# Reproducibility, Benchmarking and Scientific Standards in Artificial Intelligence and Robotics

Adrian Mercer, Leila Rahman, Daniel Okafor, Sofia Petrov

Article type: Editorial

Volume 1, Issue 1 (2026)

Read Current Issue →


## Abstract

Rapid advances in artificial intelligence and robotics are scientifically meaningful only when reported results can be independently understood, evaluated, compared, and, where feasible, reproduced. Yet contemporary research remains vulnerable to undocumented preprocessing, inaccessible datasets, hidden hyperparameter choices, random-seed sensitivity, weak baselines, benchmark overfitting, selective reporting, incomplete statistical analysis, simulation-to-reality gaps, and insufficient disclosure of computational resources. This editorial argues that reproducibility should be treated as a foundational condition of cumulative progress rather than an optional feature of exemplary research. Reproducibility, replicability, repeatability, robustness, and generalizability describe related but distinct dimensions of scientific reliability, each requiring appropriate evidence. The field should move toward transparent experimental reporting, standardized benchmark protocols, multiple-run evaluation, uncertainty quantification, stronger baseline comparisons, systematic ablation, accessible code and data where feasible, computational-resource disclosure, external validation, and real-world robotic testing. Journals, conferences, reviewers, and research institutions also shape these practices through incentives and publication standards. Scientific progress in intelligent systems should ultimately be measured not only by higher benchmark scores, but by whether claims remain credible when independently examined.


**Keywords:** reproducibility; benchmarking; artificial intelligence; robotics; scientific standards; open science; experimental methodology

## Citation

Mercer, A., Rahman, L., Okafor, D., & Petrov, S. (2026). Reproducibility, Benchmarking and Scientific Standards in Artificial Intelligence and Robotics. Journal of AI, Automation & Intelligent Systems, 1(1). https://eldranacademic.com/jaais/articles/4/

| Received | Revised | Accepted | Published |
| --- | --- | --- | --- |
| 16 March 2026 | 12 June 2026 | 24 July 2026 | 13 August 2026 |


### Copyright

© 2026 Adrian Mercer, Leila Rahman, Daniel Okafor, Sofia Petrov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) ↗. Use, distribution or reproduction in other forums is permitted, provided the original authors and copyright owners are credited and the original publication in this journal is cited in accordance with accepted academic practice.

**Disclaimer**

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, the publisher, the editors or the reviewers. Any product evaluated in this article, or claim made by its manufacturer, is not guaranteed or endorsed by the publisher.

---

# 1.0 Introduction

## 1.1 Context and Problem Framing

Artificial intelligence, machine learning, computer vision, reinforcement learning, autonomous systems, and robotics are advancing at a pace that places unusual pressure on conventional mechanisms of scientific verification. New models, benchmarks, datasets, simulation environments, software libraries, accelerators, and training procedures appear continuously, often accompanied by comparatively small performance gains that are interpreted as evidence of methodological progress. Yet scientific progress requires more than novelty or a higher score. It requires sufficient transparency for other researchers to understand what was done, determine whether comparisons are fair, and evaluate whether the reported result is stable under reasonable repetition. Concerns about these conditions are not hypothetical. A survey of 400 papers from major artificial-intelligence conferences found substantial deficiencies in documentation needed for reproducibility (Gundersen & Kjensmo, 2018), while independent reimplementation attempts have shown that apparently minor reporting omissions can determine whether published machine-learning work can be reproduced in practice (Raff, 2019). Community initiatives such as the NeurIPS reproducibility program have consequently treated transparency, code release, and structured reporting as central to research quality rather than peripheral administrative requirements (Pineau et al., 2021).

Terminological precision is important because several related concepts are frequently conflated. Following the convention adopted by the National Academies of Sciences, Engineering, and Medicine (2019), reproducibility can be understood as obtaining consistent computational results using the same data, code, methods, and analytical conditions, whereas replicability involves examining the same scientific question through new data or new experiments. Repeatability may be used more narrowly to describe repeated measurements or experimental runs under closely matched conditions. Robustness concerns whether conclusions persist when reasonable methodological assumptions or conditions vary, while generalizability concerns whether findings extend to different data, settings, populations, hardware, environments, or tasks. These distinctions matter in AI and robotics because a result may be computationally reproducible yet fail to generalize beyond a benchmark, or may appear repeatable in simulation while proving unreliable on physical hardware. Reproducibility is therefore necessary for scientific scrutiny but does not by itself establish validity.

## 1.2 Why the Issue Matters

The unresolved problem is that many contemporary research incentives reward visible performance improvements more directly than careful evidence about how those improvements were produced. Benchmark competitions encourage comparison, but they can also promote incremental optimization against repeatedly reused test sets. Complex software stacks can conceal preprocessing decisions and dependency assumptions; large computational budgets can make independent confirmation prohibitively expensive; and robotic experiments introduce additional sources of variation through hardware, sensors, calibration, environments, and human operation. Concerns about unclear attribution of empirical gains and weak methodological discipline have consequently accompanied rapid machine-learning progress (Lipton & Steinhardt, 2019). The central position advanced here is that technological progress becomes durable scientific knowledge only when the associated claims are sufficiently

transparent, statistically credible, fairly benchmarked, and open to independent evaluation. Reproducibility should therefore be treated not as a demand that every experiment produce identical numerical outcomes, but as an evidentiary framework for determining which conclusions deserve to accumulate into the scientific record.

## 2.0 Current Landscape

### 2.1 Benchmark Culture

The reproducibility landscape in artificial intelligence has improved substantially, although practice remains uneven. Early analyses emphasized that AI publications frequently omitted information required to recreate experiments and proposed structured checklists covering experimental setup, data, methods, software, and computational procedures (Gundersen et al., 2018). Open research infrastructures have also made datasets, tasks, experimental metadata, and results easier to share and compare; OpenML, for example, was explicitly designed to support networked and reusable machine-learning experimentation (Vanschoren et al., 2013). More recently, reproducibility checklists, voluntary code submission, artifact evaluation, and dedicated reproducibility challenges have become increasingly visible in major machine-learning venues (Pineau et al., 2021). A large replication study reported in 2025 further found a strong relationship between the availability of code and data and successful reproduction, reinforcing the practical value of open scientific artifacts (Gundersen et al., 2025).

### 2.2 Datasets

Experimental reporting has also received greater methodological scrutiny. Test-set performance alone can conceal the role of hyperparameter-search budgets, validation choices, and computational expenditure; Dodge et al. (2019) demonstrated that conclusions about comparative model quality can change when the resources devoted to model selection differ. Random initialization and hyperparameter variation introduce additional uncertainty. Reimers and Gurevych (2017), after evaluating tens of thousands of recurrent-network configurations, showed the importance of reporting score distributions rather than isolated best-performing runs. Bouthillier et al. (2021) similarly demonstrated that variation from data sampling, initialization, and hyperparameter choices can materially affect benchmark comparisons. Statistical testing is another unevenly applied component: Dror et al. (2018) found that statistical significance testing in natural-language-processing research was frequently absent or incorrectly used.

### 2.3 Software and Model Sharing

Benchmarking itself has become both a strength and a vulnerability of modern AI research. Shared datasets and evaluation protocols facilitate rapid comparison, but methodological weaknesses may remain hidden when communities focus excessively on leaderboard positions. Dacrema et al. (2019), for example, reproduced a subset of recent neural recommender systems and found that several reported advances were not robust to comparison with appropriately tuned simpler baselines. Musgrave et al. (2020) reported analogous concerns in deep metric learning, where improved evaluation protocols substantially reduced the apparent magnitude of progress. Such studies illustrate why benchmarking should be treated as experimental measurement rather than competition alone.

### 2.4 Evaluation Practices

Robotics introduces still greater complexity because physical experiments cannot always be duplicated through software artifacts. Real robots vary through mechanical wear, actuator behaviour, calibration, sensing noise, timing, laboratory geometry, and environmental conditions. Real-world reinforcement-learning benchmarking has demonstrated both strong hyperparameter sensitivity and difficulty transferring comparative rankings among tasks (Mahmood et al., 2018). Frameworks such as Eurobench seek standardized experimental protocols and scoring

mechanisms for physical robotic systems (Remazeilles et al., 2022), while contemporary robotics scholarship increasingly argues that reproducible methodology and objective comparison should be designed into experimental systems from the outset (Bonsignorio et al., 2025).

## 3.0 Critical Challenges

### 3.1 Data Leakage

Hidden implementation details remain among the most persistent obstacles to reproduction. Preprocessing pipelines, data exclusions, augmentation settings, stopping criteria, optimizer details, initialization procedures, software versions, and undocumented code modifications may materially affect outcomes. Reinforcement-learning research provides a particularly clear example because algorithmic stochasticity, environment randomness, reward scaling, and hyperparameter choices can produce markedly different results from ostensibly identical algorithms (Henderson et al., 2018). A single random seed can therefore provide a repeatable run without providing convincing evidence about expected performance. Multiple independent runs and explicit reporting of variance are necessary when stochasticity is material to the claim.

Data problems can be equally consequential. Leakage between training, validation, and evaluation procedures can produce severely overoptimistic conclusions, and a cross-disciplinary analysis identified leakage-related problems across hundreds of machine-learning studies (Kapoor & Narayanan, 2023). Benchmark labels themselves are not infallible. Northcutt et al. (2021) identified substantial label errors across widely used benchmark test sets and demonstrated that such errors can alter model rankings. Dataset reuse also creates the possibility that repeated optimization gradually adapts a research community to idiosyncrasies of a supposedly held-out test set. Recht et al. (2019) showed that models evaluated on newly collected ImageNet and CIFAR-10 test sets experienced notable accuracy reductions even when collection procedures were designed to resemble the originals. Dataset bias and cross-dataset generalization concerns had already been documented much earlier in computer vision (Torralba & Efros, 2011).

### 3.2 Benchmark Saturation

Benchmark culture creates a further methodological problem when a small number of datasets become proxies for entire research questions. Koch et al. (2021) found increasing concentration of dataset use within machine-learning subcommunities, raising questions about whether leaderboard progress corresponds to broader scientific generalization. Weak baselines compound this effect. When proposed methods are compared with outdated implementations, poorly tuned alternatives, or non-equivalent computational budgets, an apparent algorithmic improvement may instead reflect differences in engineering attention. This has been demonstrated empirically in recommender systems and metric learning (Dacrema et al., 2019; Musgrave et al., 2020). Ablation studies are therefore essential because they help identify which component actually produces an observed gain.

### 3.3 Incomplete Compute Reporting

Robotics adds the unresolved problem of experimental scale. Physical trials are expensive, slow, mechanically demanding, and difficult to standardize. Consequently, conclusions may be drawn from small numbers of trials or from simulations that remove many real-world sources of uncertainty. Simulation calibration research shows that transferring findings from simulation to physical systems requires explicit attention to the assumptions that create the simulation-to-reality gap (Mehta et al., 2021). Swarm-robotics evidence likewise demonstrates that simulation models differ substantially in their ability to predict real-world performance (Ligot & Birattari, 2022). Recent work

such as SIMPLER suggests that carefully constructed simulation can provide scalable and reproducible evaluation correlated with real robotic behaviour, but it also underscores the need to validate simulated benchmarks against physical evidence rather than assuming equivalence (Li et al., 2025).

### 3.4 Replication and Negative Results

Computational access presents another challenge. Hardware, software frameworks, numerical libraries, drivers, and dependencies evolve rapidly. Some methods depend on specialized accelerators or infrastructures unavailable to many research groups, creating what Hooker (2021) characterizes as a hardware lottery in which particular research directions become advantaged by prevailing computational architectures. Computational expense also affects reproducibility and fairness: large hyperparameter searches may yield performance gains inaccessible to researchers unable to duplicate equivalent search budgets. Energy and resource costs are therefore scientific reporting issues, not merely operational concerns (Strubell et al., 2019).

## 4.0 Discussion and Priorities

### 4.1 Integrative Discussion

Reproducibility problems in AI and robotics are not produced by a single methodological weakness. They emerge from interactions among publication incentives, benchmark culture, software complexity, computational inequality, proprietary resources, and rapidly changing technical environments. A research ecosystem that rewards marginal state-of-the-art improvements creates incentives to report the strongest run, choose favorable metrics, minimize negative outcomes, and devote less manuscript space to implementation details that do not appear novel. Lipton and Steinhardt (2019) argue that unclear attribution of empirical gains can compromise the intellectual foundations of machine-learning scholarship; reproducibility failures are one manifestation of this broader problem.

Benchmarks intensify these incentives because they convert multidimensional research performance into easily communicated rankings. Such simplification is useful for coordination, but a benchmark score can absorb effects from dataset curation, preprocessing, compute budget, software implementation, hardware, random initialization, and evaluation design. Dataset concentration then makes those scores disproportionately influential in determining what the field regards as progress (Koch et al., 2021). The scientific question should therefore remain larger than the benchmark used to approximate it.

Resource disclosure is similarly intertwined with scientific comparison. Reporting a new model without documenting the compute required to discover, tune, and train it can make apparently fair comparisons misleading. Green AI arguments have consequently called for efficiency and computational-resource reporting alongside predictive performance (Schwartz et al., 2020), while work on the environmental and financial costs of large-scale model development demonstrates how resource intensity can narrow participation in research (Strubell et al., 2019). Reproducibility is therefore partly an accessibility problem: theoretically available code is of limited value when execution requires unavailable hardware, proprietary data, inaccessible APIs, or prohibitive computation.

Open science can mitigate some of these constraints but cannot resolve them automatically. Code availability without dependency information, dataset provenance, version control, or execution instructions may still leave substantial ambiguity. Conversely, some data and systems cannot be released because of privacy, security, commercial, or legal obligations. The relevant standard is therefore sufficient transparency to permit credible evaluation, not indiscriminate openness. Encouragingly, recent replication evidence indicates that shared code and

data materially improve successful reproduction rates (Gundersen et al., 2025). The objective should be a research culture in which claims carry an explicit evidentiary trail linking data, implementation, statistical analysis, computational resources, and evaluation conditions.

Complementary scholarship informing this synthesis includes Gundersen (2018); and National Academies of Sciences (2019). These works broaden the evidentiary base considered across the field.

### 4.2 Research and Practice Priorities

A first priority is standardized methodological reporting. AI and robotics papers should disclose preprocessing, dataset partitions, hyperparameter-search procedures, initialization, stopping criteria, random seeds, software and hardware environments, evaluation metrics, and computational budgets. Consensus-based frameworks such as REFORMS provide a useful basis for strengthening reporting in machine-learning-based science (Kapoor et al., 2024). Reporting requirements should be proportionate to the complexity and consequences of the claim but sufficiently detailed for independent assessment.

Second, multiple-run and uncertainty-aware reporting should replace isolated best-run results wherever stochasticity is material. Means, dispersion measures, confidence intervals, or other appropriate uncertainty estimates should accompany comparative claims. Bouthillier et al. (2021) show why benchmark variance must be incorporated into model comparison, while Dror et al. (2018) demonstrate the importance of selecting statistical procedures appropriate to the evaluation structure.

Third, benchmark protocols should become more explicit and resistant to hidden adaptation. Training, validation, and final evaluation boundaries should be clearly specified; test sets should not serve as iterative development targets; and external or refreshed evaluation datasets should be incorporated when benchmark saturation becomes plausible. Strong baselines must be implemented and tuned competently, with equivalent computational assumptions wherever possible. Ablation analysis should identify whether reported improvements arise from the proposed scientific contribution or from auxiliary engineering choices.

Fourth, code, data, and executable environments should be made available where ethically, legally, and technically feasible. Repositories should include versioned dependencies, configuration files, preprocessing scripts, evaluation code, and sufficient instructions to reproduce central claims. Where data cannot be shared, authors should provide synthetic examples, detailed provenance, secure-access mechanisms, or other reasonable paths to evaluation. The NeurIPS reproducibility experience indicates that structured artifact expectations can improve research practice without making complete openness a prerequisite for every study (Pineau et al., 2021).

Fifth, preregistration should be considered for confirmatory studies, major benchmark comparisons, and replication projects where analytical flexibility creates substantial risk of selective reporting. It is not appropriate for every exploratory engineering experiment, but explicit distinction between exploratory development and confirmatory testing can reduce hindsight-driven hypothesis construction and benchmark tuning.

Sixth, robotics requires stronger physical validation standards. Simulation results should report simulator configuration, physics parameters, randomization procedures, and calibration evidence, while claims about real-world robotic performance should include sufficient independent trials, environmental variation, failure reporting, and hardware descriptions. Standardized scoring infrastructures such as Eurobench demonstrate how common protocols can support comparison across systems (Remazeilles et al., 2022). Recent calls for reproducible robotics similarly emphasize objective comparison and statistically meaningful experimental repetition (Bonsignorio et al., 2025).

Finally, replication and scientifically informative negative results should acquire greater publication value. Failed reproductions can expose undocumented assumptions; negative comparisons can prevent redundant research; and confirmation studies can establish whether influential findings are robust. Scientific maturity requires incentives

for verification as well as novelty.

### 4.3 Implications for the Journal's Scope

For the Journal of AI, Automation & Intelligent Systems (JAAIS), methodological transparency should be treated as part of scientific contribution rather than as supplementary presentation. Manuscripts reporting machine-learning, robotics, automation, computer-vision, autonomous-system, or intelligent-control results should provide enough information for reviewers and readers to understand precisely how the evidence was generated and what limitations constrain the claim.

JAAIS should encourage reproducible engineering studies, benchmark analyses, replication research, external validation, negative results of clear scientific value, standardized robotic evaluation, and studies that investigate the reliability of commonly used methodologies. The journal should also welcome contributions that improve experimental infrastructure, open benchmarking, statistical evaluation, dataset documentation, simulator validation, and reproducible software workflows. Where artifacts cannot be released, authors should explain the restriction and provide the maximum feasible methodological detail.

The standard should not be identical numerical duplication in every case. Robotics, stochastic learning, and adaptive systems inherently produce variation. The appropriate expectation is that methods are transparent, uncertainty is reported, comparisons are fair, and conclusions remain defensible under independent scrutiny. That standard is consistent with the broader movement toward reproducible AI and robotics research (Bonsignorio et al., 2025; Pineau et al., 2021).

### 4.4 Conclusion

Artificial intelligence and robotics progress through experimentation, comparison, correction, and reuse. These mechanisms function only when scientific claims can be inspected beyond their headline performance numbers. A benchmark improvement that depends on hidden preprocessing, favorable random seeds, weak baselines, leaked information, inaccessible computation, or undocumented implementation choices is difficult to interpret as cumulative scientific progress, regardless of how impressive the resulting score appears.

Reproducibility should therefore be understood as part of a wider scientific assurance framework. Computational reproducibility establishes that an analysis can be reconstructed; replication tests whether conclusions survive new experiments; repeatability characterizes stability under repeated conditions; robustness examines sensitivity to methodological variation; and generalizability determines whether findings extend beyond the original benchmark or environment. None can substitute for the others.

The next stage of AI and robotics research should place greater value on transparent methods, uncertainty-aware comparisons, rigorous benchmark design, openly documented computational resources, realistic physical validation, and publication of replication and negative results where these improve collective knowledge. Technological progress becomes scientifically durable only when other researchers can determine not merely what result was reported, but why it occurred, under what conditions it holds, and whether the underlying claim survives independent examination.

## 5.0 Declarations

### 5.1 Author Contributions

Prof. Adrian Mercer, Dr. Leila Rahman, Prof. Daniel Okafor, and Dr. Sofia Petrov contributed to the conceptual development, literature synthesis, critical interpretation, drafting, and revision of the editorial.

### 5.2 Funding

No external funding was received for the preparation of this editorial.

### 5.3 Competing Interests

The authors declare that they have no conflicts of interest or competing interests relevant to this editorial.

### 5.4 Data Availability

No original datasets were generated or analysed for this editorial.

### 5.5 AI Use Disclosure

Generative AI was used to support literature organization and language drafting during preparation of this editorial. The authors remain responsible for independently verifying the accuracy of the arguments, citations, references, and final submitted text in accordance with the publisher's policy.

## 6.0 References

Bonsignorio, F., del Pobil, A. P., & Zereik, E. (2025). Towards reproducible robotics research. Nature Machine Intelligence, 7, 1591–1592. https://doi.org/10.1038/s42256-025-01114-7

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Sepah, N., Raff, E., Madan, K., Voleti, V., Ebrahimi Kahou, S., Michalski, V., Serdyuk, D., Arbel, T., Pal, C., Varoquaux, G., & Vincent, P. (2021). Accounting for variance in machine learning benchmarks. Proceedings of Machine Learning and Systems, 3, 747–769.

Dacrema, M. F., Cremonesi, P., & Jannach, D. (2019). Are we really making much progress? A worrying analysis of recent neural recommendation approaches. Proceedings of the 13th ACM Conference on Recommender Systems, 101–109. https://doi.org/10.1145/3298689.3347058

Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2185–2194. https://doi.org/10.18653/v1/D19-1224

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The Hitchhiker's guide to testing statistical significance in natural language processing. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1, 1383–1392. https://doi.org/10.18653/v1/P18-1128

Gundersen, O. E., & Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11503

Gundersen, O. E., Cappelen, O., Mølnå, M., & Nilsen, N. G. (2025). The unreasonable effectiveness of open science in AI: A replication study. Proceedings of the AAAI Conference on Artificial Intelligence, 39(25), 26211–26219. https://doi.org/10.1609/aaai.v39i25.34818

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. https://doi.org/10.1609/aimag.v39i3.2816

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 3207–3214. https://doi.org/10.1609/aaai.v32i1.11694

Hooker, S. (2021). The hardware lottery. Communications of the ACM, 64(12), 58–65. https://doi.org/10.1145/3467017

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G., & Narayanan, A. (2024). REFORMS: Consensus-based recommendations for machine-learning-based science. Science Advances, 10(18), eadk3452. https://doi.org/10.1126/sciadv.adk3452

Koch, B., Denton, E., Hanna, A., & Foster, J. G. (2021). Reduced, reused and recycled: The life of a dataset in machine learning research. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 1.

Li, X., Hsu, K., Gu, J., Mees, O., Pertsch, K., Walke, H. R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S., Levine, S., Wu, J., Finn, C., Su, H., Vuong, Q., & Xiao, T. (2025). Evaluating real-world robot manipulation policies in simulation. Proceedings of the 8th Conference on Robot Learning, 270, 3705–3728.

Ligot, A., & Birattari, M. (2022). On using simulation to predict the performance of robot swarms. Scientific Data, 9, 788. https://doi.org/10.1038/s41597-022-01895-1

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine learning scholarship. Queue, 17(1), 45–77. https://doi.org/10.1145/3317287.3328534

Mahmood, A. R., Korenkevych, D., Vasan, G., Ma, W., & Bergstra, J. (2018). Benchmarking reinforcement learning algorithms on real-world robots. Proceedings of the 2nd Conference on Robot Learning, 87, 561–591.

Mehta, B., Handa, A., Fox, D., & Ramos, F. (2021). A user's guide to calibrating robotic simulators. Proceedings of the 2020 Conference on Robot Learning, 155, 1326–1340.

Musgrave, K., Belongie, S., & Lim, S.-N. (2020). A metric learning reality check. Computer Vision – ECCV 2020, 681–699. https://doi.org/10.1007/978-3-030-58595-2_41

National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and replicability in science. The National Academies Press. https://doi.org/10.17226/25303

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 1.

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(164), 1–20.

Raff, E. (2019). A step toward quantifying independently reproducible machine learning research. Advances in Neural Information Processing Systems, 32, 5485–5495.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? Proceedings of the 36th International Conference on Machine Learning, 97, 5389–5400.

Reimers, N., & Gurevych, I. (2017). Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 338–348. https://doi.org/10.18653/v1/D17-1035

Remazeilles, A., Dominguez, A., Barralon, P., Torres-Pardo, A., Pinto, D., Aller, F., Mombaur, K., Conti, R., Saccares, L., Thorsteinsson, F., Prinsen, E., Cantón, A., Castilla, J., Sanz-Morère, C. B., Tornero, J., & Torricelli, D. (2022). Making bipedal robot experiments reproducible and comparable: The Eurobench software approach. Frontiers in Robotics and AI, 9, 951663. https://doi.org/10.3389/frobt.2022.951663

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. https://doi.org/10.1145/3381831

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. 2011 IEEE Conference on Computer Vision and Pattern Recognition, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

Vanschoren, J., van Rijn, J. N., Bischl, B., & Torgo, L. (2013). OpenML: Networked science in machine learning. ACM SIGKDD Explorations Newsletter, 15(2), 49–60. https://doi.org/10.1145/2641190.2641198