Open Artifacts, Closed Research: How Shared Code Can Undermine Replicability

Adrian Gavornik[a], Katarina Marcincinova[a], Marek Havrila[a]

[a]Kempelen Institute of Intelligent Technologies, Ethics and Human Values in Technology, Bratislava, Slovakia

The rapid developments in AI have contributed to a significant increase in the pace and volume of research in computer science. This growth has been accompanied by evolving publication practices enabling faster dissemination of results, increasingly relying on preprints and code sharing (Peng, 2011; Zhou et al., 2025; Cavenaghi et al., 2023). Although these practices are commonly intended to enhance transparency and reproducibility, we argue that they may produce unintended effects and undermine replicability. While reproducibility is treated as the cornerstone of current computer science research, the more fundamental question is whether research results truly can be trusted. Furthermore, the described tension between reproducibility and replicability opens up broader questions about the trustworthiness of computer science research, particularly in the context of trustworthy AI. If such systems are understood as socio-technical systems, whose reliability depends on technical, organizational, and epistemic practices, then the trustworthiness of AI cannot be separated from the trustworthiness of the scientific practices through which these systems are produced and validated.

We demonstrate how the re-use of code and evaluation methodologies in reproducibility studies facilitates the propagation of inaccuracies, including logical and implementation errors. We show these effects on a case study of a coherent line of five follow-up publications on multi-objective recommender systems (Xin Xin et al., 2025; Stamenkovic et. al., 2022; Paparella et al., 2023; Labarca Silva et al., 2024; Rajapakse and Jannach, 2025). We suggest that our observations are not just another example of questionable research practice or coincidental errors. Instead, they highlight structural vulnerabilities in current experimental practices that can be better understood in light of the two general tensions present in contemporary computer science research.

First tension points to the intricate relationship between reproducibility and replicability (Plesser, 2018; Raff et al., 2025). While both contribute to the reliability of the research and reduce accidental errors, randomness, or methodological flaws, reproducibility involves re-running the original code and data, whereas replicability requires an independent reconstruction of the model or method.[1] When artifacts such as code are not shared, a study is not reproducible; however, it can still be replicable. Thus, reproducibility is not a prerequisite for replication. In this case study, we demonstrate that the availability of easy-to-use artifacts may, however,

[1] Reproducibility and replicability here are consistent with ACM definitions (ACM *Artifact Review and Badging - Current*, 2020).

discourage genuine replication. Code reuse provides real benefits in terms of time and resource savings, while simultaneously creating an illusory assurance that errors and mistakes can be prevented by refraining from implementing the method from scratch.

Second, in the context of broader discussions on the role of reproducibility in scientific research as such (Fidler and Wilcox, 2026), we suggest that practices intended to promote transparency, such as shared code, datasets, and evaluation pipelines, can function as Latourian black-box mechanisms. As Latour (1987) argues in his laboratory studies on scientific practices, black-boxing occurs when a system works reliably enough that its internal assumptions are no longer questioned. In this sense, once the artifacts, such as shared code, produce seemingly stable and publishable outputs, their internal assumptions are no longer questioned and verified. We observed that an agreement and stabilized knowledge emerged not from independent validation, but from alignment with the same erroneous artifact. This creates an illusion of improved capabilities of multistakeholder recommender systems and scientific progress in the field as such.

The discussed tensions become especially relevant in the context of Trustworthy AI, where principles such as reliability, robustness, transparency, or accountability are often emphasized (see existing documents, such as the Ethics Guidelines for Trustworthy AI by European Commission & Directorate-General for Communications Networks, 2019; Fjeld et al., 2020). However, they often focus on the AI systems themselves, paying less attention to the scientific practices through which such systems are developed and evaluated. In other words, the trustworthiness of AI cannot be meaningfully separated from the trustworthiness of the research practices that produce it. What is required is more than just improving algorithmic or evaluation metrics, but also critically examining the practices and norms of contemporary computer science research, including code reuse, reproducibility, and experimental validation.

Artifact Review and Badging—Current. (n.d.). Retrieved May 11, 2026, from https://www.acm.org/publications/policies/artifact-review-and-badging-current

Cavenaghi, E., Sottocornola, G., Stella, F., & Zanker, M. (2023). A Systematic Study on Reproducibility of Reinforcement Learning in Recommendation Systems. *ACM Transactions on Recommender Systems*, *1*(3), 1–23. https://doi.org/10.1145/3596519

European Commission, & Directorate-General for Communications Networks, C. and T. (2019). Ethics guidelines for trustworthy AI. Publications Office. https://doi.org/10.2759/177365

Fidler, F., & Wilcox, J. (2026). Reproducibility of Scientific Results. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2026). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2026/entries/scientific-reproducibility/

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (SSRN Scholarly Paper ID 3518482). Social Science Research Network. https://doi.org/10.2139/ssrn.3518482

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University press.

Paparella, V., Anelli, V. W., Boratto, L., & Di Noia, T. (2023). Reproducibility of Multi-Objective Reinforcement Learning Recommendation: Interplay between Effectiveness and Beyond-Accuracy Perspectives. *Proceedings of the 17th ACM Conference on Recommender Systems*, 467–478. https://doi.org/10.1145/3604915.3609493

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, *11*, 76. https://doi.org/10.3389/fninf.2017.00076

Raff, E., Benaroch, M., Samtani, S., & Farris, A. L. (2025). What Do Machine Learning Researchers Mean by “Reproducible”? *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(27), 28671–28683. https://doi.org/10.1609/aaai.v39i27.35093

Rajapakse, D. C., & Jannach, D. (2025). Reassessing the Effectiveness of Reinforcement Learning based Recommender Systems for Sequential Recommendation. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3306–3314. https://doi.org/10.1145/3726302.3730322

Silva, Á. L., Parra, D., & Icarte, R. T. (2024). On the Unexpected Effectiveness of Reinforcement Learning for Sequential Recommendation. *Proceedings of the 41st International Conference on Machine Learning*, 45432–45450. https://proceedings.mlr.press/v235/silva24b.html

Stamenkovic, D., Karatzoglou, A., Arapakis, I., Xin, X., & Katevas, K. (2022). Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 957–965. https://doi.org/10.1145/3488560.3498471

Xin, X., Karatzoglou, A., Arapakis, I., & Jose, J. M. (2020). Self-Supervised Reinforcement Learning for Recommender Systems. *Proceedings of the 43rd International ACM*

SIGIR Conference on Research and Development in Information Retrieval, 931–940. https://doi.org/10.1145/3397271.3401147

Zhou, K. Z., Chen, J. E., Zheng, X., Qian, Y., Xiao, Y., & Shu, K. (2025). *"Everyone Else Does It": The Rise of Preprinting Culture in Computing Disciplines* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2511.04081