POLICY BRIEF

Open Science for Artificial Intelligence: Implementing Reproducibility to Promote Trust in AI

AUTHORS:

SIMONE KOPEINIK TONY ROSS-HELLAUER

DOMINIK KOWALD

Funded by the European Union

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them.

TABLE OF CONTENTS

KEY MESSAGES

- **Reproducibility is essential** for the credibility and robustness of AI research, particularly in highly sensitive domains (e.g., medicine) and high-risk applications. However, it is not yet a priority, as technical, cultural, and institutional barriers persist.
- **Current incentives** in academia and industry favour rapid publication and innovation over transparency, reproducibility, and quality (trustworthiness).
- **Targeted policies** on Open Science and reproducibility practices, combined with dedicated funding, can significantly improve the reproducibility, reusability, and societal trustworthiness of AI systems.

1. INTRODUCTION

Artificial Intelligence (AI) is becoming ever more central to research methods, not only in computer science but across disciplines. Indeed, AI researchers were among the recipients of the 2024 Nobel Prizes for both chemistry and physics. Yet, as with other areas of innovation, optimism over the positive potential of AI to reshape workflows must be carefully tempered by recognition of the risks of technological misuse and failure. [1]

Many scientific fields are currently reckoning with crucial concerns over the reproducibility of research findings, i.e., the extent to which the repetition of parts or the whole of a study leads to the same or similar results. Apart from common challenges faced by other disciplines, the use of AI introduces unique obstacles for reproducibility, including sensitivity to model training conditions, sources of randomness, inherent nondeterminism, costs (economic and environmental) of computational resources, and the increasing use of Automated-ML (AutoML) tools. Without reproducibility, AI systems and the research they enable risk bias and endanger trust, potentially hindering scientific progress, translation of research into practice, and public confidence in AI and science.

[1] See, for example, the AI Incident Database: https://incidentdatabase.ai
[2] Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., & Kowald, D. (2025). Reproducibility in machine-learning-based research: Overview, barriers, and drivers. AI Magazine, 46(2), e70002. https://doi.org/10.1002/aaai.70002

Given the growing importance of these topics, this Policy Brief condenses key messages to aid orientation for science policy-makers in government, funding organisations, and research institutions. We build primarily on a TIER2 study recently published in AI Review [2] to highlight the critical nature of implementing principles of Open Science in AI research. We describe the multifaceted barriers that hinder the implementation of reproducibility in everyday research practice, and identify key drivers and actionable solutions, encompassing technological advancements, procedural improvements, and strategic efforts in awareness and education to support Open Science practices.

By understanding and addressing these factors, we believe we can establish a more transparent, robust, and trustworthy AI research culture, with responsibility and reproducibility as important prerequisites.

2. BARRIERS TO AI REPRODUCIBILITY

In one well-accepted definition, [3] we can distinguish different levels of reproducibility of AI research, defined through the extent of information provided (from R1, with mere textual description, through R2 (code sharing), R3 (data sharing), to R4, where the full experimental set-up is shared for re-use). As outlined in Figure 1, each level is affected by a number of challenges, which can be counteracted through the adoption of targeted actions. For an in-depth technical explanation and description of Figure 1, we refer the interested reader to our publication in the AI Magazine [2].

In more general terms, we summarise the factors hindering AI reproducibility as follows:

CULTURAL RESISTANCE

TECHNICAL OBSTACLES

INFRASTRUCTURE GAPS

POLICY FRAGMENTATION

[3] E Gundersen, O. (2021). “The Fundamental Principles of Reproducibility.” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. https://doi.org/10.1098/rsta.2020.0210

CULTURAL RESISTANCE

A "publish or perish" culture discourages sharing data, code, or methods, with competitive pressures leading to cutting corners

- Competitive research environments prioritise quantity over quality.
- Weak incentives for data/code sharing and replication work.
- Misaligned institutional metrics (e.g., reliance on journal impact factors) undermine reproducible practices.

TECHNICAL OBSTACLES

Machine learning (ML) models are often sensitive to randomness and hardware environments, and many results cannot be independently verified, even with access to data

- Lack of access to code, data, and detailed methodological descriptions.
- Inherent ML-specific issues, such as data leakage, biases, and complex dependencies in experiments.

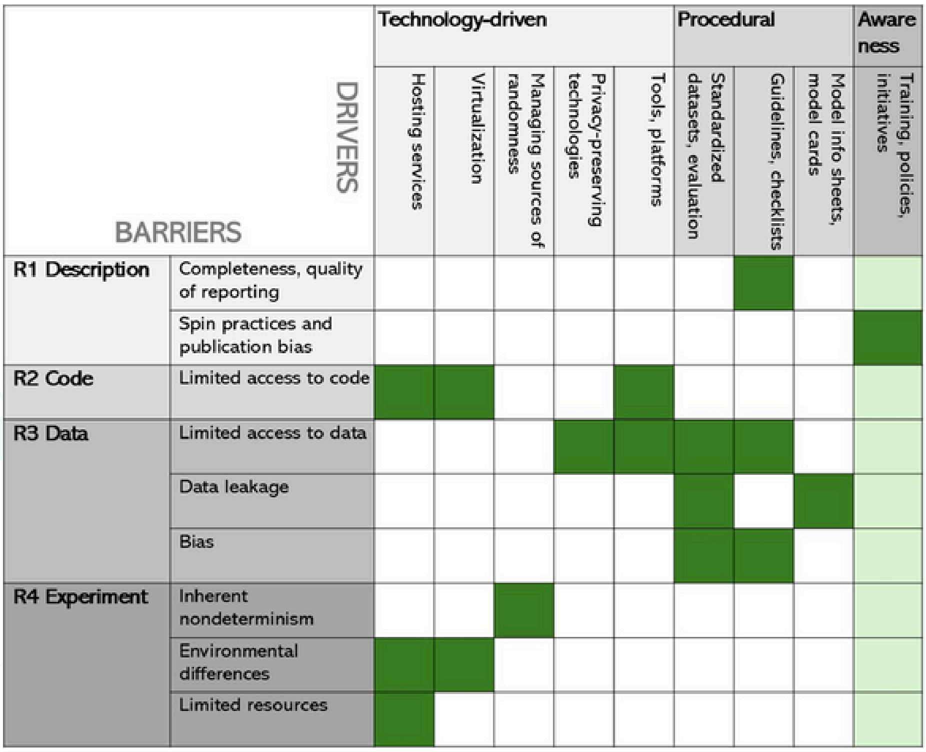
INFRASTRUCTURE GAPS

Lack of user-friendly, scalable platforms for sharing code and reproducible workflows (e.g., Docker, Code Ocean) blocks progress

- Inadequate or fragmented support for reproducible workflows.
- High cost and resource demand to replicate large-scale ML models.

POLICY FRAGMENTATION

Funders, publishers, and institutions lack coordinated, enforceable standards for reproducibility and open science



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

Figure 1: Barriers-Drivers Matrix. The color indicates to what extent a barrier can be addressed with a given driver. Thus, the light-green color of the awareness driver means that it could be used to address all barriers, but other drivers indicated with dark-green color are more effective for addressing certain barriers.

3. GENERATIVE AI IN FOCUS

Generative AI (GenAI) is a class of artificial intelligence that can create new content (e.g., text, images, music, and other forms of media). Recent years have seen great interest in the potential of these models across all sectors, including research. Another study conducted in part by TIER2 assessed the challenges and opportunities that GenAI poses for Open Science. [4] Amongst the reproducibility risks, we identified:

- Opacity of model training data, parameters, and architectures makes it difficult for others to reproduce or audit results because model behaviour is often a “black box.”
- Non-deterministic outputs — the same prompt can yield different results, and models can drift over time — undermine the ability to replicate findings consistently.
- Reliance on proprietary or closed-source models and infrastructures limits reproducibility due to restricted access, costs, and version changes.
- Poor documentation of prompts, model versions, and workflows hampers others’ ability to reconstruct how results were generated.
- Use of synthetic or manipulated datasets without clear labelling or validation introduces hidden errors and reduces confidence in reproducibility.
- Lack of attribution and traceability of training data diminishes transparency and the ability to verify or reuse resulting research outputs.

Given these serious challenges, and as GenAI evolves and becomes part of research workflows, we hence call for a concerted effort among research communities interested to investigate and address these issues, ensuring that GenAI contributes positively to the scientific community and society at large.

[4] Hosseini, M. Horbach, S.P.J.M., Holmes, K. & Ross-Hellauer, T. (2024). Open Science at the generative AI turn: An exploratory analysis of challenges and opportunities, Quantitative Science Studies, https://doi.org/10.1098/rsta.2020.0210

4. CONCLUSION

Reproducibility is not just a technical issue — it forms a foundation to trustworthy and ethical AI. We strongly believe that Open Science policies tailored to AI and supported by funding, infrastructure, and cultural incentives will ensure that AI research benefits all of society, not just a privileged few.

In order to create change, it is essential for stakeholders across the AI research ecosystem — researchers, funders, publishers, and institutions — to collaborate in mainstreaming reproducibility. This involves both cultural transformation and technical standardization. Given the fast development, we believe without strong policy action, the trustworthiness and societal value of AI innovation will be at risk.

5. RECOMMENDATIONS FOR REPRODUCIBLE AI

With the introduction of the Barriers-Drivers Matrix, we aim to emphasise the relationship between barriers and drivers, as well as the diversity of the nature of the approaches. Addressing AI reproducibility issues requires a combination of technology-driven solutions, procedural improvements, and strategies towards enhancing awareness and education. Our recommendations are listed non-hierarchically.

MAKE REPRODUCIBILITY A FUNDING REQUIREMENT

- Require grantees to share datasets, code, model cards and detailed methodology using FAIR principles and community-specific standards, respecting privacy constraints (e.g., anonymised version of the datasets in privacy-sensitive domains such as healthcare).
- Fund the development and maintenance of infrastructure for reproducibility (e.g., open-source machine learning toolkits, trusted repositories for source code, models, and datasets).

SUPPORT CULTURAL CHANGE IN RESEARCH

- Incentivise replication studies and slow, high-quality science through dedicated grant calls and tenure criteria.
- Promote reproducibility training at the graduate and postdoctoral levels.

ADOPT RESEARCH ASSESSMENT REFORM

- Move away from impact factor-based evaluations and toward transparent, process-oriented metrics (e.g., adherence to Open Science practices).
- Encourage institutions and publishers to implement open peer review and reproducibility checklists.

TAILOR OPEN SCIENCE POLICIES BY DISCIPLINE

- Avoid one-size-fits-all approaches. Recognise that reproducibility looks different in ML research vs. qualitative studies. Flexibility and community co-creation are key.

ABOUT TIER2

TIER2 is a three-year international project (2023-2025) jointly funded by the European Union (Horizon Europe) and UKRI. The project systematically investigates reproducibility across different research contexts, examining epistemological, social, and technical factors. To address these challenges, TIER2 develops and tests new tools and interventions through eight pilot activities. The project aims to bridge knowledge gaps, implement innovative solutions for managing digital objects across their full lifecycle, from data collection and research processes through to publication and preservation. The project also works to foster research communities, and influence policy to enhance reproducibility and research quality. Through co-creation methods, TIER2 actively collaborates with researchers, funders, and publishers to ensure practical and impactful outcomes.