, and D. E. Rupp. 2016. "Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly." *Journal of Business and Psychology* 31: 323–38.

Bauer, C., B. Carterette, N. Ferro, N. Fuhr, and G. Faggioli. 2023. "Frontiers of Information Access Experimentation for Research and Education (Dagstuhl Seminar 23031)." In *Dagstuhl Reports*, vol. 13. Dagstuhl, Germany, Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Beam, A. L., A. K. Manrai, and M. Ghassemi. 2020. "Challenges to the Reproducibility of Machine Learning Models in Health Care." *JAMA* 323(4): 305–6.

Belz, A., S. Agarwal, A. Shimorina, and E. Reiter. 2021. "A Systematic Review of Reproducibility Research in Natural Language Processing." In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 381–93.

Biagioli, M., and A. Lippman. 2020. *Gaming the Metrics: Misconduct and Manipulation in Academic Research*. Cambridge, MA: MIT Press.

Biderman, S., H. Schoelkopf, L. Sutawika, L. Gao, J. Tow, B. Abbasi, A. F. Aji, et al. 2024. "Lessons from the Trenches on Reproducible Evaluation of Language Models." *arXiv preprint arXiv:2405.14782*.

Boettiger, C. 2015. "An Introduction to Docker for Reproducible Research." *ACM SIGOPS Operating Systems Review*, 49(1): 71–79.

Bontempi, D., L. Nuernberg, S. Pai, D. Krishnaswamy, V. Thiriveedhi, A. Hosny, R. H. Mak, et al. 2024. "End-to-End Reproducible AI Pipelines in Radiology Using the Cloud." *Nature Communications*, 15(1): 6931.

Brockman, G., V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. 2016. "Openai Gym." *arXiv preprint arXiv:1606.01540*.

Burrows, R. 2012. "Living with the h-Index? Metric Assemblages in the Contemporary Academy." *The Sociological Review*, 60(2): 355–72.

Chen, B., M. Wen, Y. Shi, D. Lin, G. K. Rajbahadur, and Z. M. J. Jiang. 2022. "Towards Training Reproducible Deep Learning Models." In *Proceedings of the 44th International Conference on Software Engineering*, ICSE'22, 2202–14. New York, NY: ACM. ISBN 9781450392211.

Chiarelli, A., L. Loffreda, and R. Johnson. 2021. "The Art of Publishing Reproducible Research Outputs: Supporting Emerging Practices Through Cultural and Technological Innovation." Knowledge Exchange.

Chirigati, F., R. Rampin, D. Shasha, and J. Freire. 2016. "ReproZip: Computational Reproducibility with Ease." In Proceedings of the 2016 International Conference on Management of Data, *SIGMOD '16*, 2085–88. New York, NY: ACM. ISBN 978-1-4503-3531-7.

Choi, Y.-D., B. Roy, J. Nguyen, R. Ahmad, I. Maghami, A. Nassar, Z. Li, A. M. Castronova, T. Malik, S. Wang, et al. 2023. "Comparing Containerization-Based Approaches for Reproducible Computational Modeling of Environmental Systems." *Environmental Modelling & Software* 167: 105760.

Clyburne-Sherin, A., X. Fei, and S. A. Green. 2019. "Computational Reproducibility via Containers in Psychology." *Meta-Psychology*, 3: 1–9.

Colas, C., O. Sigaud, and P.-Y. Oudeyer. 2018. "How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments." *arXiv preprint arXiv:1806.08295*.

S Collaboration, O. 2015. "Estimating the reproducibility of psychological science." *Science* 349(6251): aac4716.

Collins, G. S., J. B. Reitsma, D. G. Altman, and K. G. Moons. 2015. "Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD)." *Circulation*, 131(2): 211–19.

Cremonesi, P., and D. Jannach. 2021. "Progress in Recommender Systems Research: Crisis? What Crisis?" *AI Magazine* 42(3): 43–54.

D'Antonoli, T. A., R. Cuocolo, B. Baessler, and D. P. Dos Santos. 2024. "Towards Reproducible Radiomics Research: Introduction of a Database for Radiomics Studies." *European Radiology* 34(1): 436–43.

De Cristofaro, E. 2021. "A Critical Overview of Privacy in Machine Learning." *IEEE Security & Privacy*, 19(4): 19–27.

Dror, R., G. Baumer, M. Bogomolov, and R. Reichart. 2017. "Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets." *Transactions of the Association for Computational Linguistics*, 5: 471–86.

Dwivedi, Y. K., L. Hughes, E. Ismagilova, G. Aarts, C. Coombs, T. Crick, Y. Duan, et al. 2021. "Artificial Intelligence (AI): Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy." *International Journal of Information Management* 57: 101994.

S Editorial, N. C. 2022. "Seamless Sharing and Peer Review of Code." *Nature Computational Science* 2(12): 773.

Errington, T. M., M. Mathur, C. K. Soderberg, A. Denis, N. Perfito, E. Iorns, and B. A. Nosek. 2021. "Investigating the Replicability of Preclinical Cancer Biology." *Elife* 10: e71601.

Faysse, M., P. Fernandes, N. Guerreiro, A. Loison, D. Alves, C. Corro, N. Boizard, et al. 2024. "CroissantLLM: A Truly Bilingual French-English Language Model." *arXiv preprint arXiv:2402.00786*.

Fedorov, A., W. J. Longabaugh, D. Pot, D. A. Clunie, Pieper, S. D., D. L. Gibbs, C. Bridge, et al. 2023. "National Cancer Institute Imaging Data Commons: Toward Transparency, Reproducibility, and Scalability in Imaging Artificial Intelligence." *Radiographics* 43(12): e230180.

Ferigo, D., S. Traversaro, G. Metta, and D. Pucci. 2020. "Gym-Ignition: Reproducible Robotic Simulations for Reinforcement Learning." In 2020 IEEE/SICE International Symposium on System Integration (SII), 885–90.

Fidler, F., and J. Wilcox. 2021. "Reproducibility of Scientific Results." In *The Stanford Encyclopedia of Philosophy*, edited by N. Zalta. Stanford, United States, Metaphysics Research Lab, Stanford University. Summer 2021 edition.

Freedman, L. P., I. M. Cockburn, and T. S. Simcoe. 2015. "The Economics of Reproducibility in Preclinical Research." *PLoS Biology* 13(6): e1002165.

Gibney, E. 2022. "Could Machine Learning Fuel a Reproducibility Crisis in Science?" *Nature*, 608(7922): 250–51. Bandiera_abtest: a Cg_type: News Number: 7922 Publisher: Nature Publishing Group Subject_term: Machine learning, Publishing, Mathematics and computing.

Goodman, S. N., D. Fanelli, and J. P. Ioannidis, 2016. "What Does Research Reproducibility Mean?" *Science Translational Medicine*, 8(341): 341ps12.

E Gundersen, O. 2021. "The Fundamental Principles of Reproducibility." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379(2197): 20200210.

Gundersen, O. E., K. Coakley, C. Kirkpatrick, and Y. Gil. 2023. "Sources of Irreproducibility in Machine Learning: A Review." *arXiv preprint arXiv:2204.07610*.

Gundersen, O. E., and S. Kjensmo. 2018. "State of the Art: Reproducibility in Artificial Intelligence." *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1).

Gundersen, O. E., S. Shamsaliei, and R. J. Isdahl. 2022. "Do Machine Learning Platforms Provide Out-of-the-Box Reproducibility?" *Future Generation Computer Systems* 126: 34–47.

Gundersen, O. E., S. Shamsaliei, H. S. Kjærnli, and H. Langseth. 2023. "On Reporting Robust and Trustworthy Conclusions from Model Comparison Studies Involving Neural Networks and Randomness." In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*, 37–61.

Haberl, A., and S. Thalmann. 2025. "Automated Machine Learning in Research – a Literature Review." In Proceedings of the 58th Hawaii International Conference on System Sciences.

Han, S., T. F. Olonisakin, J. P. Pribis, J. Zupetic, J. H. Yoon, K. M. Holleran, K. Jeong, N. Shaikh, D. M. Rubio, and J. S. Lee. 2017. "A Checklist is Associated with Increased Quality of Reporting Preclinical Biomedical Research: A Systematic Review." *PloS One* 12(9): e0183591.

Hardwicke, T. E., M. B. Mathur, K. MacDonald, G. Nilsonne, G. C. Banks, M. C. Kidwell, A. Hofelich Mohr, et al. 2018. "Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the journal Cognition." *Royal Society Open Science*, 5(8): 180448.

Hardwicke, T. E., J. D. Wallach, M. C. Kidwell, T. Bendixen, S. Crüwell, and J. P. Ioannidis. 2020. "An Empirical Assessment of Transparency and Reproducibility-Related Research Practices in the Social Sciences (2014–2017)." *Royal Society Open Science*, 7(2): 190806.

Hartley, M., and T. S. G. Olsson, 2020. "dtoolAI: Reproducibility for Deep Learning." *Patterns*, 1(5): 100073.

Heil, B. J., M. M. Hoffman, F. Markowetz, S.-I. Lee, C. S. Greene, and S. C. Hicks. 2021. “Reproducibility Standards for Machine Learning in the Life Sciences.” *Nature Methods*, 18(10): 1132–35.

Hong, S.-Y., M.-S. Koo, J. Jang, J.-E. E. Kim, H. Park, M.-S. Joh, J.-H. Kang, and T.-J. Oh. 2013. “An Evaluation of the Software System Dependency of a Global Atmospheric Model.” *Monthly Weather Review*, 141(11): 4165–72.

Hosseini, M., S. P. Horbach, K. Holmes, and T. Ross-Hellauer. 2024. “Open Science at the Generative AI Turn: An Exploratory Analysis of Challenges and Opportunities.” *Quantitative Science Studies* 6: 1–24.

Hu, W., M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec. 2020. “Open Graph Benchmark: Datasets for Machine Learning on Graphs.” *Advances in Neural Information Processing Systems* 33: 22118–33.

Hutson, M. 2018a. “Artificial Intelligence Faces Reproducibility Crisis Unpublished Code and Sensitivity to Training Conditions Make Many Claims Hard to Verify.” *Science*, 359(6377): 725–26.

Hutson, M. 2018b. “Missing Data Hinder Replication of Artificial Intelligence Studies.” *Science* 2018: 1–4.

P Ioannidis, J. 2005. “Why Most Published Research Findings are False.” *PLoS Medicine* 2(8): e124.

Iqbal, S. A., J. D. Wallach, M. J. Khoury, S. D. Schully, and J. P. Ioannidis. 2016. “Reproducible Research Practices and Transparency Across the Biomedical Literature.” *PLoS Biology*, 14(1): e1002333.

Ivanescu, A. E., P. Li, B. George, A. W. Brown, S. W. Keith, D. Raju, and D. B. Allison. 2016. “The Importance of Prediction Model Validation and Assessment in Obesity and Nutrition Research.” *International Journal of Obesity (2005)* 40(6): 887–94.

Johnson, A. E., L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, et al. 2023. “MIMIC-IV, a Freely Accessible Electronic Health Record Dataset.” *Scientific Data* 10(1): 1.

Rahimi, A. K., O. J. Canfell, W. Chan, B. Sly, J. D. Pole, C. Sullivan, and S. Shrapnel. 2022. “Machine Learning Models for Diabetes Management in Acute Care Using Electronic Medical rRecords: A Systematic Review.” *International Journal of Medical Informatics* 162: 104758.

Kapoor, S., and A. Narayanan. 2023. “Leakage and the Reproducibility Crisis in Machine-Learning-Based Science.” *Patterns* 4(9): 100804.

Khan, T., M. Budzys, K. Nguyen, and A. Michalas. 2024. “Wildest Dreams: Reproducible Research in Privacy-Preserving Neural Network Training.” *arXiv preprint arXiv:2403.03592*.

Khetarpal, K., Z. Ahmed, A. Cianflone, R. Islam, and J. Pineau. 2018. “RE-EVALUATE: Reproducibility in Evaluating Reinforcement Learning Algorithms.” In 2nd Reproducibility in Machine Learning Workshop at ICML.

Koenigstorfer, F., A. Haberl, D. Kowald, T. Ross-Hellauer, and S. Thalmann. 2024. “Black Box or Open Science? Assessing Reproducibility-Related Documentation in AI Research.” In Proceedings of the 57th Hawaii International Conference on System Sciences.

Kowald, D., and E. Lacic. 2022. “Popularity Bias in Collaborative Filtering-Based Multimedia Recommender Systems.” In BIAS WS at ECIR, 1–11. Springer.

Kowald, D., M. Schedl, and E. Lex. 2020. “The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study.” In 42nd European Conference on IR Research, ECIR 2020, 35–42. Springer.

Kowald, D., S. Scher, V. Pammer-Schindler, P. Müllner, K. Waxnegger, L. Demelius, A. Fessl, et al. 2024. “Establishing and Evaluating Trustworthy AI: Overview and Research Challenges.” *Frontiers in Big Data* 7: 1467222.

Kweon, S., B. Choi, M. Kim, R. W. Park, and E. Choi. 2024. “KorMedMCQA: Multi-Choice Question Answering Benchmark for Korean Healthcare Professional Licensing Examinations.” *arXiv preprint arXiv:2403.01469*.

Lariviere, V., and C. R. Sugimoto. 2019. “The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects.” In Springer Handbook of Science and Technology Indicators, 3–24. Cham: Springer.

Lee, C. J., C. R. Sugimoto, G. Zhang, and B. Cronin. 2013. “Bias in Peer Review.” *Journal of the American Society for Information Science and Technology* 64(1): 2–17.

Leventi-Peetz, A.-M., and T. Östreich. 2022. “Deep Learning Reproducibility and Explainable AI (XAI).” *arXiv preprint arXiv:2202.11452*.

Liang, P., R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, et al. 2022. “Holistic Evaluation of Language Models.”*arXiv preprint arXiv:2211.09110*.

Liesenfeld, A., A. Lopez, and M. Dingemanse. 2023. “Opening up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators.” In Proceedings of the 5th International Conference on Conversational User Interfaces, 1–6.

Liu, R., and N. B. Shah. 2023. “ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing.” *arXiv preprint arXiv:2306.00622*.

McDermott, M. B. A., S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi. 2021. “Reproducibility in Machine Learning for Health Research: Still a Ways to Go.” *Science Translational Medicine* 13(586): eabb1655.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. “A Survey on Bias and Fairness in Machine Learning.” *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. 2019. “Model Cards for Model Reporting.” In Proceedings of the Conference on Fairness, Accountability, and Transparency, *FAT*'19*, 220–29. New York, NY: ACM. ISBN 978-1-4503-6125-5.

Mongan, J., L. Moy, and C. E. Kahn. 2020. “Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers.” *Radiology: Artificial Intelligence* 2(2): e200029.

Moreau, D., K. Wiebels, and C. Boettiger. 2023. “Containers for Computational Reproducibility.” *Nature Reviews Methods Primers* 3(1): 50.

Muellner, P., D. Kowald, and E. Lex. 2021. “Robustness of Meta Matrix Factorization Against Strict Privacy Constraints.” In 43rd European Conference on IR Research, ECIR 2021, 107–19. Springer.

Munafò, M. R., B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. P. du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. 2017. “A Mnifesto for Reproducible Science.” *Nature Human Behaviour* 1(1): 1–9.

Nagarajan, P., G. Warnell, and P. Stone. 2018. “Deterministic Implementations for Reproducibility in Deep Reinforcement Learning.” *arXiv preprint arXiv:1809.05676*.

Leija, N., K. Shiptoski O. S., R. G. Scott, B. Wang, N. Renner, R. R. Newton, and J. Devietti. 2020. "Reproducible Containers." In Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems, 167–82. Lausanne: ACM. ISBN 978-1-4503-7102-5.

Nosek, B. A., E. D. Beck, L. Campbell, J. K. Flake, T. E. Hardwicke, D. T. Mellor, A. E. van't Veer, and S. Vazire. 2019. "Preregistration Is Hard, And Worthwhile." *Trends in Cognitive Sciences* 23(10): 815–18.

Nosek, B. A., T. E. Hardwicke, H. Moshontz, A. Allard, K. S. Corker, A. Dreber, F. Fidler, et al. 2022. "Replicability, Robustness, and Reproducibility in Psychological Science." *Annual Review of Psychology*, 73(1): 719–48.

Ooi, K.-B., G. W.-H. Tan, M. Al-Emran, M. A. Al-Sharafi, A. Capatina, A. Chakraborty, Y. K. Dwivedi, et al. 2023. "The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions." *Journal of Computer Information Systems*, 65: 1–32.

Panaretos, D., E. Koloverou, A. C. Dimopoulos, G.-M. Kouli, M. Vamvakari, G. Tzavelas, C. Pitsavos, and D. B. Panagiotakos. 2018. "A Comparison of Statistical and Machine-Learning Techniques in Evaluating the Association Between Dietary Patterns and 10-Year Cardiometabolic Risk (2002–2012): The ATTICA Study." *British Journal of Nutrition*, 120(3): 326–34.

Paullada, A., I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. 2021. "Data and Its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research." *Patterns*, 2(11): 100336.

Peng, R. 2015. "The Reproducibility Crisis in Science: A Statistical Counterattack." *Significance*, 12(3): 30–32.

Pineau, J., P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, and H. Larochelle. 2021. "Improving Reproducibility in Machine Learning Research (a Report from the NeurIPS 2019 Reproducibility Program)." *The Journal of Machine Learning Research*, 22(1): 164:7459–164:7478.

Pletzl, S., A. Haberl, T. Ross-Hellauer, and S. Thalmann. 2024. "Reproducible AutoML: An Assessment of Research Reproducibility of No-Code AutoML Tools." In Wirtschaftsinformatik 2024 Proceedings.

Pontika, N., T. Klebel, A. Correia, H. Metzler, P. Knoth, and T. Ross-Hellauer. 2022. "Indicators of Research Quality, Quantity, Openness, and Responsibility in Institutional Review, Promotion, and Tenure Policies Across Seven Countries." *Quantitative Science Studies* 3(4): 888–911.

Pouchard, L., Y. Lin, and H. Van Dam. 2020. "Replicating Machine Learning Experiments in Materials Science." In Parallel Computing: Technology Trends, 743–55. Amsterdam: IOS Press.

Provenzano, D., Y. J. Rao, S. Goyal, S. Haji-Momenian, J. Lichtenberger, and M. Loew. 2021. "Radiologist vs Machine Learning: A Comparison of Performance in Cancer Imaging." In 2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1–10. ISSN: 2332-5615.

Pushkarna, M., A. Zaldivar, and O. Kjartansson. 2022. "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI." In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, *FAccT'22*, 1776–826. New York, NY: ACM. ISBN 978-1-4503-9352-2.

Quaranta, L., F. Calefato, and F. Lanubile. 2022. "A Taxonomy of Tools for Reproducible Machine Learning Experiments." In CEUR Workshop Proceedings.

Raff, E.. 2019. "A Step Toward Quantifying Independently Reproducible Machine Learning Research." In *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.

Raste, S., R. Singh, J. Vaughan, and V. N. Nair. 2022. "Quantifying Inherent Randomness in Machine Learning Algorithms." *SSRN Electronic Journal. arXiv preprint arXiv:2206.12353*.

Rijcke, S. d., P. F. Wouters, A. D. Rushforth, T. P. Franssen, and B. Hammarfelt. 2016. "Evaluation Practices and Effects of Indicator Use—a Literature Review." *Research Evaluation* 25(2): 161–69.

Rougier, N. P., K. Hinsen, F. Alexandre, T. Arildsen, L. A. Barba, F. C. Y. Benureau, C. T. Brown, et al. 2017. "Sustainable Computational Science: The ReScience Initiative." *PeerJ Computer Science* 3: e142.

Rouzrokh, P., B. Khosravi, S. Faghani, M. Moassefi, D. V. Vera Garcia, Y. Singh, K. Zhang, G. M. Conte, and B. J. Erickson. 2022. "Mitigating Bias in Radiology Machine Learning: 1. Data Handling." *Radiology: Artificial Intelligence* 4(5): e210290.

Saidi, P., G. Dasarathy, and V. Berisha. 2024. "Unraveling Overoptimism and Publication Bias in ML-Driven Science." *arXiv:2405.14422*.

Schacherer, D. P., M. D. Herrmann, D. A. Clunie, H. Höfener, W. Clifford, W. J. Longabaugh, S. Pieper, R. Kikinis, A. Fedorov, and A. Homeyer. 2023. "The NCI Imaging Data Commons as a Platform for Reproducible Research in Computational Pathology." *Computer Methods and Programs in Biomedicine* 242: 107839.

Schlegel, M., and K.-U. Sattler. 2023. "Management of Machine Learning Lifecycle Artifacts: A Survey." *ACM SIGMOD Record* 51(4): 18–35.

Semmelrock, H., S. Kopeinik, D. Theiler, T. Ross-Hellauer, and D. Kowald. 2023. "Reproducibility in Machine Learning-Driven Research." *arXiv preprint arXiv:2307.10320*.

Shahriari, M., R. Ramler, and L. Fischer. 2022. "How Do Deep-Learning Framework Versions Affect the Reproducibility of Neural Network Models?" *Machine Learning and Knowledge Extraction* 4(4): 888–911.

Smaldino, P. E., and R. McElreath. 2016. "The Natural Selection of Bad Science." *Royal Society Open Science* 3(9): 160384.

Soiland-Reyes, S., P. Sefton, M. Crosas, L. J. Castro, F. Coppens, J. M. Fernández, D. Garijo, et al. 2022. "Packaging Research Artefacts with RO-Crate." *Data Science*, 5(2): 97–138.

Strømland, E. 2019. "Preregistration and Reproducibility." *Journal of Economic Psychology* 75: 102143.

Tassa, Y., Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, et al. 2018. "Deepmind Control Suite." *arXiv preprint arXiv:1801.00690*.

Tatman, R., J. VanderPlas, and S. Dane. 2018. "A Practical Taxonomy of Reproducibility for Machine Learning Research." In CEUR Workshop Proceedings.

Trosten, D. J. 2023. "Questionable Practices in Methodological Deep Learning Research." In *Proceedings of the Northern Lights Deep Learning Workshop*, vol. 4.

Walonoski, J., M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan. 2018. "Synthea: An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record." *Journal of the American Medical Informatics Association: JAMIA* 25(3): 230–38.

Wiggins, B. J., and C. D. Christopherson. 2019. "The Replication Crisis in Psychology: An Overview for Theoretical and Philosophical Psychology." *Journal of Theoretical and Philosophical Psychology* 39: 202–17.

Wolff, R. F., K. G. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett. 2019. "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies." *Annals of Internal Medicine*, 170(1): 51–58.

Xu, R., N. Baracaldo, and J. Joshi. 2021. "Privacy-Preserving Machine Learning: Methods, Challenges and Directions." *arXiv preprint arXiv:2108.04417*.

Yildiz, B., H. Hung, J. Krijthe, C. Liem, M. Loog, G. Migut, F. Oliehoek, et al. 2021. "ReproducedPapers.org: Openly Teaching and Structuring Machine Learning Reproducibility." In *International Workshop on Reproducible Research in Pattern Recognition*, 3–11. Springer.

Young, K., and T. Tian. 2019. "MinAtar: An Atari-Inspired Testbed for Thorough and Reproducible Reinforcement Learning Experiments." *arXiv preprint arXiv:1903.03176*.

**How to cite this article:** Semmelrock, H., T. Ross-Hellauer, S. Kopeinik, D. Theiler, A. Haberl, S. Thalmann, and D. Kowald. 2025. "Reproducibility in machine-learning-based research: Overview, barriers, and drivers." *AI Magazine* 46: e70002. https://doi.org/10.1002/aaai.70002

## AUTHOR BIOGRAPHIES

**Harald Semmelrock**, B.Sc., is a graduate student pursuing an M.Sc. in Computer Science at ETH Zurich, Switzerland, majoring in Machine Intelligence with a minor in Data Management. He previously earned his B.Sc. in Computer Science with distinction from Graz University of Technology, Austria. His research interests lie in the fields of artificial intelligence, machine learning, and data science.

**Dr. Tony Ross-Hellauer** is leader of the Open and Reproducible Research Group at Know Center Research GmbH Graz, Austria. He has a PhD in Information Studies (University of Glasgow, 2012), as well as degrees in Information and Library Studies and Philosophy. His research focuses on a range of issues related to open science evaluation, skills, policy, governance, monitoring and infrastructure.

**Dr. Simone Kopeinik** is a Key Researcher and Project Manager at the Know-Center, where she serves as the Deputy Research Area Manager in the Fair AI Department. She graduated from Graz University of Technology with a degree in Computer Science. Dr. Kopeinik's research interests include fairness and nondiscrimination in AI, addressing biases in data and systems, human-centered computing, recommender systems, and user and cognitive modeling.

**Dieter Theiler**, M.Sc., is a Software Engineer in the FAIR AI group at Know Center Research GmbH Graz, Austria. He holds a MSc and BSc in Computer Science from Graz University of Technology, Austria. He mainly works in the fields of recommender systems and reproducible AI.

**Armin Haberl**, B.Sc. and M.Sc., is a University Assistant at the Business Analytics and Data Science-Center at the University of Graz, where he teaches and conducts research as part of his doctoral studies in Information Systems. He holds both an M.Sc. and a B.Sc. in Business Administration from the University of Graz. His research focuses on reproducible research using low-code and no-code automated machine learning tools.

**Univ. Prof. Dr. Stefan Thalmann**, is a Professor for Business Analytics and Data Science as well as a Director of the BANDAS Center at the University of Graz, Austria. He holds a habilitation and a PhD in Information Systems from University of Innsbruck, Austria. His research interests are in the fields of No and Low Code AI, Data-Driven Decision Support, Smart Regulation and auditing of AI.

**Priv. Doz. Dr. Dominik Kowald** is Research Area Manager of the FAIR AI team at Know Center Research GmbH Graz, Austria. He holds a habilitation (Priv.-Doz.) in Applied Computer Science, as well as a Ph.D. (with hons), M.Sc. (with hons), and B.Sc. in Computer Science from Graz University of Technology, Austria. His research interests are in the fields of trustworthy and reproducible AI, recommender systems, algorithmic fairness, as well as responsible machine learning and data science.