Reproducibility: The New Frontier in AI Governance

Israel Mason-Williams [1] **Gabryel Mason-Williams** [2]

Abstract

AI policymakers are responsible for delivering effective governance mechanisms that can provide safe, aligned and trustworthy AI development. However, the information environment offered to policymakers is characterised by an unnecessarily low Signal-To-Noise Ratio, favouring regulatory capture and creating deep uncertainty and divides on which risks should be prioritised from a governance perspective. We posit that the current publication speeds in AI combined with the lack of strong scientific standards, via weak reproducibility protocols, effectively erodes the power of policymakers to enact meaningful policy and governance protocols. Our paper outlines how AI research could adopt stricter reproducibility guidelines to assist governance endeavours and improve consensus on the AI risk landscape. We evaluate the forthcoming reproducibility crisis within AI research through the lens of crises in other scientific domains; providing a commentary on how adopting preregistration, increased statistical power and negative result publication reproducibility protocols can enable effective AI governance. While we maintain that AI governance must be reactive due to AI's significant societal implications we argue that policymakers and governments must consider reproducibility protocols as a core tool in the governance arsenal and demand higher standards for AI research. Code to replicate data and figures: https://github.com/IFMW01/reproducibility-the-new-frontier-in-ai-governance

1. Introduction

AI is often regarded as a technology that will have an unprecedented impact on technological development with speculated impacts on society including, but not limited to, scientific research advances (Abramson et al., 2024; Cory-Wright et al., 2024), changes to global economics (Trammell & Korinek, 2023), up-ending job markets (Kulveit et al., 2025; Eloundou et al., 2023), new cyber security threats and opportunities (Dash et al., 2022), and revolutionising healthcare (Lee & Yoon, 2021). With the increasing economic, scientific and societal interest in this multi-purpose technology, many perspectives have emerged on where current trajectories will lead us, with prominent voices arguing both for the imminent arrival of Artificial General Intelligence (AGI) (Grace et al., 2024) and against its theoretical plausibility (Van Rooij et al., 2024). Despite the lack of consensus within the scientific community on the trajectory of AI, due to hype dynamics of AGI, there is increasing pressure for regulators and policymakers to respond to the range of risks and prepare for potential futures offered by AI advancements. Furthermore, given that current AI systems propagate and amplify complex biases (Caliskan, 2023; Kotek et al., 2023), it is crucial that the Signal-To-Noise Ratio for AI research is increased, to make it easier to assess AI capabilities and better position regulators and policymakers to act on accurate and trustworthy research.

> **Signal-To-Noise Ratio:** The quantity of research papers that contain genuine/reproducible **(signal)** insights compared to the number of papers that contain stochastic/irreproducible **(noise)** findings.

The current standard of scientific research in AI has led many prominent AI researchers to warn of a reproducibility crisis (Kapoor & Narayanan, 2023; Ball, 2023; Gundersen, 2020; Gundersen & Kjensmo, 2018; Tran et al., 2021). The definition of reproducibility is somewhat contested in literature so we explicitly define reproducibility, in line with the B2 definition provided by (Desai et al., 2025). It is important to note that the Open Science Collaboration makes no distinction between these terms (Collaboration, 2012) and the Association for Computing Machinery used these definitions until later updating them [1]. Due to the AI communities

[1]UKRI Safe and Trusted AI, Imperial College London and King's College London, London, United Kingdom [2]Independent Researcher, London, United Kingdom. Correspondence to: Israel Mason-Williams <israel.mason-williams@kcl.ac.uk>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada. Copyright 2025 by the author(s).

[1]ACM Artifact Review and Badging: https://www.acm.org/publications/policies/artifact-review-and-badging-current

use of the term reproducibility [2] [3] we have opted to center our work under this term to enable the AI community to access this work.

> **Replicability**: An independent group of researchers obtain the same result of a study using independently developed artifacts in-line with the original experimental setup **(different team, same experimental setup)**.
>
> **Reproducibility**: An independent group of researchers obtain the same findings of a study using independently developed artifacts under a modified methodological approach and/or dataset **(different team, different experimental setup)**.

It is broadly accepted that at the start of the 21st century, many different research domains, such as, Economics, Cancer Biology, and Psychology, experienced such reproducibility crises, the fallout of which has led to ineffective economic policy, opportunity cost, loss of life and ineffective medical treatments. In Figure 1 we visualise an **indicative** plot of scientific domains current publication speed versus reproducibility efforts and their projected growth, as represented in Table 1. We argue that the trajectory for AI research can be improved with the introduction of strong reproducibility protocols such as **preregistration, statistical leverage and negative result reporting**. In this paper we contextualise our recommendations and the importance of reproducibility in science by discussing how other domains have dealt with similar reproducibility issues. Through our insights we hope to bolster transparency and trust in AI research such that effective governance strategies for AI can be enacted.

2. Learning from Past Reproducibility Crises

Many empirical scientific fields have faced a reproducibility crisis (Christensen & Miguel, 2018; Errington et al., 2021a; Open Science Collaboration, 2015). In this section we cover some of the landmark reproducibility crises discussing the impacts of poor science and how each field attempted to create more robust reproducibility protocols to mitigate irreproducibility. It is important to understand core case studies for irreproducibility such that we can contextualise the impact that irreproducible AI research can have; as well as the potential harm propagation if better standards are not adopted. Furthermore failure to replicate/reproduce can foster overconfidence, underestimate uncertainty, and hinder

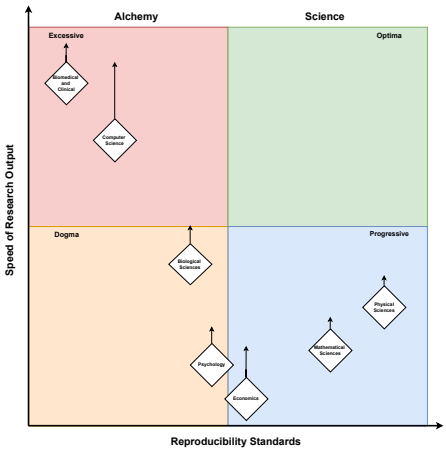


Figure 1. Indicative plot of the speed of publication versus reproducibility standards in scientific domains with average publication trajectories (↑) over the last five years. Please see Appendix Section A.1 for the methodology used to produce this plot.

scientific progress (Errington et al., 2021b) and thus, thwart policy and governance efforts.

2.1. Economics - Reproducibility

Economic research is an important factor in policy, therefore, it is crucial that rigorous scientific practices are upheld to ensure genuine advances in knowledge are made (O'Brien, 1992) . In 2006 a study on the Journal of Money, Credit, and Banking, which has mandatory data sharing requirements, found that most authors did not adhere to these practices with only 25.56% (68/266) having archival entries where required (McCullough et al., 2006). Furthermore, it was found that only 22.58% (14/62) of empirical studies could be replicated, this study sparked further inquiries into reproducibility in economics. More recent studies in economics have shown low replicability with successful replication of less than half of a selected group of papers published across top economic journals (Chang & Li, 2015).

A core case study that typifies the impact of irreproducible findings is found in the "Growth in The Time of Debt" paper published in the American Economic Review which explored the systemic relationship between public debt, growth and inflation (Reinhart & Rogoff, 2010). The paper asserted "When external debt reaches 60 percent of GDP [Gross Domestic Product], annual growth declines by about two percent; for higher levels, growth rates are roughly cut in half" (Reinhart & Rogoff, 2010). In the wake of the 2008 financial crash the relation made between external debt and growth had major impacts on governmental perspectives to economic policy. Governments sought austerity policies which seeks to reduce budget deficits, and therefore reliance on external debt, by leveraging spending

[2] AI4Europe Reproducibility Initiative: https://www.ai4europe.eu/ethics/articles/ai4europe-reproducibility-initiative

[3] ML Reproducibility Challenge: https://reproml.org

cuts for public services and/or tax increases. It has been argued that the Eurozone used the evidence to support "The Treaty on Stability, Coordination and Governance" which stated that member states should not exceed debt in excess of 60% of GDP (Compact, 2012) and that it was used to support austerity policy in the United Kingdom. An Oxfam case study report for the United Kingdom revealed that austerity measures led to an increase in inequality and created an environment for the rich to get richer (Oxfam, 2013) and further studies have linked austerity policies in the United Kingdom to hundreds of thousands of excess deaths (Walsh et al., 2022).

An attempt to replicate the results of the "Growth in the Time of Debt" paper failed due to missing data and existing errors in the calculations of the original work (Bell et al., 2014). When correctly analysing the data the replicators asserted, that there was no trend that the OECD countries conformed with regard to debt and growth. This result meant that there was no evidence to support adopting austerity policy from the original study's findings; demonstrating how over-reliance on specific irreproducible findings can have deep unintended consequences for societies. In an attempt to mitigate the harms of non replicable or reproducible economic research, there has been an increased emphasis on the importance of research design, preregistration, disclosure standards, and open sharing of data and materials (Christensen & Miguel, 2018) to improve the transparency and credibility of research outputs in this domain.

2.2. The Reproducibility Project: Cancer Biology

In 2021 the Center for Open Science concluded an eight-year-long study to replicate 193 experiments from 53 high-impact preclinical papers in cancer biology published between 2010 and 2012 (Errington et al., 2021a). Preclinical papers provide the foundation for determining which treatments to give further evaluation and testing in clinical trials on humans. While this is an important stage of treatment development, it can represent a large opportunity cost to participants when other known treatment routes are available (Kane & Kimmelman, 2021). When considering cancer patients, where time can be limited, this is of particular concern. In general it is reported that 19 of 20 cancer drugs used for clinical studies do not demonstrate enough safety, efficacy or commercial promise to achieve license, which incurs a significant financial and opportunity cost (Kane & Kimmelman, 2021). The Open Science reproduction study for cancer biology found that only 2% of studies had open data, 0% of the studies had pre-requisite protocols (a detailed plan for conducting a research study) that allowed for replication and that of the experiments that could be successfully reproduced (50/193) the effect sizes were 85% smaller on average than the original findings (Errington et al., 2021a). The alarming results provides a particularly harrowing perspective on the importance for rigorous reproducibility efforts, especially when considering the integration of AI into such domains. Furthermore, the low Signal-To-Noise Ratio means it is challenging to identify ideal cancer drug candidates in the future, pushing back scientific progress. However, it is important to note that cancer will not merely be solved via strong scientific standards alone. Following findings from this report (and other earlier studies) reproducibility recommendations for cancer biology were made that include expert statistician evaluation of experiments, preregistration, preprinting with public comments (to avoid publication bias) and transparent data and code availability (Rodgers & Collings, 2021).

2.3. Reproducibility Project: Psychology

In 2012 a study conducted by the Open Science Collaboration (OSC) commenced to examine the reproducibility of psychology research (Open Science Collaboration, 2015). In the study, the OSC attempted to replicate 100 randomly selected studies from three prestigious psychology journals. In 2015, following a three-year-long project, the results were released. To replicate the studies they used the original materials and high-powered designs. They discovered that only 36% of the studies "successfully" replicated had significance in the same direction as the original studies, but that effect sizes were half that of the effect size reported in the original studies (Open Science Collaboration, 2015). The study largely pointed towards cultural issues surrounding pressure to publish and argued that incentives for individual scientists prioritise novelty over reproduction. Such studies have prompted the development of the Transparency and Openness Promotion Guidelines (Nosek et al., 2016) which introduces a TOP Factor metric that reports a journals alignment with promoting core scholarly norms of transparency and reproducibility (Center for Open Science, 2020).

2.4. Reflections for AI

While AI is regarded as the most important emerging technology (IEE, 2023) like any other empirical science, it remains vulnerable to reproducibility pitfalls without robust research practices and protocols. Furthermore, it is important to note that given publications in AI have been growing on average at a circa 50% faster rate between 2019 and 2024 compared to most domains in the last five years, as shown in Table 1, there is a strong requirement to establish robust scientific practices before the number of publications exceeds a critical threshold where the Signal-To-Noise Ratio is too low. Without intervention increasing the ratio could represent a significant challenge beyond the scope of any individual or group, of researchers, publishing bodies and policymakers. In the absence of consensus on AI reproducibility and research standards, it is possible that industry actors will be able to capitalise on a polluted in-

Table 1. Publication trends across scientific domains over the last five years, as categorised by Dimensions database as of April 2025.

Research Domain	2019	2020	2021	2022	2023	2024	Percentage Growth 2019-2024
Biomedical & Clinical	1,170,895	1,345,291	1,417,197	1,433,960	1,435,700	1,478,650	26.284%
Information & Computer Science	475,933	520,807	590,753	639,524	723,629	818,642	72.008%
Biological	388,231	439,843	475,523	488,074	478,308	487,150	25.479%
Physical	284,936	296,182	311,913	299,285	303,834	318,808	11.888%
Mathematical	187,573	197,366	203,721	208,431	207,464	212,854	13.478%
Psychology	146,967	160,994	171,286	173,419	176,259	176,589	20.156%
Economics	81,421	90,407	95,953	101,201	106,581	109,335	34.284%

formation environment which creates a threat of regulatory capture through asymmetric information (Baron & Besanko, 1984) and could undermine AI governance endeavours and democratic systems.

3. Mitigation Strategies for The Reproducibility Crisis in AI

In recent years a few prominent conferences have engaged in practices to increase reproducibility with examples including the pre-registration workshop at NeurIPS in 2020 (Bertinetto et al.) and 2021, and the Reproducibility Challenge which has run since 2018 (Pineau et al., 2019). However, these represent small initiatives with few submissions compared to the main conferences and largely there is no consensus on how to address reproducibility in AI. In Figure 2 we present the number of papers at NeurIPS that mention GitHub. We use this as a proxy for replicability of papers, while this is an imperfect measure as discussed in Appendix Section A.2, it can be observed that the trend of mentions has increased over the past five years. The most notable changes occurs when NeurIPS introduced the Datasets and Benchmarks track in 2021. The results indicates that more papers are sharing code bases, however, there exists a large number of papers over the last five years that do not provide any mention of code bases. Furthermore, simply replicating a result does not mean that its findings are reproducible in nature, so while this is indicative of reproducibility it is a limited analysis. These findings somewhat support the belief that there is a reproducibility crisis in the field that will slow progress and propagate harm (Ball, 2023). Given the broad adoption of AI and its growing importance across domains it is of the utmost importance that reproducibility protocols are strengthened.

Of particular relevance when considering practical steps to improve reproducibility standards in AI are **preregistration, improved statistical experimental design** and finally **negative result reporting**. Each of these pragmatic reproducibility protocols can greatly improve the information environment and have shown success in other scientific domains. We discuss how leveraging policy can improve reproducibility and enable effective governance outcomes,

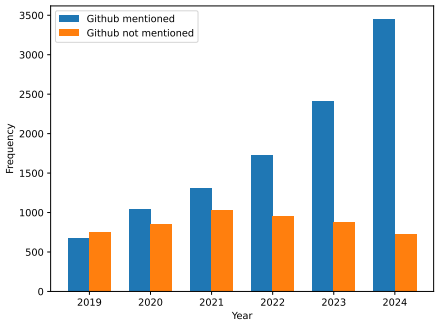


Figure 2. Number of NeurIPS publications that mention GitHub between 2019-2024. We use GitHub mentions as a proxy for replicability. The motivations and limiations of this approach are described in Appendix Section A.2

this is visualised in Figure 3 and discussed in Appendix Section B.

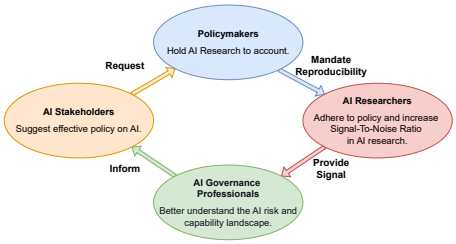


Figure 3. Preposed impact of reproducibility protocols.

3.1. Preregistration

Postdiction can occur in science when a researcher forms a new, or improved, hypothesis to explain their results due to the observation of new data; it represents a typical manifestation of hindsight bias (Roese & Vohs, 2012). Reliance on such postdictions can lead to overconfidence and inflate the likelihood of false positive results. As a result, this can hinder scientific progress as presenting postdictions as predictions reduces the communication of uncertainties which harms reproducibility (Nosek et al., 2018). Preregistration has been introduced across domains to improve research

standards and strengthen peer review mechanisms. Preregistration enables reviewers to distinguish between prediction and postdiction as the predictions and corresponding experiments to test the hypothesis are publicly registered before the experiment is conducted. (Nosek et al., 2018). However, it should be noted that while researchers consider preregistration a net-positive, it can induce more stress and cause longer project durations, but, overall, researchers recommend the practice (Sarafoglou et al., 2022). For AI, existing preregistration practices can be adopted from other fields such that researchers and policymakers have better estimations over prediction capacities for AI research. Currently research venues for AI do not require preregistration of experiments, however due to its central role in communicating uncertainties within research it should be adopted for AI research. We see that a preregistration mandate is a viable avenue for increasing the Signal-To-Noise Ratio of AI research, however this would require agreement from publishing venues and reviewers.

3.2. Statistical Leverage

In many research domains such as psychology, clinical trails and biology, studies are dependent upon voluntary participation to conduct experiments, this leads to issues surrounding sample size which can impact the robustness of findings. Typically AI research (outside of human centred studies) does not have a participation bottleneck, making it possible to conduct experiments using high numbers of samples where greater statistical power can be leveraged, such as the robust analysis conducted in physics (Lyons, 2013). Despite this, many research venues do not have requirements or guidance surrounding sample sizes used in studies. Without consensus, many papers employ a varying number of sample sizes or omit reporting altogether which increases uncertainty in findings as small sample sizes are unreliable (Cao et al., 2024). Furthermore, introducing guidance on sample sizes for AI research and appropriate use of statistics such as Standard Error of the Mean (Belia et al., 2005) would enable improved analysis reducing statistical errors. The use of large sample sizes has seen benefits in understanding knowledge transfer phenomenon in neural networks (Mason-Williams et al., 2024) and has also been suggested to improve capability reporting for evaluations of LLMs (Miller, 2024), but can and should be applied to AI research more broadly. Possible governance solutions exist in providing open access compute, GPU and CPU, resources from AI Factories or Gigafactories [4] that are dedicated to enabling improved statistical power of compute based experiments, to champion statistical significance of findings in AI.

[4]AI Factories: https://digital-strategy.ec.europa.eu/en/policies/ai-factories

3.3. Negative Result Reporting

Scientific domains often suffer from publication bias which favors the reporting of only significant or positive results. This is often due to rejection of studies with negative results, opportunity cost of writing up negative results and lack of citation incentive for negative results (Mlinarić et al., 2017); as well as conflicts with funding which favors positive outcomes (Nair, 2019). As a result, negative result publication is seldom practiced in AI, this means that AI researchers and governance experts have limited oversight in the current limitations of AI and our understanding of it. Moreover, this can lead to an over-reliance on positive results which can impact policymaking as a full scientific picture cannot be presented leading to ineffective policy implementation (Sharma & Verma, 2019). Given the predicted influence of AI and its broad range of stakeholders it is crucial that scientists, policymakers and the public demand full transparency on the state-of-play of AI capabilities through the publication of negative results. A notable step in the right direction is represented by the Science for Deep Learning NeurIPS Workshop which actively incentivised a debunking challenge to interrogate common wisdom in the field [5]. Research which publishes negative results in AI (Zhang et al., 2016; Dinh et al., 2017; McGreivy & Hakim, 2024; Mason-Williams et al., 2025) has potential to shift perspectives on long-held beliefs which can lead to innovative approaches. In AI negative results reporting can be supported with increased funding towards workshops at conferences that support such open science initiatives to incentivise the write up of negative results.

4. Conclusion

Improving reproducibility standards for AI is central to empowering policymakers to execute meaningful and effective governance mechanisms. Given the potential of AI to revolutionise numerous sectors of society it is of the utmost importance that collective action ensures scientific studies in AI are held to the highest standard to avoid the common pitfalls attributed to empirical science. By increasing awareness and calling for consensus on reproducibility protocols it is possible to increase the Signal-To-Noise Ratio of AI. However without cohesive action to address reproducibility there is a high likelihood that the many harms AI can propagate will be actualised. Thus, it is the collective responsibility of scientists, policymakers and governments alike to address reproducibility as a new frontier in AI governance.

[5]Debunking Challenge at NeurIPS 2024: https://scienceofdlworkshop.github.io/challenge/

References

The impact of technology in 2024 and beyond: an ieee global study, Oct 2023. URL https://transmitter.ieee.org/impact-of-technology-2024/.

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024. URL https://www.nature.com/articles/s41586-024-07487-w.

Baker, M. 1,500 scientists lift the lid on reproducibility, 2016. URL https://www.nature.com/articles/533452a.

Ball, P. Is ai leading to a reproducibility crisis in science? *Nature*, 624(7990):22–25, 2023. URL https://www.nature.com/articles/d41586-023-03817-6.

Baron, D. P. and Besanko, D. Regulation, asymmetric information, and auditing. *The RAND Journal of Economics*, pp. 447–470, 1984. URL https://www.jstor.org/stable/2555518?seq=1.

Belia, S., Fidler, F., Williams, J., and Cumming, G. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389, 2005. URL https://psycnet.apa.org/buy/2005-16136-002.

Bell, A., Johnston, R., and Jones, K. Stylised fact or situated messiness? the diverse effects of increasing debt on national economic growth. *Journal of Economic Geography*, 15(2):449–472, 2014. URL https://academic.oup.com/joeg/article-abstract/15/2/449/928375?redirectedFrom=fulltext.

Bertinetto, L., Henriques, J. F., Albanie, S., Paganini, M., and Varol, G. (eds.). *NeurIPS 2020 Workshop on Preregistration in Machine Learning*, volume 148 of *Proceedings of Machine Learning Research*. PMLR. URL https://proceedings.mlr.press/v148/.

Bordg, A. A replication crisis in mathematics? *The mathematical intelligencer*, pp. 1–5, 2021. URL https://link.springer.com/article/10.1007/s00283-020-10037-7.

Caliskan, A. Artificial intelligence, bias, and ethics. In *IJCAI*, pp. 7007–7013, 2023. URL https://www.ijcai.org/proceedings/2023/0799.pdf.

Cao, Y., Chen, R. C., and Katz, A. J. Why is a small sample size not enough? *The oncologist*, 29(9):761–763, 2024. URL https://academic.oup.com/oncolo/article/29/9/761/7700046.

Center for Open Science. New measure rates quality of research journals' policies to promote transparency and reproducibility. 2020. URL https://www.cos.io/about/news/new-measure-rates-quality-research-journals-policies-promote-transparency-and-reproducibility.

Chang, A. C. and Li, P. Is economics research replicable? sixty published papers from thirteen journals say'usually not'. 2015. URL https://www.federalreserve.gov/econresdata/feds/2015/files/2015083pap.pdf.

Christensen, G. and Miguel, E. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–980, 2018. URL https://www.aeaweb.org/articles?id=10.1257/jel.20171350.

Cobey, K. D., Ebrahimzadeh, S., Page, M. J., Thibault, R. T., Nguyen, P.-Y., Abu-Dalfa, F., and Moher, D. Biomedical researchers' perspectives on the reproducibility of research. *PLoS biology*, 22(11):e3002870, 2024. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3002870.

Collaboration, O. S. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on psychological science*, 7(6):657–660, 2012. URL https://journals.sagepub.com/doi/full/10.1177/1745691612462588.

Collberg, C. and Proebsting, T. A. Repeatability in computer systems research. *Communications of the ACM*, 59(3): 62–69, 2016. URL https://dl.acm.org/doi/10.1145/2812803.

Compact, F. Treaty on stability, coordination and governance in the economic and monetary union. *T/scg/en*, 1, 2012. URL https://www.consilium.europa.eu/media/20399/st00tscg26_en12.pdf.

Cory-Wright, R., Cornelio, C., Dash, S., El Khadir, B., and Horesh, L. Evolving scientific discovery by unifying data and background knowledge with ai hilbert. *Nature Communications*, 15(1):5922, 2024. URL https://www.nature.com/articles/s41467-024-50074-w.

Dash, B., Ansari, M. F., Sharma, P., and Ali, A. Threats and opportunities with ai-based cyber security intrusion detection: a review. *International Journal of Software Engineering & Applications (IJSEA)*, 13(5), 2022. URL

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4323258.

de Oliveira Andrade, R. Huge reproducibility project fails to validate dozens of biomedical studies, Apr 2025. URL https://pubmed.ncbi.nlm.nih.gov/40281293/.

Desai, A., Abdelhamid, M., and Padalkar, N. R. What is reproducibility in artificial intelligence and machine learning research? *AI Magazine*, 46(2):e70004, 2025. URL https://onlinelibrary.wiley.com/doi/10.1002/aaai.70004.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. Gpts are gpts: An early look at the labor market impact potential of large language models. arxiv. *arXiv preprint arXiv:2303.10130*, 2023. URL https://arxiv.org/abs/2303.10130.

Errington, T., Denis, A., Perfito, N., Iorns, E., and Nosek, B. Reproducibility in cancer biology: Challenges for assessing replicability in preclinical cancer biology. elife, 10, article e67995, 2021a. URL https://elifesciences.org/articles/67995.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. Investigating the replicability of preclinical cancer biology. *elife*, 10: e71601, 2021b. URL https://elifesciences.org/articles/71601.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., and Brauner, J. Thousands of ai authors on the future of ai. *arXiv preprint arXiv:2401.02843*, 2024. URL https://arxiv.org/pdf/2401.02843.

Gundersen, O. E. The reproducibility crisis is real. *AI Magazine*, 41(3):103–106, 2020. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/5318.

Gundersen, O. E. and Kjensmo, S. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11503.

Gundersen, O. E., Cappelen, O., Mølnå, M., and Nilsen, N. G. The unreasonable effectiveness of open science in ai: A replication study. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26211–26219, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/34818.

Kane, P. B. and Kimmelman, J. Is preclinical research in cancer biology reproducible enough? *Elife*, 10:e67527, 2021. URL https://elifesciences.org/articles/67527.

Kapoor, S. and Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4 (9), 2023. URL https://www.sciencedirect.com/science/article/pii/S2666389923001599.

Kotek, H., Dockum, R., and Sun, D. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, pp. 12–24, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701139. doi: 10.1145/3582269.3615599. URL https://doi.org/10.1145/3582269.3615599.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. Gradual disempowerment: Systemic existential risks from incremental ai development. *arXiv preprint arXiv:2501.16946*, 2025. URL https://arxiv.org/abs/2501.16946.

Lee, D. and Yoon, S. N. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International journal of environmental research and public health*, 18(1):271, 2021. URL https://www.mdpi.com/1660-4601/18/1/271.

Lyons, L. Discovering the significance of 5 sigma. *arXiv preprint arXiv:1310.1284*, 2013. URL https://arxiv.org/pdf/1310.1284.

Mason-Williams, G., Mason-Williams, I., and Dahlqvist, F. Data free metrics are not reparameterisation invariant under the critical and robust layer phenomena. In *High-dimensional Learning Dynamics 2025*, 2025. URL https://openreview.net/forum?id=J0iFy6BxlC.

Mason-Williams, I., Mason-Williams, G., and Sandler, M. Knowledge distillation: The functional perspective. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=Cgo73ZnAQc.

McCullough, B. D., McGeary, K. A., and Harrison, T. D. Lessons from the jmcb archive. *Journal of Money, Credit and Banking*, pp. 1093–1107, 2006. URL https://www.jstor.org/stable/3838995.

McGreivy, N. and Hakim, A. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature Machine Intelligence*, 6(10):1256–1269, 2024. URL https://www.nature.com/articles/s42256-024-00897-5.

Miller, E. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024. URL https://arxiv.org/pdf/2411.00640.

Mlinarić, A., Horvat, M., and Šupak Smolčić, V. Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia medica*, 27(3): 447–452, 2017. URL https://www.biochemia-medica.com/en/journal/27/3/10.11613/BM.2017.030201.

Nair, A. S. Publication bias-importance of studies with negative results! *Indian journal of anaesthesia*, 63(6): 505–507, 2019. URL https://journals.lww.com/ijaweb/fulltext/2019/63060/Publication_bias___Importance_of_studies_with.18.aspx.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S., Breckler, S., Buck, S., Chambers, C., Chin, G., Christensen, G., et al. Transparency and openness promotion (top) guidelines. 2016. URL https://osf.io/9f6gx/.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. URL https://www.pnas.org/action/oidcStart?redirectUri=%2Fdoi%2F10.1073%2Fpnas.1708274114.

O'Brien, D. P. Economists and data. *British Journal of Industrial Relations*, 30(2):253–285, 1992. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8543.1992.tb00774.x.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. doi: 10.1126/science.aac4716.

Oxfam. True cost of austerity 2013, Sep 2013. URL https://www-cdn.oxfam.org/s3fs-public/file_attachments/cs-true-cost-austerity-inequality-uk-120913-en_0.pdf.

Pineau, J., Sinha, K., Fried, G., Ke, R. N., and Larochelle, H. ICLR Reproducibility Challenge 2019. *ReScience C*, 5(2):5, May 2019. doi: 10.5281/zenodo.3158244. URL https://zenodo.org/record/3158244/files/article.pdf.

Reinhart, C. M. and Rogoff, K. S. Growth in a time of debt. *American economic review*, 100(2):573–578, 2010. URL https://www.aeaweb.org/articles?id=10.1257/aer.100.2.573.

Rodgers, P. and Collings, A. What have we learned?, 2021. URL https://elifesciences.org/articles/75830.

Roese, N. J. and Vohs, K. D. Hindsight bias. *Perspectives on psychological science*, 7(5):411–426, 2012. URL https://journals.sagepub.com/doi/abs/10.1177/1745691612454303.

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., and Aczel, B. A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, 9(7):211997, 2022. URL https://royalsocietypublishing.org/doi/full/10.1098/rsos.211997.

Sharma, H. and Verma, S. Is positive publication bias really a bias, or an intentionally created discrimination toward negative results? *Saudi Journal of Anaesthesia*, 13(4): 352–355, 2019. URL https://pubmed.ncbi.nlm.nih.gov/31572081/.

Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W., et al. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025. URL https://arxiv.org/pdf/2504.01848.

Trammell, P. and Korinek, A. Economic growth under transformative ai. Technical report, National Bureau of Economic Research, 2023. URL https://www.nber.org/system/files/working_papers/w31815/w31815.pdf.

Tran, D., Valtchanov, A. V., Ganapathy, K. R., Feng, R., Slud, E. V., Goldblum, M., and Goldstein, T. An open review of openreview: A critical analysis of the machine learning conference review process, 2021. URL https://openreview.net/forum?id=Cn706AbJaKW.

Van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., and Rich, P. Reclaiming ai as a theoretical tool for cognitive science. *Computational Brain & Behavior*, 7(4): 616–636, 2024. URL https://link.springer.com/article/10.1007/s42113-024-00217-5.

Walsh, D., Dundas, R., McCartney, G., Gibson, M., and Seaman, R. Bearing the burden of austerity: how do changing mortality rates in the uk compare between men and women? *J Epidemiol Community Health*, 76(12): 1027–1033, 2022. URL https://jech.bmj.com

/content/jech/early/2022/09/26/jech-2022-219645.full.pdf.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. URL https://arxiv.org/pdf/1611.03530.

A. Methodology for Figures

In this section, we detail the methodology employed to create Figures 1 and 2; we would like to highlight that these figures are entirely indicative and are not complete reconstructions of the quantities of interest. In the following subsections, we identify the limitations of our analysis and why they should be considered best attempts at capturing quantities of interest, as discussed in the main body of the paper.

A.1. Speed of Research and Reproducibility Matrix

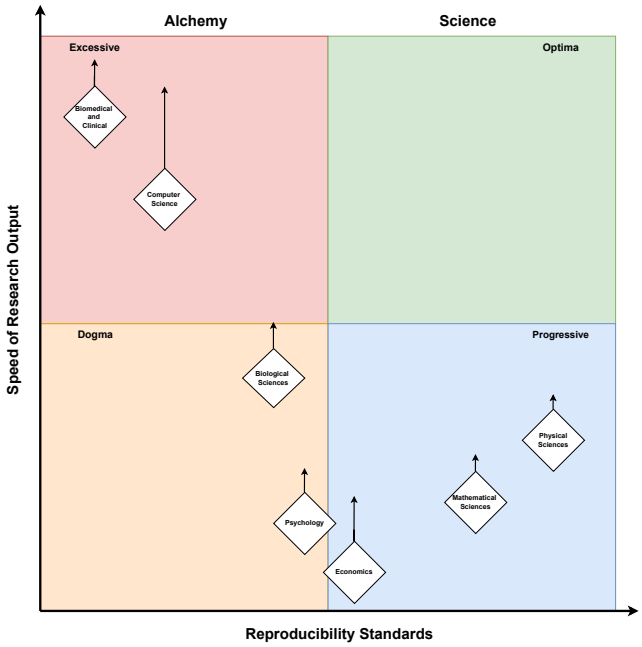


Figure 4. Blown up plot of publication speed versus reproducibility standards.

Speed of Research Outputs: For Figure 1, we use data from the research database Dimensions [6] from the 11th April 2025, which provides publishing analysis of different scientific domains. For the results in the figure, we extract the number of publications recorded for each domain in 2024. Table 1 presents the number of publications for each domain in the last five years between 2024 and 2019. The Information and Computer Science domain representing Artificial Intelligence has the most considerable average year-on-year change, double that of any other domain presented. With this trend set to continue, it underscores the prevalence of research in AI. It shows how important it is to establish strong reproducibility standards for a technology that can be applied across research domains.

Reproducibility Ratings: The positions on the reproducibility axis are primarily determined by previous reproduction analysis per domain. For the Biomedical domains previous studies have highlighted that 72% of surveyed researchers agreed there was a reproducibility crisis in biomedicine, with 27% stating the crisis was significant (Cobey et al., 2024), with a recent study in Brazil showing that only 21% of the experiments were replicable (de Oliveira Andrade, 2025). Following this we place Computer Science based on the analysis we conducted and presented in the main body in Table 1 and Figure 2; alongside previous studies which have shown low reproducibility (Gundersen & Kjensmo, 2018; Gundersen, 2020; Gundersen et al., 2025) and repeatability (Collberg & Proebsting, 2016) for the Computer Science domain. Additionally, more recent studies have shown low replicability of top-rated papers that gave Oral or Spotlight talks at ICML 2024 with 24% replication by LLMs at significant computational cost and less than 50% reproducibility by PhD students (Starace et al., 2025). A survey on Biological Sciences in 2016 found that 70% of researchers could not reproduce the findings of other scientists and circa 60% of researchers could not reproduce their findings (Baker, 2016), given this was almost a decade ago and attention has been paid to this we increased its reproducibility score. There are initiatives to improve reproducibility,

[6]Research database Dimensions: https://www.dimensions.ai/dimensions-data/

such as the ASCB Report on Reproducibility and the American Type Culture Collection (ATCC) (Cell and the Microbial Authentication Services and Programs). We believe this is an ongoing issue within Biology, but it is receiving attention from the field. We give Psychology and Economics a moderate reproducibility score due to the implementation of pre-registration practices discussed in the paper's main body. For mathematics, we provide one of the highest reproducibility standards; this is because, by nature of the field, mathematics does not depend on empirical study but rather proof and verifications, which reduces the avenues for error that are observed in more empirical domains (Bordg, 2021). Finally, we provide the highest reproduction score to the Physical Sciences as it primarily focuses on the creation of theories, and for empirical particle physics experiments, the 5-sigma significance is adopted to ensure exact findings, but with calls to tailor this for the experiment being conducted (Lyons, 2013). As a result, physics, namely via particle physics, has the best reproducibility standards due to high significance and strict reporting standards.

A.1.1. Categories:

We create four categories to describe research outputs based on the speed of outputs and reproducibility ratings. The overarching categories are Alchemy and Science and they are defined in this context as the following.

> **Alchemy:** Research that largely follows the scientific process but where findings are not reproducible across other settings, or replicable in the original setting, this research is characterised by a **low Signal-To-Noise Ratio**.
>
> **Science:** Research that follows the scientific process but where findings are reproducible across other settings and/or replicable in the original setting, this research is characterised by a **high Signal-To-Noise Ratio**.

The subcategories are defined as follows:

> **Dogma:** This category should be interpreted from the Greek definition of "something that seems true"; we believe that research domains in this category do not output research quickly but also have low reproducibility standards, which makes them susceptible to dogma.
>
> **Excessive:** Excessive research is characterised by high-speed research outputs with low reproducibility standards; the research is created quickly, but the findings do not last.
>
> **Progressive:** Research in this category has a slow publication speed but has high reproducibility standards. In this category progress can be slow-moving but each contribution can have high-impact.
>
> **Optima:** Characterised by a high publication speed and strong reproducibility practices. Research produced by scientific fields in this category represents the research holy grail where there is no trade-off between the reproducibility of findings and speed of advancement.

Both subcategories **Dogma** and **Excessive** fit into the **Alchemy** category. We have allocated subcategories **Progressive** and **Optima** in the larger **Science** category as we argue that the ability to reproduce findings separates alchemy and scientific endeavours.

A.1.2. Limitations:

The limitations of these results are that we have not conducted an exhaustive analysis across research databases; this may mean that other databases may represent other publication trends; however, we believe that this database largely represents publication trends. Furthermore, we know that AI does not encompass all CS research. However, we recognise it as one of the most active research areas, so we decided that the trends for this research domain would describe trends for AI research.

A.2. Replicability Proxy for NeurIPS

Below, we provide the code we employed to get the count of papers published at NeurIPS between 2019 and 2024, mentioning GitHub. To have a proxy for replicability, we count the number of accepted papers with GitHub links in the

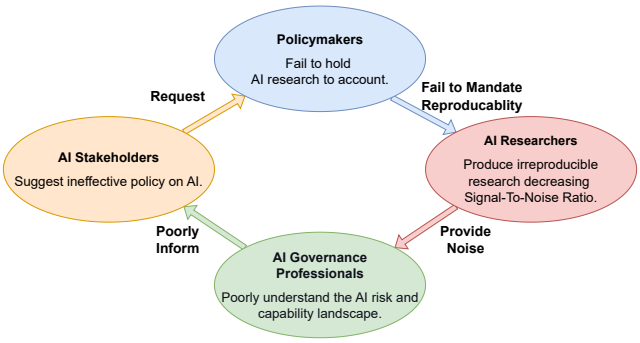
main tracks (2019-2021) and the dataset and benchmark track (2022-2024). It has been argued that having access to code bases can improve the replicability of scientific studies in Computer Science (Gundersen et al., 2025). So we feel this is an apt proxy. We limit this analysis to NeurIPS as it is rated as the top publication venue for AI[7]. Furthermore, as replicability is required for reproducibility we believe that this proxy is somewhat indicative for reproducibility.

Limitations: It is important to note that our proxy for replicability does not represent the exact number of papers that contain the repositories for their code; it is indeed possible and plausible that papers can contain references to GitHub without providing the code to replicate their work and simply providing code does not guarantee that work can always be replicated let alone their findings reproduced. Finally, our data does not represent all of the papers displayed at NeurIPS, and this count excludes Workshop papers an where the PDF analysis resulted in an error. Furthermore, not all studies are empirical and do not require code links to asses if their work is replicable/reproducible. We view replicability as a prerequisite for reproducibility generally and, as a result, we believe that insights on replicability provide a weak but indicative gauge on the status of reproducibility in AI.

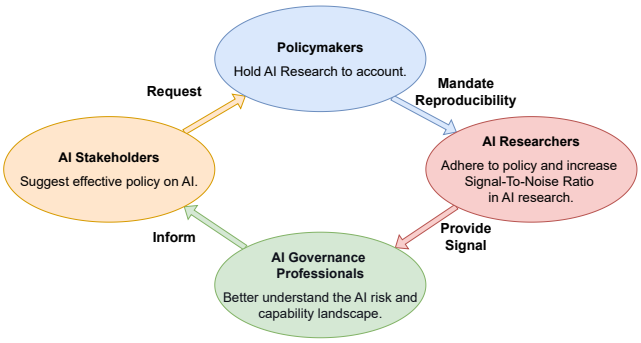
Code to replicate data and figures[8]

B. Theory of Change

Currently, irreproducible findings in AI research act as a bottleneck to effective governance. We visualise this in Sub Figure 5(a); here, we argue that Policymakers fail to hold AI research to account because they fail to mandate better standards in AI research, which in turn leads to a decreased Signal-To-Noise-Ratio of outputs, which provides Governance Professionals a more complex challenge in understanding the AI capability and risk landscape, leading to poorly informed AI Stakeholders who suggest policy that will be ultimately ineffective for holding AI research, or AI in general, to account. In Sub Figure 5(b), we show how our suggestions could improve this feedback loop. Introducing mandates of reproducibility standards on AI research, or at least research used for policy, can increase the Signal-To-Noise-Ratio of research outputs, which will provide an improved signal to Governance Professionals who will have a stronger understanding of the capability and risk landscape, which can be effectively communicated to AI stakeholders, such that effective policy on AI is championed. While there are other methods for increasing research standards that research publishers and conferences can take, we believe collaboration between policymakers and these bodies could be most effective at enacting this change. It is important to note that increased reproducibility protocols do come with trade-offs that can slow down research projects, increase burdens for researchers and require more compute (GPU or CPU hours). However, due to the projected impact of AI, we believe this is a necessary step to create a healthy information environment for AI characterised by a high Signal-To-Noise Ratio, which will, in turn, improve governance endeavours.



(a) Current Information Flow.



(b) Proposed Information Flow.

Figure 5. Reproducibility: Theory of Change.

[7]Google Scholar rating of NeurIPS: https://scholar.google.co.uk/citations?view_op=top_venues&hl=en&vq=eng_artificialintelligence

[8]Code for replication: https://github.com/IFMW01/reproducibility-the-new-frontier-in-ai-governance