# Improving reproducibility in AI research: four mechanisms adopted by JAIR

Gundersen, O.E.; Helmer, M.; Hoos, H.H.

Journal of Artificial Intelligence Research 81 (2024) 1019-1041 Submitted 07/2024; published 12/2024

# Improving Reproducibility in AI Research: Four Mechanisms Adopted by JAIR

**Odd Erik Gundersen** ODDERIK@NTNU.NO
*Norwegian University of Science and Technology*

**Malte Helmert** MALTE.HELMERT@UNIBAS.CH
*University of Basel*

**Holger Hoos** HH@AIM.RWTH-AACHEN.DE
*RWTH Aachen University & Universiteit Leiden*

## Abstract

**Background:** Lately, the reproducibility of scientific results has become an increasing worry in the scientific community. Several studies show that artificial intelligence research is not spared from reproducibility issues.

**Objectives:** As a pioneer in open and transparent research published on the Internet, the *Journal of Artificial Intelligence Research* (JAIR) seeks to promote good research practices and close the feedback loop between the original researchers and those reproducing their research.

**Methods:** Four different mechanisms will be adopted immediately by JAIR. These are: 1) reproducibility checklists, 2) structured abstracts, 3) reproducibility badges and 4) reproducibility reports.

**Results:** All authors submitting articles to JAIR fill out a reproducibility checklist and are encouraged to use structured abstracts. Articles that fulfill certain criteria will receive reproducibility badges, and reproducibility reports can be submitted by anyone for any article published in JAIR.

**Conclusions:** We believe that adopting the four mechanisms outlined in this paper will improve the reproducibility of research published in JAIR and thus make a contribution to addressing the broader reproducibility issue in artificial intelligence. We hope that JAIR's reproducibility initiative will inspire similar efforts at other top-tier journals.

## 1. Introduction

Recently, the research community has become increasingly aware of what is called the reproducibility crisis (Baker, 2016): the notion that a large portion of published research is false in the sense that its claims are not supported by evidence when experiments are repeated by third parties or even by the original authors themselves. While the premise of a crisis is disputed (Fanelli, 2018), several meta-studies in areas related to artificial intelligence (AI) indicate that AI researchers should not dismiss it out of hand (Lucic et al., 2018; Dacrema et al., 2019; Henderson et al., 2018; Reimers & Gurevych, 2017; Melis et al., 2018). There are several reasons for taking it seriously (Gundersen, 2020).

Although many counterexamples exist, such as theory and position papers, progress in AI is to a substantial degree driven by empirical research. Therefore, dealing with reproducibility challenges is of high importance to our field. All empirical work faces challenges in reproducing experiments and empirical findings. Ioannidis (2005) suggests that AI research

is more prone to research findings being false than physics, which is generally considered to have the fewest false findings. He places AI on par with psychology, at the other end of the spectrum, with a very high rate of false findings being reported, partly because the field of AI is relatively young and its research methodology less mature.

Ostensibly, reproducibility appears easier to achieve in computer science than in other empirical sciences because of the ability to completely specify and fully observe the algorithms and their implementations used in computational experiments. Unfortunately, the complexity of the hardware and software involved in empirical AI work at least partially negates this advantage, as discussed in more detail in Section 2. Moreover, AI research also includes facets of engineering and human-computer interaction that go beyond purely computational aspects (McGuffee, 2000; Comer et al., 1989). Therefore, AI inherits methodological challenges from both the harder and softer sciences.

Moreover, researchers conduct research in an environment that provides incentives that sometimes work against reproducibility. Researchers setting high standards for themselves must surmount additional obstacles. Transparency in the form of sharing code and data requires additional effort during and after publication, while holding back on code and data can yield a competitive advantage for follow-up research. Clearly expressing the expectations of what constitutes acceptable detail levels for research to be published will hopefully reduce differences in documentation diligence that is observed for example between academia and industry (Gundersen, 2019).

The publication culture in AI is strongly biased towards novel findings and supposed improvements over the state of the art, putting pressure on researchers to strive for and report on significant empirical improvements, sometimes at the expense of sound experimental methodology. Researchers that seek to reproduce and validate the work of others find it difficult to publish these efforts in reputable journals or high-profile conferences, and as a consequence reproducing published results is not a high priority in the AI community.

Reproducing published results is an important aspect of the scientific method, as it helps ensure that these results are correct. Moreover, scientific progress requires that researchers can build on existing work, and the openness and transparency that go hand in hand with reproducibility are key to making this possible. Ensuring true findings and high-quality research is not only the responsibility of individual researchers, but rather a shared responsibility of the community. Peer-reviewed scientific journals and conferences are the principal gateways that research must pass to be broadly disseminated. In order to facilitate the progress of science, they should provide mechanisms that encourage publishing reproducible results as well as reproducing the results of others. One such mechanism that has been widely adopted recently is to allow the sharing of code and data (Gundersen, 2020). JAIR has published code and data accompanying scientific articles since it was launched in 1993, but much more can be done to facilitate reproducibility.

We have decided to adopt four novel mechanisms for JAIR to help make AI research more reproducible: 1) reproducibility checklists, 2) structured abstracts, 3) reproducibility badges and 4) reproducibility reports. The goal of reproducibility checklists is to ensure that researchers provide the required information to make their research reproducible. They are intended to facilitate the work of the reviewers and make it easier for authors to follow good research practices. Structured abstracts seek to provide the readership of JAIR with a quick overview of published research and to thus make it more accessible. Reproducibility

badges help third parties locate research that is easy to reproduce and therefore, in many cases, also easier to build upon. (Of course, decisions on the desirability of building on a given piece of work depend on factors other than reproducibility.) They also provide a mark of distinction for authors that follow the principles of transparency and good experimental practices. Finally, reproducibility reports seek to close the feedback loop so that less effort is spent on invalid claims and introduce incentives for researchers to reproduce research.

The remainder of this article is structured as follows. Section 2 describes the challenges of reproducibility in more detail. Section 3 introduces reproducibility checklists as well as the intention behind them. Section 4 presents structured abstracts, followed by a discussion of reproducibility badges in Section 5. Section 6 introduces reproducibility reports, and finally in Section 7 we provide some general conclusions.

## 2. Reproducibility Challenges

All science is critically based on the notion of reproducibility, and all empirical science is based on the idea that the experiments designed to challenge hypotheses, their results, and the insights gained from these results can all be reproduced. This requirement of reproducibility gives rise to challenges, including the need to

- document experiments in sufficient detail to permit them to be replicated by others;
- have access to all resources (materials, infrastructure and expertise) required for carrying out a particular experiment;
- assess whether the results of an experiment can reasonably be considered to agree with those it is supposed to reproduce.

At least in principle, computational experiments, i.e., experiments conducted entirely by executing programs on computer hardware, offer advantages with respect to these challenges, since:

- the behaviour of algorithms and computer programs is, by definition and in fact, prescribed with mathematical precision;
- the execution environment of a computational experiment (i.e., the hardware and software stack on which it is run) is usually deterministic and under full control of the experimenter;
- the internal workings of software can be observed in full detail;
- experiments (along with the analysis of results) can be fully automated;
- computer hardware and software as well as data is often much more cheaply and readily available than the materials and infrastructure required for experiments in other disciplines.

However, upon closer examination, these advantages do not always fully apply, and empirical results in AI research can be difficult to reproduce, for a variety of reasons (see also Gundersen et al., 2022).

Firstly, many algorithms used in AI (and other areas of computer science) are randomised; this is notably the case for machine learning algorithms such as gradient boosted trees (Pouchard et al., 2020) and stochastic gradient descent (Reimers & Gurevych, 2017). When implemented and run on standard hardware, such algorithms almost always make use of pseudo-random number generators (PRNGs), and it has been demonstrated in some contexts that the behaviour thus obtained is indistinguishable from that resulting from the use of a true random number source (see, e.g., Hoos & Tompkins, 2006). While this does not preclude fully reproducible behaviour if precisely the same PRNG and random number seeds are used, other factors, including non-determinism due to parallel processing as well as interaction with virtual or physical environments, cannot be fully controlled. Furthermore, in many cases – especially when algorithms are run for a specific amount of time (e.g, fixed-time pre-processing in the case of AI planning, or fixed time budgets for automated configuration in the case of AutoML) – even small changes in the execution environment can impact the results of computational experiments. In any case, results produced by randomised methods or affected by other random factors must be evaluated with care, using appriopriately chosen statistical techniques such as hypothesis tests (Pham et al., 2020; Zhuang et al., 2022; Gundersen et al., 2023).

Secondly, AI software tends to be complex and typically exhibits complex behaviour. Even with full access to the inner workings of an algorithm or its implementation, it can be difficult to fully understand its functioning, e.g., to the point where performance on slightly modified input data can be predicted reliably (see, e.g., Hutter, Xu, Hoos, & Leyton-Brown, 2014). It is for this reason that performance measurements can be substantially impacted by the random seeds used in conjunction with pseudo-random number generators (Zhuang et al., 2022; Bouthillier et al., 2021), as well as by compiler settings and by the operating system (Hong et al., 2013). The use of parallel processing gives rise to additional challenges (see, e.g., Nagarajan et al., 2019; Pinto et al., 2021; Pham et al., 2020; Zhuang et al., 2022), as does the presence of unknown bugs (Gundersen et al., 2022). Furthermore, pseudo-code descriptions of algorithms, as are commonly found in publications, rarely capture sufficient detail to permit faithful reimplementation.

In addition, specifications of parameter settings are often missing; the execution environment in which experiments have been performed is described incompletely or not at all; and experiments are rarely fully automated, which introduces a risk of human error in manual steps, such as running scripts or transferring data for subsequent analysis.

Finally, experiments with AI algorithms and systems can be resource-intensive. This is especially the case in recent work in deep learning and AutoML. Even though such large-scale experiments may be reproducible in principle, in practice even those who conducted them originally may not be able to rerun them.

Additional challenges apply to a very common type of computational experiment, whose purpose is to provide evidence that a new algorithm or system advances the state of the art in solving a given problem. Arguably most serious among these is the risk of biased comparisons due to unfair optimisation of implementations or parameter settings, or due to cherry-picking of results. The former occurs when more effort is spent in optimising the implementation or parameters settings of an algorithm whose superiority is desired to be demonstrated than for the baselines against which it is compared. The latter refers to selectively reporting results from experiments that make one algorithm look better than

another, while neglecting to report the unfavourable outcomes of other experiments. Both of these issues can arise more easily in computer science than in other disciplines, since performance optimisation tends to be easier to achieve (manually or automatically), and since it is often also easier to perform relatively large numbers of experiments, once sufficient computational resources have been secured. Poor choice of baselines, benchmarking scenarios or evaluation metrics can also affect the validity of conclusions drawn from the results of empirical performance comparisons (Henderson et al., 2018; Makridakis et al., 2018).

Another fundamental challenge arises from the fact that standard computer hardware and software operates on discrete data, and real-valued operations or results are ultimately approximations. This implies that commonly used statistical methods that are based on continuous probability distributions – including statistical hypothesis tests, such as the two-sample t-test – are strictly speaking not applicable (Rardin & Uzsoy, 2001). Furthermore, since computational experiments often permit the use of large sample sizes, it is often possible to establish the statistical significance of small and practically irrelevant performance differences. This complicates the use of statistical hypothesis tests as a means for assessing reproducibility of empirical results.

Significant challenges arise when using data sets, especially in AI research involving the use of machine learning techniques. Here, data set bias (Torralba & Efros, 2011), as well as the specific techniques used for preprocessing (Dacrema et al., 2021), augmenting (Bouthillier et al., 2021), batching or splitting data sets (Makridakis et al., 2018) can all have substantial effects on empirical results.

Further challenges are encountered in the context of computational experiments that involve the simulation of complex physical (or virtual) environments. While the use of such simulations can greatly facilitate reproducibility, it usually gives rise to a simulation gap, i.e., a loss of transferrability of results from the simulated to the real environment. Challenges arising from such simulation gaps are well-known in robotics, where simulations of robot bodies and of the physical environment they interact with are widely used (see, e.g., Jakobi, Husbands, & Harvey, 1995). Similar challenges arise when simulating the behaviour of humans (individuals or groups) or of phenomena within the internet (see, e.g., Doersch & Zisserman, 2019; Ma et al., 2016; Huo et al., 2021).

Finally, although computational experiments play a very significant role in AI and other areas of computer science, the empirical pillar of the field remains relatively weakly developed compared to mature empirical sciences such as physics, medicine or psychology. This is reflected in the fact that formal training in empirical methods is largely absent from computer science curricula, and there are still no widely accepted standards and best practices for empirical work, although efforts in this direction have been ongoing for several decades (Cohen & Howe, 1988; Cohen, 1995). Recently, several workshops and tutorials have been organized on related topics, such as the AAAI 2020 Workshop on Evaluating Evaluation of AI Systems, the Reproducible AI workshops at AAAI in 2019 and 2020, and the tutorials on reproducibility at IJCAI 2019 and 2020. Nonetheless, well-known issues such as HARKing – hypothesising after results are known (see, e.g., Cockburn et al., 2020) – are still relatively common.

Evidently, meeting these challenges for a particular experiment or study requires knowledge and diligence by the scientists involved in it and will vary from case to case. However,

there are some mechanisms which we believe will encourage scientists to address reproducibility challenges arising in AI research, and which are likely to increase the reproducibility of AI research published in JAIR. In the following, we introduce and discuss four such mechanisms.

## 3. Reproducibility Checklist

A checklist is a sequence of steps to be carried out in order to ensure that a set of criteria or conditions crucial for accomplishing a complex task is satisfied. Checklists are used in many different domains to ensure adherence to formal requirements or best practices. They help to reduce the risk of forgetting to carry out important checks, of avoiding bias or error in making complex decisions, and of spuriously inferring a pattern or explanation one is looking for in a mess of data (Scriven, 2000).

Checklists were first introduced to aid pilots in flying increasingly complex airplanes (Gawande, 2010). Nowadays, they help reduce errors not only in aviation, but also in production, critical care (Hales & Pronovost, 2006), in the clinic (Charlesworth et al., 2013) and in the boardroom (for self-assessment of the board's performance) (Gill et al., 2005). Checklists help establish a higher standard of baseline performance, and they have been demonstrated to be useful in solving relatively simple and very complex problems (Gawande, 2010).

The computational experiments carried out in AI research are often complex. As a result, it can be challenging to reproduce the observations obtained from an experiment and difficult to ascertain that these observations support the conclusions drawn from them in a given publication. These challenges are known as verification and validation, respectively (Ivie & Thain, 2018); checklists can help authors in AI to address these issues and thus improve the reproducibility of their research.

As far as we know, reproducibility checklists have so far not been used by any journals in AI. They have, however, been introduced at most top-tier conferences in AI over the last few years, including Association for the Advancement of AI Conferences (AAAI), the International Joint Conferences on AI (IJCAI), the Conferences on Neural Information Processing Systems (NeurIPS), the International Conferences on Machine Learning (ICML), and the Conferences on Empirical Methods in Natural Language Processing (EMNLP). The International Conferences on Learning Representations (ICLR) has not introduced a checklist, but instead encourages authors to write reproducibility statements, in which authors should discuss the efforts that have been made to ensure reproducibility. A reproducibility statement provides flexibility for the authors and allows them to emphasise the parts that they believe to be most important when it comes to reproducing their research. However, the flexibility does not ensure that all parts generally regarded as important for reproducibility are described when authors choose what to emphasise. In this model, it is up to the individual reviewers to decide whether the reproducibility statement is of sufficient quality, which makes it difficult to achieve consistency. By using a checklist, it is easier to ensure adherence to a common set of best practices; furthermore, since the checklist explicitly outlines these practices, it becomes easier to discuss and, if necessary, revise or amend them.

One possible drawback of introducing checklists and other mechanisms to promote reproducibility comes from the additional burden placed on authors. Pineau et al. (2021) reported that after introducing reproducibility checklists at NeurIPS 2019, code sharing, which was voluntary but covered by the checklist, increased from less than 50% to nearly 75% compared to 2018. Furthermore, the additional effort associated with the checklist and the items covered in it did not appear to reduce the eagerness to submit papers to NeurIPS, as submissions increased by 40% compared to the previous year, although authors could have elected to submit to other top-tier venues where there were no expectations related to reproducibility. This suggests that the benefits associated with reproducibility checklists outweigh the potential drawbacks.

The reproducibility checklist that we now introduce for JAIR covers both theoretical and empirical research. It is based on the checklist developed for AAAI, which again was based on the recommendations presented in Gundersen et al. (2018). Furthermore, in developing it, we found useful guidance in Joelle Pineau's machine learning reproducibility checklist v2.0 (Pineau, 2020).

The checklist, shown in full in Appendix A, comprises four parts. The first part is related to methodology and covers claims, limitations and pseudo-code. It has to be completed for all submissions. The second part is only required for submissions that are partly or fully theoretical. It requires formal statements of assumptions, limitations and claims, as well as proofs and proof sketches of novel claims and results and citations to theoretical tools used. The third part is relevant for research articles that report on computational experiments. It covers whether code is made publicly available, whether the method for setting the seeds of methods that depend on randomness is described, whether the execution environment is clearly described, whether evaluation metrics are described, whether the number of algorithm runs that were used to compute the result have been specified, whether the analysis of results goes beyond the single-dimensional summaries and includes statistics on variation, whether all parameters are described and (where applicable) whether the process of finding hyperparameters is described. The fourth and final part is required for papers that rely on one or more datasets, and the items in this part of the list cover whether datasets are public, whether they are properly documented with metadata and analyses as well as how data is divided into training, validation and test sets if applicable.

Authors as well as reviewers are supposed to make use of the reproducibility checklist. Authors are required to complete it when submitting a paper, and reviewers should consider whether the paper covers the items in the checklist in sufficient detail to ensure that the research can be reproduced. It is important to emphasise that full adherence to the checklist is not a necessary requirement for an article to be accepted. However, it should be clear from reading the article why certain checklist items are not relevant, why data or code cannot be shared, or why specific details are not described. The reproducibility checklist will regularly be revisited and adapted as needed. This will be done under the leadership of the editor-in-chief and associate editor-in-chief, based on input elicited from the associate editors and editorial board members.

## 4. Structured Abstracts

A simple and relatively widely used way of nudging authors and readers towards paying attention to key components of solid empirical research (and of the scientific method underlying it) is the use of *structured abstracts*, i.e., the practice of providing and explicitly labelling somewhat standardised sections within the abstract of a publication. These sections prominently include the objectives of a given study, the methods used to achieve them, the results provided in this context, and their broader significance. All of these aspects are related, directly or indirectly, to the reproducibility of the work described in the publication.

In addition, structured abstracts have been shown to be more readable (Hartley & Sydes, 1997). They also make it easier to automatically process abstracts and hence facilitate future efforts in using AI tools within the scientific process (e.g., for literature search and analysis).

Structured abstracts are relatively widely used in other sciences. For example, many articles published in Science provide structured abstracts, and the same is true for other publication venues, such as *Nature Journal of Exposure Science & Environmental Epidemiology*, *Computer Methods and Programs in Biomedicine* and *IEEE Transactions on Biomedical Engineering* (see, e.g., Sears et al., 2022; Nistal-Nuño, 2022; Phan et al., 2022). There is also at least one instance in which a structured abstract has been used in a AAAI conference paper (Gundersen & Kjensmo, 2018). While the structure suggested or required to be used in these publication venues varies slightly and is not always used consistently, it generally includes the previously mentioned key components.

We believe that a substantial number of mostly empirical studies published in JAIR will benefit from the use of structured abstracts, and we therefore strongly encourage authors to adopt the following structure whenever possible and applicable:

- Background
- Objectives
- Methods
- Results
- Conclusions

The background section of the abstract should motivate the research and situate it relative to prior work. The objectives section should explicitly and concretely state what the research aims to achieve; for empirical work, this includes a crisp statement of what the computational experiments are designed to demonstrate. The methods section should outline, at a high level, the approach taken to achieve the objectives. The results section provides a brief summary of the key findings, and the conclusions situate these findings in a broader context, typically by commenting on their significance and impact beyond the work being presented in the article. To illustrate how a structured abstract could be written, we provide an example of one that we have made for an article that won the 2010 IJCAI-JAIR Best Paper Prize. It is shown in Figure 1.

We realise that structured abstracts may not be suitable for all types of work published in artificial intelligence. For example, mostly or purely theoretical contributions may require a different type of structure or be best summarised without using a prescribed structure.

**Background:** It has been widely observed that there is no single "dominant" SAT solver; instead, different solvers perform best on different instances.

**Objectives:** Rather than following the traditional approach of choosing the best solver for a given class of SAT instances, we aim to make this decision fully automatically, online and on a per-instance basis, with the goal of solving a broad range of SAT instances more efficiently in terms of running time.

**Methods:** We describe SATzilla, an automated approach for constructing per-instance algorithm portfolios for SAT that use so-called empirical hardness models to choose among their constituent solvers. This approach takes as input a distribution of problem instances and a set of component solvers, and constructs a portfolio optimizing a given objective function (such as mean running time, percent of instances solved, or score in a competition). In this article, we go well beyond earlier versions of SATzilla, by making the portfolio construction scalable and completely automated, and improving it by integrating local search solvers as candidate solvers, by predicting performance score instead of running time, and by using hierarchical hardness models that take into account different types of SAT instances.

**Results:** The excellent performance of SATzilla was independently verified in the 2007 SAT Competition, where our SATzilla07 solvers won three gold, one silver and one bronze medal. We demonstrate the effectiveness of the new techniques introduced here in extensive experimental results on data sets including instances from the most recent SAT competition.

**Conclusions:** The effectiveness of the SATzilla approach demonstrated in this article suggests that per-instance automated algorithm selection may also be possible for NP-hard problems other than SAT. We expect this to pave the way for achieving substantial improvements in the state of the art in solving other important problems in AI and beyond.

Figure 1: Example of a structured abstract that we made for L. Xu et al., F. Hutter, H. H. Hoos, K. Leyton-Brown: SATzilla – Portfolio-based Algorithm Selection for SAT. Journal of Artificial Intelligence Research 32: 565–606, 2008.

Nonetheless, we encourage authors to consider the use of structured abstracts, and reviewers as well as associate editors to suggest it when they feel it would be useful for a given piece of work. In general, structured abstracts should not be used in a mechanistic way, but should make it easier for readers to gain a clear understanding of the contributions made in an article and of the way in which the validity of these contributions has been established. Additional considerations apply with respect to the clarity and readability of abstracts. For example, it is generally desirable to minimise the use of abbreviations and to avoid including references in an abstract, regardless of whether the structure suggested here is followed.

## 5. Reproducibility Badges

In order to incentivise authors to strive towards more transparency and reproducibility and to make it easier for readers to find papers that follow these principles, JAIR has decided to introduce four types of *badges* related to reproducibility for articles published in the journal:

- *Open data:* All data used in the paper is shared in a public repository, following best practices for long-term accessibility. A problematic option is to host data on institutional webpages of the authors, which often cease to exist when the institutional relationship ends. A good option is to have the data hosted by JAIR itself as one or more online appendices. Another good option is to use a data repository that is specifically aimed at long-term archiving of research data such as Zenodo.

- *Open source:* All code used for experiments in the paper is shared in a public repository, following best practices for long-term accessibility as well as for source code readability and documentation. This includes both code developed by the authors as part of their scientific contribution and, as far as possible, code dependencies and code for related work that the paper compares to experimentally. For repositories that host more than one code version, the exact code version used in the paper is made clear.

- *Reproducible research:* The criteria for the open data and open source badges are satisfied, and all other aspects of the reproducibility checklist (Section 3) are addressed in a satisfactory manner.

- *Independently reproduced:* The findings of the paper have been independently reproduced by a third party to a satisfactory level, and the outcomes of this independent effort have been published in the form of a *reproducibility report*, as discussed in Section 6.

Badges are awarded to individual papers and are displayed alongside these papers on the JAIR website. *Open data* and *open source* badges are accompanied by links to the corresponding data/code, and *independently reproduced* badges are accompanied by links to reproducibility reports. The badges are a seal of approval that publicly acknowledges the authors' efforts towards reproducibility as well as the journal's efforts towards this goal. They are also a form of metadata that can help researchers and other readers decide whether they want to invest time in reading a paper. Ceteris paribus, researchers working on similar problems or intending to leverage the results of a JAIR article in their own work might value a contribution with the *open source* badge more highly because they know that they

will be able to use the software artifacts related to the paper. For papers with the *open data* badge, readers know that they are not limited to the authors' interpretations and summaries of experimental findings if they want to dig deeper. The *reproducible research* and *independently reproduced* badges combine these benefits and lift them to a higher level.

It is clear from the preceding discussions that reproducibility is not a binary property. Reproducibility has many facets, and also within each of these facets, reproducibility is usually measured on a graduated scale rather than being a matter of "yes" or "no". For example, for the *open source* badge, even if the authors make all of their own code available, it might depend on third-party code (such as operating systems like Linux, macOS or Windows; programming language environments such as Python or Java; or libraries such as TensorFlow or CPLEX) that may or may not have licenses that permit redistribution within the context of an academic paper. For the *reproducible research* badge, there are best practices that authors should be encouraged to implement to maximise the chance of successfully reproducing results (e.g., controlling dependencies on randomness or machine architecture as much as possible; limiting dependencies to easily accessible software with defined version numbers; using container technology to minimise the effort of reproducing an experiment environment), but we believe it is important to set the barrier for a reproducibility badge at an appropriate level: high enough to encourage authors to carefully consider questions of reproducibility and adapt their research practices over time in order to make their work as reproducible as possible, but not so high that it becomes practically insurmountable for a large fraction of papers.

This also implies that awarding reproducibility badges involves judgement calls. Tools such as the reproducibility checklist (Section 3) are helpful in guiding and streamlining this process, but do not replace careful deliberation. Rather, the checklist is a major input to the decision of whether or not a paper satisfies the criteria for reproducibility. A paper can be reproducible without ticking all the boxes. However, conversely, it should not happen that a paper "ticks all the boxes" of the checklist, but still does not count as reproducible – this would be an indication that the checklist needs revision.

While we believe that reproducibility badges provide substantial benefits for authors, readers and our research community as a whole, they come with a cost. Reproducibility needs to be evaluated carefully and fairly, which requires effort. Given the already very high demand for reviewing resources in our research community, this process should be lightweight to be accepted widely and to be seen as a positive improvement rather than an additional burden. The proposed model for JAIR is to integrate the awarding of the *open data*, *open source* and *reproducible research* badges into the regular review process. The reproducibility checklist plays a key role in this process. The authors use it to indicate their own view of the reproducibility of their research and the reviewers can use it as a guideline to either confirm the authors' claim of reproducibility, or to clearly communicate in their reviews where they disagree and revisions are needed before a reproducibility badge can be awarded. The ultimate decision which badges should be awarded to a paper is at the discretion of the associate editor in charge of the review process. While this model clearly adds some cost to the review process, we believe it allows for smooth integration into the existing review model, so that this additional cost is justified in light of the benefits of fostering reproducibility.

One consequence of integrating the awarding of the *open data*, *open source* and *reproducible research* badges into the regular review process is that these badges will not be awarded retroactively for papers submitted before these badges were introduced. While it would be desirable for past papers to have the same chance of being recognised for reproducibility as future papers, opening up the awarding of these badges to all past JAIR papers would come with a very steep cost, and any fixed past cutoff date would be as arbitrary as setting the cutoff date to "now". However, all past JAIR papers are eligible for the *independently reproduced* badge as described in the following section.

## 6. Reproducibility Reports

Not just dessert aficionados know that the proof of the pudding is in the eating. The crowning achievement of reproducible research is for someone to go ahead and independently reproduce it. This is routinely done, for example, for important empirical findings in physics, such as the discovery of the Higgs boson (see, e.g., Franklin, 2018). In our current research environment, there are few incentives for researchers to reproduce existing research rather than making a novel research contribution of their own, and the outcomes of such reproduction efforts are not very visible. Most conferences and journals focus on publishing *novel* research, and therefore reproducing someone else's research results can be a substantial amount of work with comparatively little reward. (However, it is worth pointing out that exceptions to the "novel research only" rule exist, including highly cited ones, for example in the form of survey articles. JAIR also publishes viewpoint articles in its AI & Society track.)

For this reason, JAIR is adding *reproducibility reports* to the classes of papers it publishes. A JAIR reproducibility report is a paper whose sole purpose it is to reproduce the findings of another JAIR paper, for example by providing and evaluating an independent implementation of an algorithm described in a published JAIR paper or independently reproducing experiments that another paper reported on.[1] In this context, the notion of reproducing a result can take various forms, as discussed later in this section in more detail. A reproducibility report describes the challenges and outcomes of a reproducibility effort and ultimately provides independent confirmation of the results of the original paper, or possibly shows that the results of the original paper cannot be reproduced.

Especially in the case of a failure to reproduce a result, this can be a somewhat delicate matter. Perhaps a result cannot be reproduced because a key detail was not mentioned in the original publication, but after some clarifications provided by the original authors,

1. The decision to limit the scope to reproducing work published in JAIR is the result of an extensive discussion because reproducibility reports would also be valuable for work published elsewhere. There are some advantages to limiting the scope to JAIR papers, some of which are practical in nature and others that are based on more fundamental considerations. JAIR has all information about the papers being reproduced including confidential aspects such as the identities of associate editors and reviewers, which makes the process of recruiting competent reviewers manageable. Most importantly, it allows JAIR to link papers and their reproducibility reports both ways. JAIR carries the responsibility for maintaining standards of scientific integrity for JAIR papers, but not for other venues. Authors will only be challenged regarding the reproducibility by the venue that published their paper, JAIR in this case, and not some other venue that does not have any relations to the published work. We hope for the initiative to set an example that other journals or conferences might follow. If this does not happen, it is possible to that JAIR may extend the scope to other venues in the future.

a confirmation of the original findings is easily possible. To minimise misunderstandings while also maximising the benefit to the community, authors of reproducibility reports are encouraged to contact the original authors where necessary to facilitate reproducing the results, but also to indicate in their reports where external help was needed to achieve the desired confirmation. Moreover, prior to publication of the reproducibility report, the original authors are given a chance to provide a response or comment that is then published alongside the report; the authors' response could provide clarifications or corrections of the original publication and possibly invalidate it if the authors have made mistakes.

Like other JAIR submissions, reproducibility reports undergo a peer review process with three possible outcomes: "accept for publication" (possibly with minor revisions), "reject with encouragement to resubmit" (after substantial revision), or "reject". Compared to regular submissions, this peer review process will often be more lightweight and might only involve a single reviewer, depending on the nature and depth of the work. Because the purpose of reproducibility reports is independent confirmation of results, none of the authors of a reproducibility report must be in any conflict-of-interest situation, as per the definition usually applied in the context of reviewing, with the authors of the paper whose results are attempted to be reproduced. Reproducibility reports are expected to follow the same standards with respect to clarity and technical correctness as other articles published in JAIR.

Accepted reproducibility reports are published as papers in their own right and also link to the papers they try to reproduce on the JAIR website. If the reproducibility effort is deemed successful, the original paper receives the *independently reproduced* badge. The process is as follows:

- **Select a JAIR article to reproduce:** Any article published in JAIR that reports on any experiments can be selected as a target for a reproducibility report. Also, theoretical results can be addressed in reproducibility reports, by filling non-trivial gaps or exposing flaws (potentially, but not necessarily along with providing fixes). While it might be particularly attractive to reproduce work for which there is evidence of broad impact, in principle there can also be significant value in selecting less prominent articles as targets for reproducibility reports, for example if they describe work of interest in the context of a specialised research area.

- **Reproduce results:** Conduct a reproducibility experiment (or equivalent for theoretical results) and document the steps taken to reproduce results as well as the results generated by the reproducibility effort.

- **Report results as a reproducibility report:** Reproducibility reports should follow the template presented below. Abstracts should follow the same structure.

- **Submit reproducibility report:** Reproducibility reports are submitted through JAIR's regular submission system with the article type set to "reproducibility report".

- **Review:** Whenever feasible, reproducibility reports are handled by the original associate editor and reviewed by at least one of the original reviewers.

- **Author response:** The associate editor contacts the original authors and gives them the opportunity for a response that will be published along with the reproducibility

report. Authors will be given access to the full reproducibility report and will be expected to respond within four weeks (unless specific other arrangements have been made with the associate editor).

- **Acceptance decision:** Acceptance decisions follow the standard process of JAIR. A reproducibility report is rejected if its technical quality or the quality of presentation does not meet the standards for publication in JAIR. The associate editor also makes sure that the reproducibility report and a possible response by the original authors are worded in a professional and respectful manner.

- **Publish:** Reproducibility reports are published in JAIR and linked from the web page of the original article.

A central aspect of the question of reproducibility is acknowledging that the scientific process is not infallible, and peer-reviewed scientific works can be incorrect. This observation also applies to the reproducibility reports themselves. It is possible, and in our view healthy, for the scientific process that a first reproducibility report concludes that a scientific work is reproducible, and a later second report comes to the opposite conclusion, or vice versa. It is therefore possible for an "independently reproduced" badge to be withdrawn in the light of new findings, although we expect this to be a rare occurrence.

Reproducibility reports have a uniform title of the form *Reproducibility Report for "Title of Paper to be Reproduced"*. The abstract and the report itself are structured as follows:

- **Scope of reproducibility:** Describe which research is the target of the reproducibility experiment. Refer to the paper that is being reproduced and also which claims or hypotheses are the targets of the reproducibility experiment.

- **Methodology:** Specify the type of reproducibility experiment that is described in the report (see also Gundersen, 2021). Does the reproducibility experiment rely on the paper alone, so that the code has to be re-implemented and new data collected; does it rely on the paper and the code, so that the same code is executed on new data; does it rely on the paper and the data, so that the code has to be re-implemented; or does the reproducibility experiment rely on the paper and executing the same code on the same data?

- **Results:** Present the results and discuss whether the claims or hypotheses that were the targets of the reproducibility experiment have been reproduced. This could be done by specifying the degree to which the results have been reproduced (Gundersen, 2021). If the reproducibility experiment produces the exact same output, the experiment is *outcome reproducible*. If the reproducibility experiment results in a different outcome, but the same set of analyses yield the same conclusion, then the experiment is *analysis reproducible*. Finally, if a different analysis ends up with the same conclusion, then the experiment is *interpretation reproducible*.

- **What was easy:** What did go smoothly when conducting the reproducibility experiment? Describe which parts of the experiment were easy to reproduce.

- **What was difficult:** Describe any issues that were encountered when conducting the reproducibility experiment. Were any parts of the experiment not covered in enough detail in the target article? Were any important unstated assumptions made? Describe any ambiguities and assumptions that you had to make because of these.

- **Communication with authors:** Did you have to communicate with the authors of the article that was the target of the reproducibility experiment? Did the communication contribute to the success of the reproducibility experiment?

The structure is based on the one used in the Machine Learning Reproducibility Challenge, which was first organised at ICLR 2019 (Pineau et al., 2019) and later the same year at NeurIPS 2019. Now, the reproducibility challenge is organised twice a year (Sinha et al., 2021). A more permanent solution than such challenges are fully-fledged reproducibility *tracks*, which have been added to conferences including SIGIR 2022, NAACL 2022, ECIR 2021 and ACM Recommender Systems 2022. Such tracks focus on reproducing past research published at the conference series and sometimes also articles discussing methodological issues. However, we are not aware of dedicated reproducibility tracks having yet been organised at top-tier AI conferences, such as AAAI, ICML, IJCAI and NeurIPS.

An important question arises in the context of how reproducibility reports should be published. For good or bad, automated bibliometric indicators, such as the various definitions of impact factor, cannot be ignored by scientific journals. Despite the utility of reproducibility reports for our research community, we expect that most of them will receive significantly fewer citations than typical JAIR papers reporting on or surveying original research. Hence, care must be taken to make sure that publishing reproducibility reports does not negatively impact JAIR's bibliometric indicators. Naturally, other journals also publish content other than original research and survey articles, which tends to get cited less, and have managed to establish ways to avoid dilution of key bibliometric indicators.

We considered the option of creating a new sister journal, "JAIR Reproducibility Reports", to avoid the dilution of JAIR's impact factor by reproducibility reports. However, such a solution also comes with drawbacks. In particular, publishing reproducibility reports in JAIR itself sends a clear signal that we recognise these reports as valuable scientific contributions in their own right. Therefore, we decided to publish reproducibility reports in JAIR itself rather than a new sister journal. We will monitor their impact on JAIR's bibliometrics and may revisit this decision in the future.

## 7. Conclusion

Reproducibility is of key importance to all sciences. Research in AI gives rise to specific reproducibility challenges, in addition to more general issues that also arise in other areas of computer science and beyond, as outlined earlier in this article. Reproducibility challenges encompass the design of computational experiments, the design and implementation of algorithms, the way empirical observations (such as performance measurements) are made, specifics of how data collected from experiments are evaluated, as well as the documentation of all aspects of empirical studies (see, e.g., Gundersen et al., 2022).

JAIR, as one of the most visible publication venues in AI, is adopting four concrete mechanisms to enhance the reproducibility of the research contributions published in the journal:

reproducibility checklists, structured abstracts, reproducibility badges and reproducibility reports. Reproducibility checklists encourage authors to adhere to widely accepted standards of reproducible research. Structured abstracts nudge authors and readers towards paying attention to key components of solid empirical research. Reproducibility badges incentivise the sharing of code and data, as well as the adoption of other measures that make it easier for others to reproduce research results published in JAIR. Finally, reproducibility reports document and incentivise serious attempts at independently reproducing results presented in JAIR articles.

Together, we hope that these measures will substantially increase the extent to which others can confidently build on results (especially of empirical nature) published in JAIR. Obviously, as we gain experience with these mechanisms, adjustments may be necessary, very much in line with the general principle in empirical science to revise current models and theories in light of new evidence. We therefore aim to review reproducibility challenges in AI research and the specific mechanisms used by JAIR to address them at least every two years, and to revise them as needed, based on discussions with the associate editors and members of the editorial board.

Nevertheless, it is important to note that methodologically sound research might still lead to incorrect findings even when fully reproducible. For example, a data set that does not properly represent the underlying phenomenon would lead to false conclusions every time it is used to analyse the phenomenon, unless maybe when combined with new data. Furthermore, empirical work in AI often aims to establish results in the form of universal statements, such as "algorithm A performs better than algorithm B on problem instances of type X", and with few exceptions, such statements typically cannot be verified by means of experiments. Instead, experiments should be designed as meaningful attempts at falsifying the claim (Popper, 1934). If the results of such experiments fail to invalidate the claim, this does not imply the truth of the statement, but merely increases our confidence that it might be true.

We believe that it is important for authors, reviewers and the AI community at large to keep in mind that scientific progress is not always served best by a narrow focus on beating baseline algorithms on a given set of benchmarks. While the merit of new ideas needs to be demonstrated, through thorough empirical or theoretical analysis, improvements over the state of the art of solving a challenging problem are neither the only, nor always the best way to make noteworthy contributions to AI research.

Progress in science depends on the rigorous investigation of bold claims, and publication venues such as JAIR have an important role to play in this context. In this spirit, we hope that JAIR's reproducibility initiative outlined in this article will increase the degree to which others can build on AI research published in JAIR, without discouraging the kind of bold speculation that advances science, in AI and beyond.

## Acknowledgments

It is JAIR's policy not to publish research articles by the editor-in-chief or associate editor-in-chief during their tenure to avoid conflicts of interest, and also because they have access to all reviewing information for the journal. This article involving the current editor-in-

chief (Malte Helmert) and editor-in-chief at the time it was first written (Holger Hoos) has an editorial character, as it concerns the running of the journal itself and is explicitly not a research article. The peer review process for this article was conducted in a non-anonymous fashion, by the associate editors of JAIR, and the final decision to publish the article was made by the current associate editor-in-chief (Chris Beck) and the managing director (Steven Minton) in agreement with JAIR's advisory board.

## Appendix A. Reproducibility Checklist for JAIR

This appendix lists all questions that articles submitted to JAIR have to answer in the submission process.

**All articles:**

1. All claims investigated in this work are clearly stated. [yes/partially/no]
2. Clear explanations are given how the work reported substantiates the claims. [yes/partially/no]
3. Limitations or technical assumptions are stated clearly and explicitly. [yes/partially/no]
4. Conceptual outlines and/or pseudo-code descriptions of the AI methods introduced in this work are provided, and important implementation details are discussed. [yes/partially/no/NA]
5. Motivation is provided for all design choices, including algorithms, implementation choices, parameters, data sets and experimental protocols beyond metrics. [yes/partially/no]

**Articles containing theoretical contributions:**

Does this paper make theoretical contributions? [yes/no]
If yes, please complete the list below.

1. All assumptions and restrictions are stated clearly and formally. [yes/partially/no]
2. All novel claims are stated formally (e.g., in theorem statements). [yes/partially/no]
3. Proofs of all non-trivial claims are provided in sufficient detail to permit verification by readers with a reasonable degree of expertise (e.g., that expected from a PhD candidate in the same area of AI). [yes/partially/no]
4. Complex formalism, such as definitions or proofs, is motivated and explained clearly. [yes/partially/no]
5. The use of mathematical notation and formalism serves the purpose of enhancing clarity and precision; gratuitous use of mathematical formalism (i.e., use that does not enhance clarity or precision) is avoided. [yes/partially/no]
6. Appropriate citations are given for all non-trivial theoretical tools and techniques. [yes/partially/no]

**Articles reporting on computational experiments:**

Does this paper include computational experiments? [yes/no]
If yes, please complete the list below.

1. All source code required for conducting experiments is included in an online appendix or will be made publicly available upon publication of the paper. The online appendix follows best practices for source code readability and documentation as well as for long-term accessibility. [yes/partially/no]

2. The source code comes with a license that allows free usage for reproducibility purposes. [yes/partially/no]

3. The source code comes with a license that allows free usage for research purposes in general. [yes/partially/no]

4. Raw, unaggregated data from all experiments is included in an online appendix or will be made publicly available upon publication of the paper. The online appendix follows best practices for long-term accessibility. [yes/partially/no]

5. The unaggregated data comes with a license that allows free usage for reproducibility purposes. [yes/partially/no]

6. The unaggregated data comes with a license that allows free usage for research purposes in general. [yes/partially/no]

7. If an algorithm depends on randomness, then the method used for generating random numbers and for setting seeds is described in a way sufficient to allow replication of results. [yes/partially/no/NA]

8. The execution environment for experiments, i.e., the computing infrastructure (hardware and software) used for running them, is described, including GPU/CPU makes and models; amount of memory (cache and RAM); make and version of operating system; names and versions of relevant software libraries and frameworks. [yes/partially/no]

9. The evaluation metrics used in experiments are clearly explained and their choice is explicitly motivated. [yes/partially/no]

10. The number of algorithm runs used to compute each result is reported. [yes/no]

11. Reported results have not been "cherry-picked" by silently ignoring unsuccessful or unsatisfactory experiments. [yes/partially/no]

12. Analysis of results goes beyond single-dimensional summaries of performance (e.g., average, median) to include measures of variation, confidence, or other distributional information. [yes/no]

13. All (hyper-) parameter settings for the algorithms/methods used in experiments have been reported, along with the rationale or method for determining them. [yes/partially/no/NA]

14. The number and range of (hyper-) parameter settings explored prior to conducting final experiments have been indicated, along with the effort spent on (hyper-) parameter optimisation. [yes/partially/no/NA]

15. Appropriately chosen statistical hypothesis tests are used to establish statistical significance in the presence of noise effects. [yes/partially/no/NA]

**Articles using data sets:**

Does this work rely on one or more data sets (possibly obtained from a benchmark generator or similar software artifact)? [yes/no]

If yes, please complete the list below.

1. All newly introduced data sets are included in an online appendix or will be made publicly available upon publication of the paper. The online appendix follows best practices for long-term accessibility with a license that allows free usage for research purposes. [yes/partially/no/NA]

2. The newly introduced data set comes with a license that allows free usage for reproducibility purposes. [yes/partially/no]

3. The newly introduced data set comes with a license that allows free usage for research purposes in general. [yes/partially/no]

4. All data sets drawn from the literature or other public sources (potentially including authors' own previously published work) are accompanied by appropriate citations. [yes/no/NA]

5. All data sets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. [yes/partially/no/NA]

6. All new data sets and data sets that are not publicly available are described in detail, including relevant statistics, the data collection process and annotation process if relevant. [yes/partially/no/NA]

7. All methods used for preprocessing, augmenting, batching or splitting data sets (e.g., in the context of hold-out or cross-validation) are described in detail. [yes/partially/no/NA]

**Explanations on any of the answers above (optional):**

[Text here; please keep this brief.]

## References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., et al. (2021). Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, *3*.

Charlesworth, G., Burnell, K., Hoe, J., Orrell, M., & Russell, I. (2013). Acceptance checklist for clinical effectiveness pilot trials: a systematic approach. *BMC medical research methodology*, *13*(1), 1–7.

Cockburn, A., Dragicevic, P., Besançon, L., & Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, *63*(8), 70–79.

Cohen, P. R. (1995). *Empirical methods for artificial intelligence*, Vol. 139. MIT press Cambridge, MA.

Cohen, P. R., & Howe, A. E. (1988). How evaluation guides AI research: The message still counts more than the medium. *AI magazine*, *9*(4), 35–35.

Comer, D. E., Gries, D., Mulder, M. C., Tucker, A., Turner, A. J., & Young, P. R. (1989). Computing as a discipline. *Communications of the ACM*, *32*(1), 9–23.

Dacrema, M. F., Boglio, S., Cremonesi, P., & Jannach, D. (2021). A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)*, *39*(2), 1–49.

Dacrema, M. F., Cremonesi, P., & Jannach, D. (2019). Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109.

Doersch, C., & Zisserman, A. (2019). Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in Neural Information Processing Systems*, *32*, 12929—12941.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to?. *Proceedings of the National Academy of Sciences*, *115*(11), 2628–2631.

Franklin, A. (2018). *Is It the 'Same' Result: Replication in Physics*. 2053-2571. Morgan & Claypool Publishers.

Gawande, A. (2010). *The Checklist Manifesto: How to get Things Right*. Picador.

Gill, M., Flynn, R. J., & Reissing, E. (2005). The governance self-assessment checklist: An instrument for assessing board effectiveness. *Nonprofit Management and Leadership*, *15*(3), 271–294.

Gundersen, O. E. (2019). Standing on the feet of giants—reproducibility in ai. *Ai Magazine*, *40*(4), 9–23.

Gundersen, O. E. (2020). The reproducibility crisis is real. *AI Magazine*, *41*(3), 103–106.

Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A*, *379*(2197), 20200210.

Gundersen, O. E., Coakley, K., Kirkpatrick, C., & Gil, Y. (2022). Sources of irreproducibility in machine learning: A review. *arXiv preprint arXiv:2204.07610*.

Gundersen, O. E., Gil, Y., & Aha, D. W. (2018). On reproducible ai: Towards reproducible research, open science, and digital scholarship in ai publications. *AI Magazine*, *39*(3), 56–68.

Gundersen, O. E., & Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence.. In *AAAI*, pp. 1644–1651.

Gundersen, O. E., Shamsaliei, S., & Isdahl, R. J. (2022). Do machine learning platforms provide out-of-the-box reproducibility?. *Future Generation Computer Systems*, *126*, 34–47.

Gundersen, O. E., Shamsaliei, S., Kjærnli, H. S., & Langseth, H. (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*, pp. 37–61.

Hales, B. M., & Pronovost, P. J. (2006). The checklist—a tool for error management and performance improvement. *Journal of critical care*, *21*(3), 231–235.

Hartley, J., & Sydes, M. (1997). Are structured abstracts easier to read than traditional ones?. *Journal of research in reading*, *20*(2), 122–136.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Hong, S.-Y., Koo, M.-S., Jang, J., Kim, J.-E. E., Park, H., Joh, M.-S., Kang, J.-H., & Oh, T.-J. (2013). An evaluation of the software system dependency of a global atmospheric model. *Monthly Weather Review*, *141*(11), 4165–4172.

Hoos, H., & Tompkins, D. (2006). On the quality and quantity of random decisions in stochastic local search for SAT. In *Advances in Artificial Intelligence, 19th Conference of the Canadian Society for Computational Studies of Intelligence*, Vol. 4013 of *Lecture Notes in Artificial Intelligence (LNAI)*, pp. 146–158. Springer, Heidelberg, Germany.

Huo, L., Jiang, D., Qi, S., Song, H., & Miao, L. (2021). An ai-based adaptive cognitive modeling and measurement method of network traffic for eis. *Mobile Networks and Applications*, *26*(2), 575–585.

Hutter, F., Xu, L., Hoos, H., & Leyton-Brown, K. (2014). Algorithm runtime prediction: Methods & evaluation. *Artificial Intelligence*, *206*, 79–111.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, *2*(8), e124.

Ivie, P., & Thain, D. (2018). Reproducibility in scientific computing. *ACM Computing Surveys (CSUR)*, *51*(3), 1–36.

Jakobi, N., Husbands, P., & Harvey, I. (1995). Noise and the reality gap: The use of simulation in evolutionary robotics. In *European Conference on Artificial Life*, pp. 704–720. Springer.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2018). Are GANs created equal? a large-scale study. In *Advances in neural information processing systems*, pp. 700–709.

Ma, Y., Lee, E. W. M., & Yuen, R. K. K. (2016). An artificial intelligence-based approach for simulating pedestrian movement. *IEEE Transactions on Intelligent Transportation Systems*, *17*(11), 3159–3170.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889.

McGuffee, J. W. (2000). Defining computer science. *ACM SIGCSE Bulletin*, *32*(2), 74–76.

Melis, G., Dyer, C., & Blunsom, P. (2018). On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*.

Nagarajan, P., Warnell, G., & Stone, P. (2019). Deterministic implementations for reproducibility in deep reinforcement learning..

Nistal-Nuño, B. (2022). Developing machine learning models for prediction of mortality in the medical intensive care unit. *Computer Methods and Programs in Biomedicine*, *216*, 106663.

Pham, H. V., Qian, S., Wang, J., Lutellier, T., Rosenthal, J., Tan, L., Yu, Y., & Nagappan, N. (2020). Problems and opportunities in training deep learning software systems: An analysis of variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, pp. 771–783.

Phan, H., Mikkelsen, K. B., Chen, O., Koch, P., Mertins, A., & De Vos, M. (2022). Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 1–1.

Pineau, J. (2020). The Machine Learning Reproducibility Checklist (v2.0). `https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf`. Accessed: 2022-11-26.

Pineau, J., Sinha, K., Fried, G., Ke, R. N., & Larochelle, H. (2019). ICLR Reproducibility Challenge 2019. *ReScience C*, *5*(2), 5.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *Journal of Machine Learning Research*, *22*.

Pinto, W. G., Alguacil, A., & Bauerheim, M. (2021). On the reproducibility of fully convolutional neural networks for modeling time-space evolving physical systems. *arXiv preprint arXiv:2105.05482*.

Popper, K. (1934). *Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft*. Julius Springer.

Pouchard, L., Lin, Y., & Van Dam, H. (2020). Replicating machine learning experiments in materials science. In *Parallel Computing: Technology Trends*, pp. 743–755. IOS Press.

Rardin, R. L., & Uzsoy, R. (2001). Experimental evaluation of heuristic optimization algorithms: A tutorial. *Journal of Heuristics*, *7*, 261–304.

Reimers, N., & Gurevych, I. (2017). Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348.

Scriven, M. (2000). The logic and methodology of checklists..

Sears, C. G., Lanphear, B. P., Xu, Y., Chen, A., Yolton, K., & Braun, J. M. (2022). Identifying periods of heightened susceptibility to lead exposure in relation to behavioral problems. *Journal of exposure science & environmental epidemiology*, *32*(1), 1–9.

Sinha, K., Dodge, J., Luccioni, S., Forde, J., Raparthy, S. C., Mercier, F., Pineau, J., & Stojnic, R. (2021). ML Reproducibility Challenge 2021. `https://paperswithcode.com/rc2021`. Accessed: 2022-06-19.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE.

Zhuang, D., Zhang, X., Song, S., & Hooker, S. (2022). Randomness in neural network training: Characterizing the impact of tooling. *Proceedings of Machine Learning and Systems*, *4*.