Treat societally impactful scientific insights as open-source software artifacts

Liem, Cynthia C. S.; Demetriou, Andrew M.

DOI
10.1109/ICSE-SEIS58686.2023.00020

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 2023 IEEE/ACM 45th International Conference on Software Engineering

2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS) | 979-8-3503-2261-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSE-SEIS58686.2023.00020

Treat societally impactful scientific insights as open-source software artifacts

Cynthia C. S. Liem
Multimedia Computing Group
Delft University of Technology
Delft, The Netherlands
c.c.s.liem@tudelft.nl
https://orcid.org/0000-0002-5385-7695

Andrew M. Demetriou
Multimedia Computing Group
Delft University of Technology
Delft, The Netherlands
a.m.demetriou@tudelft.nl
https://orcid.org/0000-0002-0724-2278

***Abstract*—So far, the relationship between open science and software engineering expertise has largely focused on the open release of software engineering research insights and reproducible artifacts, in the form of open-access papers, open data, and open-source tools and libraries. In this position paper, we draw attention to another perspective: scientific insight itself is a complex and collaborative artifact under continuous development and in need of continuous quality assurance, and as such, has many parallels to software artifacts. Considering current calls for more open, collaborative and reproducible science; increasing demands for public accountability on matters of scientific integrity and credibility; methodological challenges coming with transdisciplinary science; political and communication tensions when scientific insight on societally relevant topics is to be translated to policy; and struggles to incentivize and reward academics who truly want to move into these directions beyond traditional publishing habits and cultures, we make the parallels between the emerging open science requirements and concepts already well-known in (open-source) software engineering research more explicit. We argue that the societal impact of software engineering expertise can reach far beyond the software engineering research community, and call upon the community members to proactively help driving the necessary systems and cultural changes towards more open and accountable research.**

***Index Terms*—open science, software engineering, open source, transdisciplinary research, responsible research practice**

I. Introduction

This article is a 'paper'[1]. At the moment it will reach broader readership with a formal citation attached, it will have passed peer review, and be part of a referenceable collection of proceedings of the ICSE 2023 Software Engineering in Society Track. This form and workflow have been the traditional template for communicating scientific outcomes, where getting papers accepted at prestigious venues has traditionally been treated as the major indicator of academic achievement.

Academic research has been operating under scarcity, both regarding job and research funding security. As a consequence, (not) getting major publications accepted and sufficiently cited thus has great career consequences. Still, for a long time, research communities have been acknowledging that contributions of scientific insight extend much beyond a paper, and proposals for open science have emerged, including ventures into open access, open and FAIR (Findable, Accessibile, Interoperable, Reusable) data, and open-source software.

The software engineering research community has been acting upon this [1], with open science policies now being explicit parts of well-respected venues like ICSE and the Empirical Software Engineering Journal, open-source tools with artifact badging being explicitly encouraged, and the option to submit registered reports entering several sub-communities such as the Conference on Mining Software Repositories. Software engineering researchers also have actively contributed to discussions on applying FAIR principles to research software [2].

With this position paper, we wish to inspire the software community to look even beyond this. More specifically, considering empirical scientific insights in the broad sense (i.e., insights requiring empirical observation of phenomena, often expressed in the form of data measurements), we will argue that making these insights more open will require infrastructure and quality assurance mechanisms similar to those needed in developing complex open-source software artifacts.

II. Arguments for open science beyond the paper

Already in 1942, Robert K. Merton noted that anti-intellectualism was rising and the integrity of science was under attack. In response, four 'institutional imperatives' were formulated as comprising the ethos of modern science: *universalism* (the acceptance or rejection of claims entering the lists of science should not depend on personal or social attributes of the person bringing in these claims), *"communism"* [sic] (common ownership of scientific findings, with the imperative to communicate findings, as opposed to secrecy), *disinterestedness* (upholding scientific integrity by not having self-interested motivations), and *organized skepticism* (judgment on the scientific contribution should be suspended until detached scrutiny is performed, according to institutionally accepted criteria) [3]. Many scientists still subscribe to these norms today [4]. These imperatives also implicitly echo in today's calls, manifestos and proposals for open science and open access [5], [6], which push for better science, which more people can access—but with which more people also can actively interact. Below, we further elaborate on several arguments and

[1]Most likely, it will not reach the reader on paper, but as a digital PDF.

DOI 10.1109/ICSE-SEIS58686.2023.00020

initiatives that argue that open science should not stop at a paper that more people can read.

A. Insufficient quality control on papers

Open-access publishing may stimulate academic and societal uptake, transform the business models of publishers, and allow for publicly funded knowledge to be publicly available. Still, open access is only an aspect of open science, and insights and methods reported in a paper may not trivially be reproducible or replicable[2], either because common specifications are not sufficiently detailed [9], or because claims may be outright false [10]. While researchers have been divided on which domains suffer from reproducibility crises [11], generally, many well-published works have failed to replicate in psychology [12] and cancer biology [13], and many concerns are arising about the replicability of machine learning outcomes [14], [15]. This leads to credibility crises, in which it is unclear whether results can actually be trusted and built upon. When policy-makers seek to base decisions on scientific insights, this can have severe consequences to human health and public trust [16], [17].

Officially, science should be self-correcting; through peer review and active continuous scrutiny processes, illegitimate claims should be detected and corrected. However, in practice, self-correction turns out painfully slow and reluctant [18], [19]. This may have to do with 'publish or perish' cultures being too strong in institutions, leading to unhealthy working environments [20]–[22], incentivizing Questionable Research Practices [23], and de-incentivizing investment in Responsible Research Practices [24].

B. Joint resource investments for collaborative momentum

With machine learning research, growing power and resource imbalances are observed between large industrial labs, and small labs in public institutions. A researcher at a university will likely not have sufficient computational resources and comprehensive data access to easily be able to replicate results as reported by big tech industry. Thus, joint investments in shared computation infrastructure are needed [25].

In psychological science, joint efforts have been coordinated into massive replication projects, where multiple teams tried to replicate canonically reported outcomes in parallel. Good examples of this are the five 'Many Labs' large-scale replication projects [26]–[30].

For such efforts, the joint investments need to focus on technical and intellectual infrastructure: i.e., the efforts required to reach a joint insight or paper, in such a way that many can indeed participate, without the transaction costs of getting started growing too large on an individual party. In other words, the focus needs to be on facilitating a shared process, rather than claiming limited-ownership output, which our present-day incentive systems still appear to push for.

[2]Definitions of 'reproducibility' and 'replicability' have not always been used in crisp ways; e.g., compare the former [7] and current [8] ACM definitions, in which definitions are swapped. Generally spoken, in the current discussion, we do not need a sharp distinction, and rather want to refer to the overall concept that repeating an experiment should give consistent results.

C. Challenges when crossing disciplines

When research becomes interdisciplinary or even transdisciplinary [31], methodology and consequent quality assurance mechanisms become more ambiguous than in the case of monodisciplinary work. While in the software engineering community, the SIGSOFT empirical standards [32] help articulating and standardizing what a reviewer should expect for different types of methodological contributions, when multiple disciplines are represented at the same time, a discipline-specific reviewer may only be capable of doing a thorough quality assessment for the parts of the contribution within their expertise, but not of the full intellectual work.

In case of transdisciplinary work, a broader spectrum of stakeholders (that may not be academics) will be involved. This again causes ambiguity on how work should be reviewed and evaluated. At the same time, for societally relevant application domains, it has been argued that broader participation of stakeholders can help getting out of credibility crises with regard to modeling choices [17]. Furthermore, if academic insights are to be implemented in society, it is not unreasonable to not only push the view of academics, but also actively involve the perspectives and experiences of non-academic societal stakeholders who will be experiencing the impact of this implementation.

D. Societal relevance causes vulnerability

Research on urgent, societally relevant challenges (e.g., climate change, public health) tend to be situated in dynamic, complex, socio-technical contexts, and require transdisciplinary approaches [31]. Problems of relevance may be wicked [33] or even super wicked [34], meaning that there is ambiguity on how the problem should be framed (while the solution depends on the framing), and one can assess whether a solution is 'better' or 'worse', but there are no hard binary outcomes of whether a result is absolutely 'true' or 'false'. In case of super wicked problem, there is high urgency and time is running out, while there is a lack of central authority.

Acting under such dynamic uncertainty comes with challenges. While fast open publishing and knowledge-sharing can be further enabled through open science, too-hasty conclusions that have not been deeply reviewed may cause hazards to human safety [16], [19]. Furthermore, while the general public will demand high accountability on societally impactful outcomes, at the same time, ambiguity, uncertainty, and dynamically changing insights make it impossible to end up with static, firm insights. Potentially contradicting readings on topics requiring deeper expertise can cause feelings of uncertainty in people, harming credibility of scientific work and leading to distrust [35]. Distrust in science causes vulnerability to credibility attacks. Indeed, in Big Tobacco, health, climate change, and AI, concerted delegitimizing efforts have been taking place as part of lobbying processes towards non-public interests [36]–[38]. Here again, more public transparency on how insights were obtained may help in sustaining trust and facilitating broader public scrutiny.

III. More holistic open science: from tools to conceptual parallels to open-source software

In response to movements towards more open science, in recent years, a plethora of process improvements with supporting platforms and tools have emerged, that support releasing a more holistic scientific artifact than a paper alone. These include pre-registration (e.g., The Center for Open Science (COS)[3], AsPredicted[4]), pre-print publication (e.g., arXiv[5], COS), storage of additional materials beyond the PDF (e.g., COS, data repositories such as Zenodo[6] and ResearchHub[7]), the co-publication of research code or software artifacts (e.g., Papers with Code[8])), decomposed publication (e.g., Octopus[9], ResearchEquals[10], Desci Foundation[11]), open peer review (e.g., F1000Research[12]) and pre-print / post publication peer review (e.g., PubPeer[13], PREreview[14], Sciety[15]). Organizations like the COS and Psychological Science Accelerator[16] have coordinated big-team data collection efforts. In parallel, traditional publication venues have started accepting more modern publication formats, such as registered reports [39]. This tooling space is presently fragmented, capturing different aspects that should improve openness in science. At a higher level, as discussed below, we however see clusters of intended functionality, that are very close to well-researched topics in software engineering research.

A. Inclusive contributorship with credit

As opposed to the traditional authorship model of publication (where author names in a list denote some undisclosed contribution to the work, the list of authors is final, and author order may imply local hierarchies that are specific to a research sub-community), there is a need to be more specific and transparent about collaborators' contributions to the intellectual work. In the publication world, the Contributor Roles Taxonomy (CRediT) has been proposed and increasingly adopted as a possible taxonomy for this, with an explicit change from authorship to contributorship [40].

Models of contributorship have naturally been implemented, facilitated and acknowledged in open-source software. In case multiple contributors work on the same artifact, version control systems (typically, Git) will be employed that help tracking the degree and provenance of changes (i.e., who contributed what at what time on the development timeline). Contributors may work in parallel, both working on main features needing priority, but also on more experimental features. Through branches, this can be done while there still is a consensus of what currently is a working non-breaking artifact on the main branch. While parallel work may be done, version control systems have protocols for resolving potential conflicts arising from parallel contributions and changes. Regardless of the status of the branch, the history of contributions will always be transparent. In addition, they allow for 'orphan' components of unfinished projects to also be gathered and transparently disclosed. In psychology, attention has for long been drawn to the 'file drawer' problem [41]. Here, many studies with non-significant results may never have been reported, but still provide useful insights, and can help meta-scientific understanding of whether results reported as significant are indeed significant, or may have resulted from sampling bias.

We can see a similar parallel to the building of scientific knowledge: a main branch can represent current stable insights, where other branches may represent work in progress, that down the road can make the overall artifact better. Where in software engineering, code review practices ensure quality control whenever a change is to be committed (regardless of whether this is on a main or experimental branch), the same can hold for peer review, where elevated reviewing safeguards can be implemented for merging into the main branch and 'pushing to production'. As we will discuss further down, the concept of the 'main branch' and versioned releases has parallels to scientific consensus of current state-of-the-art.

Where in terms of ownership, public open-source repositories may have an active team of maintainers and owners of an artifact, other people not in these groups are explicitly welcome to raise issues or feature requests if they see points for improvement, and implement and suggest contributions themselves, that the maintainers and owners may choose to incorporate. Similarly, in scientific insight, a core team may work on a particular project, but other researchers and interested parties may suggest changes or improvements that could be incorporated with visible provenance.

Where open-source projects that actively seek public contributors will have clear documentation and guidelines on how to get started and contribute if one is an outsider, similar inclusion-facilitating practices can transfer to scientific research projects, as already have been demonstrated in e.g. the Many Labs large-scale replication projects.

B. Decomposition into maintainable units

As discussed, potential reviewers to scientific work may not naturally be equipped to thoroughly review every aspect of a complete paper, especially if this paper reflects interdisciplinary work. Generally spoken, it seems unnatural to only review a complex intellectual contribution only at release time. With pre-registration and registered reports, publishing cultures already tried to solicit such feedback earlier, with positive effects on research quality and integrity [42]; however, this still involves the review of complete experimental setups.

In the software engineering world, it has generally been seen as an example of good practice to organize a complex software

[3]https://www.cos.io/
[4]https://aspredicted.org
[5]https://arxiv.org/
[6]https://zenodo.org/
[7]https://www.researchhub.com/
[8]https://paperswithcode.com/
[9]https://www.octopus.ac
[10]https://www.researchequals.com
[11]https://descifoundation.org/
[12]https://f1000research.com/
[13]https://pubpeer.com
[14]https://prereview.org/
[15]https://sciety.org
[16]https://psysciacc.org/

artifact into smaller, clearly scoped modules and functions. When committing code contributions to this overall artifact, commits also would be organized in smaller, logical contributions with a clear focus, and code review would iteratively be solicited on these small contributions. This reviewing model resembles the tools facilitating decomposed publication. In software engineering, we have already seen that decomposition will help in fostering maintainability of the overall artifact, and making it easier on new contributors to quickly get onboarded on the parts of the artifact where they wish to contribute.

We explicitly want to note that this model could work at the level of scientific artifacts (effectively, a digitally enriched form of work that currently only manifests as a paper), but also one step up, at the level of scientific insight that may source from different papers and other intellectual contributions. In scientific insight, we wish to stand on the shoulders of giants, and build upon earlier work. As such, we may source from other insights, similarly to how open-source software may make use of existing other libraries. Furthermore, again looking at functionalities offered by Git, if serious new contributions to an existing repository possibly warrant branching-off into a new strand of independent development, forking functionality allows for this, while again still keeping a living reference to the original repository.

Where in science, the insights we build upon may still be under active research, and there are chances they still may change and update, the same holds for open-source software libraries. This may create a dependency hell, for which software engineering research is actively researching best practices to still make a complex artifact building upon other artifacts as maintainable as possible. We argue that a translation of these best practices will be beneficial in navigating how scientific insight building upon other insights can best be organized and updated, in case all insights dynamically will keep evolving.

As for how to decompose and (re)organize complex code, the software engineering research community has consolidated a rich body of best practices or practices to avoid, consolidated in the form of software engineering methods, design patterns and smells. Equivalents can be formulated for the organization of scientific insight: what sub-experiments or analyses can be modularized or refactored for better reuse? Here, we would like to point out that software engineering methods tend to be taught as advanced-level programming knowledge, and as such may not as actively be part of the skillset of non-computer scientists who took an introductory programming courses—while we believe they are essential in thinking strategically about overall information organization.

C. Intermediate releases with consensus, and organizational safety to find weaknesses, iterate and improve

When developing a software artifact, pushing code to production, and having formal versioned releases, we naturally agree we have not yet reached The Ultimate and Optimal Final Product—rather, what currently is running may be a Minimum Viable Product that is iterated upon, but that will likely still have many imperfections in need of improvement. In scientific publishing, we may acknowledge this in text, but there is less incentive to demonstrate progressive improvement over subsequent contributions. Furthermore, as we will argue below, it may culturally be unsafe to admit weaknesses and visibly correct them, as this could lead to retractions and consequences on citation track records. In software development, this however is no problem, as changes and releases that can be referenced by others are more clearly separated.

To us, a scientific paper could be seen as a versioned release—a larger, but coherent collection of changes and reviewer consensus that can be frozen and referred to. Similarly, through containerization, we can freeze, save and share the entirety of a computational environment associated to a contribution. If available on a cloud-based platform, this allows for reproducibility, as well as immediate, rapid progress on both the review of material and its reuse and further development, since installation overhead will be reduced. At the same time, these freezes do not signal the end of development, and development can still actively continue.

In our argument to not only organize scientific artifacts as open-source artifacts, but even group them at a higher level of scientific insight, the concept of currently agreed-on consensus can also be taken one level up, similar to how knowledge is established in Wikipedia: for a research problem that many people work on in parallel, a meta-scientific overview of what the collective insight and consensus currently is can be consolidated. Consensus-focused publications aim to condense the overall state of a thread of research, reporting first any consensus, while also indicating ambiguities or research opportunities. One might further conceive 'living' consensus-focused articles, i.e. systematic reviews, in a model similar to Wikipedia articles, where the review stays current, as authors continuously update it. This especially will be relevant for topics with increasingly unwieldy numbers of associated publications, where there is a clear benefit to finding a means to condense scientific information; even more so, when the consolidated insights may be looked-at in informing policy (e.g., with regard to climate change or public health).

As software is developed under pressure and with short development timelines, compromises and simplifications will be made. This may lead to technical debt, in which issues needing deeper attention may pile up without being prioritized—up to the point that major and expensive fixes may be needed. Similarly, we would argue that current problems with self-correction in science and questionable insights may be a consequence of intellectual debt, as also suggested in [43]. Again, creating cultures in which this is as actively mitigated as possible will be beneficial.

Here, we already mentioned problems with organizational and cultural safety in admitting weakness and implementing corrections in scientific contributions. With software artifacts, we up-front acknowledge that programmers are competent [44], but bugs, errors and issues may still have occurred. Through testing (and ideally, test-driven development) we include safeguards that help reducing the amount of problems that will need fixing—or otherwise will help us signaling and

fixing them as early as possible. Still, we will never know whether a program will be fully bug-free. However, this does not prevent us from having justifiable perspectives on when code artifacts can be published and released. We feel that the culture of encouraging and appreciating testing and quality assurance in software engineering may be very inspirational to discussions on fostering Responsible Research Practices and academic integrity, without this being seen as a reputational threat or attack on character.

Finally, the software engineering community has increasingly been acknowledging the social and contextual organizational surroundings of a software artifact, with emerging strands of research studying how team interactions and organizational policies will affect the quality of a software artifact, as well as the efficiency and effectiveness of the process leading to its development. Similarly, these social and organizational insights will be beneficial in efforts to address the culture of scientific research itself (as well as constructive directions for systems changes, if we indeed would decide to move beyond our current ways of sharing insights through static papers).

IV. Conclusion

In this work, we have outlined how the development of scientific insights parallels the development of software. The shift from traditional publishing to open science involves challenging culture and systems changes. As the software engineering community has noticed in its own open science endeavors, investing in this is a serious, expensive and so far under-rewarded investment [1], [45], [46]. Yet, as we argued, the strong expertise of software engineering experts in acknowledging contributorship on complex larger collaborations, designing for robust maintainability, and developing based on iterative improvements, can more broadly benefit the development of scientific insight subscribing to the Mertonian norms of science—and be beneficial to society at large when complex societal challenges are addressed.

We therefore call upon our software engineering colleagues committed to open science to both think more boldly in how academic incentives can be improved beyond the focus on output, and even look beyond the software engineering research field alone. As for the first, beyond current (commendable) efforts to integrate open science principles in the publication process of software engineering venues, it will be worthwhile to think of what a 'Many Labs' equivalent in software engineering may look like. As one thought, may it take aspects of current tool competitions and benchmarks, that also focus on collective understanding, but rather from the start be framed as a joint collaborative and iterative effort?

As for the second, we invite our colleagues to join existing scientific reform movements, help developing and increasing interoperability of current tools, and critically reflect on what software engineering skills can best be taught outside of the own curriculum. As one possible thought experiment, what would a re-framing of the state-of-the-art in climate science as a complex software-like artifact look like? Which insights would need to be decomposed? Who reviews what, and what would a review discussion look like if multiple disciplines get involved? How can we allow for public scrutiny, while not feeding into public distrust?

As one example, we as authors of this article have been actively attending meta-scientific and science improvement events (such as the meetings of the Society for the Improvement of Psychological Science[17]), and have started prototyping the idea of turning scientific publication processes into Git-supported software artifacts [47]. First prototypical development towards the latter mission was performed in the form of a software development project, which several bachelor students in Computer Science and Engineering at our institute took up as part of their software engineering coursework. The resulting work was presented as a non-archival contribution at a Scientific Progress Seminar [48]. While this was not yet a formal publication, it was an excellent way to get bachelor-level software engineering students interested in research processes, and many of them enthusiastically attended the seminar, that was highly interdisciplinary, also e.g. involving epistemological philosophical work. Currently, we are working with our local Open Science community and advertising new student projects to further develop this project.

However, we ourselves are no software engineering researchers, and we are certain our colleagues with deeper expertise in the subject matter can push such developments much further. In doing this, we would argue that software engineering expertise can have even broader societal and scientific impact than it already does today.

CRediT author statement

Cynthia C. S. Liem: Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing; **Andrew M. Demetriou**: Conceptualization, Investigation, Resources, Writing – original draft, Writing – review & editing.

References

[1] D. Mendez, D. Graziotin, S. Wagner, and H. Seibold, "Open science in software engineering," in *Contemporary Empirical Methods in Software Engineering*, M. Felderer and G. H. Travassos, Eds. Cham: Springer International Publishing, 2020, pp. 477–501. [Online]. Available: https://doi.org/10.1007/978-3-030-32489-6_17

[2] A.-L. Lamprecht, L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. Martin Del Pico, V. Dominguez Del Angel, S. van de Sandt, J. Ison, P. A. Martinez, P. McQuilton, A. Valencia, J. Harrow, F. Psomopoulos, J. L. Gelpi, N. Chue Hong, C. Goble, and S. Capella-Gutierrez, "Towards FAIR principles for research software," *Data Science*, vol. 3, no. 1, pp. 37–59, 2020. [Online]. Available: https://doi.org/10.3233/DS-190026

[3] R. K. Merton *et al.*, "Science and technology in a democratic order," *Journal of legal and political sociology*, vol. 1, no. 1, pp. 115–126, 1942.

[4] M. S. Anderson, E. A. Ronning, R. DeVries, and B. C. Martinson, "Extending the Mertonian Norms: Scientists' Subscription to Norms of Research," *Journal of Higher Education*, vol. 81, pp. 366–393, 2010. [Online]. Available: https://doi.org/10.1353/jhe.0.0095

[5] M. Munafò, B. Nosek, D. Bishop, K. Button, C. Chambers, N. Percie Du Sert, U. Simonsohn, E. Wagenmakers, J. Ware, and J. Ioannidis, "A manifesto for reproducible science," *Nature Human Behaviour*, vol. 1, no. 1, Jan. 2017. [Online]. Available: https://doi.org/10.1038/s41562-016-0021

[17] https://improvingpsych.org/

[6] J. Tennant, F. Waldner, D. Jacques, P. Masuzzo, L. Collister, and C. Hartgerink, "The academic, economic and societal impacts of open access: an evidence-based review [version 3; peer review: 4 approved, 1 approved with reservations]," *F1000Research*, vol. 5, no. 632, 2016. [Online]. Available: https://doi.org/10.12688/f1000research.8460.3

[7] ACM, "Artifact Review and Badging – Version 1.0 (not current)," 2019. [Online]. Available: https://www.acm.org/publications/policies/artifact-review-badging

[8] ——, "Artifact Review and Badging Version 1.1," 2020. [Online]. Available: https://www.acm.org/publications/policies/artifact-review-and-badging-current

[9] B. McFee, J. W. Kim, M. Cartwright, J. Salamon, R. Bittner, and J. P. Bello, "Open-Source Practices for Music Signal Processing Research," *IEEE Signal Processing Magazine*, vol. 36, pp. 128–137, 2019. [Online]. Available: https://doi.org/10.1109/MSP.2018.2875349

[10] J. P. A. Ioannidis, "Why Most Published Research Findings Are False," *PLoS Medicine*, vol. 2, 2005. [Online]. Available: https://doi.org/10.1371/journal.pmed.1004085

[11] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, pp. 452–454, 2016. [Online]. Available: https://doi.org/10.1038/533452a

[12] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015. [Online]. Available: https://doi.org/10.1126/science.aac4716

[13] C. G. Begley and L. M. Ellis, "Raise standards for preclinical cancer research," *Nature*, vol. 483, pp. 531–533, 2012. [Online]. Available: https://doi.org/10.1038/483531a

[14] B. Yildiz, H. Hung, J. H. Krijthe, C. C. S. Liem, M. Loog, G. Migut, F. A. Oliehoek, A. Panichella, P. Pawełczak, S. Picek, M. de Weerdt, and J. van Gemert, "ReproducedPapers.Org: Openly Teaching and Structuring Machine Learning Reproducibility," in *Reproducible Research in Pattern Recognition: Third International Workshop, RRPR 2021, Virtual Event, January 11, 2021, Revised Selected Papers*, 2021, pp. 3—-11. [Online]. Available: https://doi.org/10.1007/978-3-030-76423-4_1

[15] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in ML-based Science," 2022. [Online]. Available: https://arxiv.org/abs/2207.07048

[16] J. M. Lawrence, G. Meyerowitz-Katz, J. A. J. Heathers, N. J. L. Brown, and K. A. Sheldrick, "The lesson of ivermectin: meta-analyses based on summary data alone are inherently unreliable," *Nature Medicine*, vol. 27, pp. 1853–1854, 2021. [Online]. Available: https://doi.org/10.1038/s41591-021-01535-y

[17] J. P. van der Sluijs, "A way out of the credibility crisis of models used in integrated environmental assessment," *Futures*, vol. 34, pp. 133–146, mar 2002. [Online]. Available: https://doi.org/10.1016/S0016-3287(01)00051-9

[18] S. Vazire and A. O. Holcombe, "Where Are The Self-Correcting Mechanisms In Science?" *Review of General Psychology*, vol. 26, pp. 212–223, 2022. [Online]. Available: https://doi.org/10.1177/10892680211033912

[19] L. Besançon, E. Bik, J. Heathers, and G. Meyerowitz-Katz, "Correction of scientific literature: Too little, too late!" *PLoS Biology*, vol. 20, 2022. [Online]. Available: https://doi.org/10.1371/journal.pbio.3001572

[20] S. Rawat and S. Meena, "Publish or perish: Where are we heading?" *J Res Med Sci*, vol. 19, pp. 87–89, feb 2014. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24778659

[21] M. de Rond and A. N. Miller, "Publish or Perish: Bane or Boon of Academic Life?" *Journal of Management Inquiry*, vol. 14, 2005. [Online]. Available: https://doi.org/10.1177/1056492605276850

[22] H. P. van Dalen and K. Henkens, "Intended and Unintended Consequences of a Publish-or-Perish Culture: A Worldwide Survey," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 7, pp. 1282—-1293, 2012. [Online]. Available: https://doi.org/10.1002/asi.22636

[23] G. Gopalakrishna, G. ter Riet, G. Vink, I. Stoop, J. M. Wicherts, and L. M. Bouter, "Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands," *PLOS ONE*, vol. 17, no. 2, 02 2022. [Online]. Available: https://doi.org/10.1371/journal.pone.0263023

[24] G. Gopalakrishna, J. Wicherts, G. Vink, I. Stoop, O. van den Akker, G. ter Riet, and L. Bouter, "Prevalence of responsible research practices among academics in The Netherlands [version 2; peer review: 2 approved]," *F1000Research*, vol. 11, no. 471, 2022. [Online]. Available: https://doi.org/10.12688/f1000research.110664.2

[25] V. J. Hellendoorn and A. A. Sawant, "The Growing Cost of Deep Learning for Source Code," *Communications of the ACM*, vol. 65, no. 1, pp. 31—-33, dec 2021. [Online]. Available: https://doi.org/10.1145/3501261

[26] R. A. Klein, K. A. Ratliff, M. Vianello, R. B. Adams, v. Bahník, M. J. Bernstein, K. Bocian, M. J. Brandt, B. Brooks, C. C. Brumbaugh, Z. Cemalcilar, J. Chandler, W. Cheong, W. E. Davis, T. Devos, M. Eisner, N. Frankowska, D. Furrow, E. M. Galliani, F. Hasselman, J. A. Hicks, J. F. Hovermale, S. J. Hunt, J. R. Huntsinger, H. IJzerman, M.-S. John, J. A. Joy-Gaba, H. Barry Kappes, L. E. Krueger, J. Kurtz, C. A. Levitan, R. K. Mallett, W. L. Morris, A. J. Nelson, J. A. Nier, G. Packard, R. Pilati, A. M. Rutchick, K. Schmidt, J. L. Skorinko, R. Smith, T. G. Steiner, J. Storbeck, L. M. Van Swol, D. Thompson, A. E. van 't Veer, L. Ann Vaughn, M. Vranka, A. L. Wichman, J. A. Woodzicka, and B. A. Nosek, "Investigating Variation in Replicability," *Social Psychology*, vol. 45, no. 3, pp. 142–152, 2014. [Online]. Available: https://doi.org/10.1027/1864-9335/a000178

[27] R. A. Klein, M. Vianello, F. Hasselman, B. G. Adams, J. Reginald B. Adams, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Štěpán Bahník, R. Batra, M. Berkics, M. J. Bernstein, D. R. Berry, O. Bialobrzeska, E. D. Binan, K. Bocian, M. J. Brandt, R. Busching, A. C. Rédei, H. Cai, F. Cambier, K. Cantarero, C. L. Carmichael, F. Ceric, J. Chandler, J.-H. Chang, A. Chatard, E. E. Chen, W. Cheong, D. C. Cicero, S. Coen, J. A. Coleman, B. Collisson, M. A. Conway, K. S. Corker, P. G. Curran, F. Cushman, Z. K. Dagona, I. Dalgar, A. D. Rosa, W. E. Davis, M. de Bruijn, L. D. Schutter, T. Devos, M. de Vries, C. Doğulu, N. Dozo, K. N. Dukes, Y. Dunham, K. Durrheim, C. R. Ebersole, J. E. Edlund, A. Eller, A. S. English, C. Finck, N. Frankowska, M. Ángel Freyre, M. Friedman, E. M. Galliani, J. C. Gandi, T. Ghoshal, S. R. Giessner, T. Gill, T. Gnambs, Ángel Gómez, R. González, J. Graham, J. E. Grahe, I. Grahek, E. G. T. Green, K. Hai, M. Haigh, E. L. Haines, M. P. Hall, M. E. Heffernan, J. A. Hicks, P. Houdek, J. R. Huntsinger, H. P. Huynh, H. IJzerman, Y. Inbar, Åse H. Innes-Ker, W. Jiménez-Leal, M.-S. John, J. A. Joy-Gaba, R. G. Kamiloğlu, H. B. Kappes, S. Karabati, H. Karick, V. N. Keller, A. Kende, N. Kervyn, G. Knežević, C. Kovacs, L. E. Krueger, G. Kurapov, J. Kurtz, D. Lakens, L. B. Lazarević, C. A. Levitan, J. Neil A. Lewis, S. Lins, N. P. Lipsey, J. E. Losee, E. Maassen, A. T. Maitner, W. Malingumu, R. K. Mallett, S. A. Marotta, J. Meedović, F. Mena-Pacheco, T. L. Milfont, W. L. Morris, S. C. Murphy, A. Myachykov, N. Neave, K. Neijenhuijs, A. J. Nelson, F. Neto, A. L. Nichols, A. Ocampo, S. L. O'Donnell, H. Oikawa, M. Oikawa, E. Ong, G. Orosz, M. Osowiecka, G. Packard, R. Pérez-Sánchez, B. Petrović, R. Pilati, B. Pinter, L. Podesta, G. Pogge, M. M. H. Pollmann, A. M. Rutchick, P. Saavedra, A. K. Saeri, E. Salomon, K. Schmidt, F. D. Schönbrodt, M. B. Sekerdej, D. Sirlopú, J. L. M. Skorinko, M. A. Smith, V. Smith-Castro, K. C. H. J. Smolders, A. Sobkow, W. Sowden, P. Spachtholz, M. Srivastava, T. G. Steiner, J. Stouten, C. N. H. Street, O. K. Sundfelt, S. Szeto, E. Szumowska, A. C. W. Tang, N. Tanzer, M. J. Tear, J. Theriault, M. Thomae, D. Torres, J. Traczyk, J. M. Tybur, A. Ujhelyi, R. C. M. van Aert, M. A. L. M. van Assen, M. van der Hulst, P. A. M. van Lange, A. E. van 't Veer, A. Vásquez-Echeverría, L. A. Vaughn, A. Vázquez, L. D. Vega, C. Verniers, M. Verschoor, I. P. J. Voermans, M. A. Vranka, C. Welch, A. L. Wichman, L. A. Williams, M. Wood, J. A. Woodzicka, M. K. Wronska, L. Young, J. M. Zelenski, Z. Zhijia, and B. A. Nosek, "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 4, pp. 443–490, 2018. [Online]. Available: https://doi.org/10.1177/2515245918810225

[28] C. Ebersole, O. Atherton, A. Belanger, H. Skulborstad, J. Allen, J. Banks, E. Baranski, M. Bernstein, D. Bonfiglio, L. Boucher, E. Brown, N. Budiman, A. Cairo, C. Capaldi, C. Chartier, J. Chung, D. Cicero, J. Coleman, J. Conway, W. Davis, T. Devos, M. Fletcher, K. German, J. Grahe, A. Hermann, J. Hicks, N. Honeycutt, B. Humphrey, M. Janus, D. Johnson, J. Joy-Gaba, H. Juzeler, A. Keres, D. Kinney, J. Kirshenbaum, R. Klein, R. Lucas, C. Lustgraaf, D. Martin, M. Menon, M. Metzger, J. Moloney, P. Morse, R. Prislin, T. Razza, D. Re, N. Rule, T. Sacco, K. Sauerberger, E. Shrider, M. Shultz, C. Siemsen, K. Sobocko, R. Sternglanz, A. Summerville, K. Tskhay, Z. van Allen, L. Vaughn, R. Walker, A. Weinberg, J. Wilson, J. Wirth, J. Wortman, and B. Nosek, "Many Labs 3: Evaluating participant

pool quality across the academic semester via replication," *Journal of Experimental Social Psychology*, vol. 67, pp. 68–82, Nov. 2016. [Online]. Available: https://doi.org/10.1016/j.jesp.2015.10.012

[29] R. A. Klein, C. L. Cook, C. R. Ebersole, C. Vitiello, B. A. Nosek, J. Hilgard, P. H. Ahn, A. J. Brady, C. R. Chartier, C. D. Christopherson, S. Clay, B. Collisson, J. T. Crawford, R. Cromar, G. Gardiner, C. L. Gosnell, J. Grahe, C. Hall, I. Howard, J. A. Joy-Gaba, M. Kolb, A. M. Legg, C. A. Levitan, A. D. Mancini, D. Manfredi, J. Miller, G. Nave, L. Redford, I. Schlitz, K. Schmidt, J. L. M. Skorinko, D. Storage, T. Swanson, L. M. Van Swol, L. A. Vaughn, D. Vidamuerte, B. Wiggins, and K. A. Ratliff, "Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement," *Collabra: Psychology*, vol. 8, no. 1, 04 2022, 35271. [Online]. Available: https://doi.org/10.1525/collabra.35271

[30] C. R. Ebersole, M. B. Mathur, E. Baranski, D.-J. Bart-Plange, N. R. Buttrick, C. R. Chartier, K. S. Corker, M. Corley, J. K. Hartshorne, H. IJzerman, L. B. Lazarević, H. Rabagliati, I. Ropovik, B. Aczel, L. F. Aeschbach, L. Andrighetto, J. D. Arnal, H. Arrow, P. Babincak, B. E. Bakos, G. Baník, E. Baskin, R. Belopavlović, M. H. Bernstein, M. Białek, N. G. Bloxsom, B. Bodroža, D. B. V. Bonfiglio, L. Boucher, F. Brühlmann, C. C. Brumbaugh, E. Casini, Y. Chen, C. Chiorri, W. J. Chopik, O. Christ, A. M. Ciunci, H. M. Claypool, S. Coary, M. V. Čolić, W. M. Collins, P. G. Curran, C. R. Day, B. Dering, A. Dreber, J. E. Edlund, F. Falcão, A. Fedor, L. Feinberg, I. R. Ferguson, M. Ford, M. C. Frank, E. Fryberger, A. Garinther, K. Gawryluk, K. Ashbaugh, M. Giacomantonio, S. R. Giessner, J. E. Grahe, R. E. Guadagno, E. Hałasa, P. J. B. Hancock, R. A. Hilliard, J. Hüffmeier, S. Hughes, K. Idzikowska, M. Inzlicht, A. Jern, W. Jiménez-Leal, M. Johannesson, J. A. Joy-Gaba, M. Kauff, D. J. Kellier, G. Kessinger, M. C. Kidwell, A. M. Kimbrough, J. P. J. King, V. S. Kolb, S. Kołodziej, M. Kovacs, K. Krasuska, S. Kraus, L. E. Krueger, K. Kuchno, C. A. Lage, E. V. Langford, C. A. Levitan, T. J. S. de Lima, H. Lin, S. Lins, J. E. Loy, D. Manfredi, Łukasz Markiewicz, M. Menon, B. Mercier, M. Metzger, V. Meyet, A. E. Millen, J. K. Miller, A. Montealegre, D. A. Moore, R. Muda, G. Nave, A. L. Nichols, S. A. Novak, C. Nunnally, A. Orlić, A. Palinkas, A. Panno, K. P. Parks, I. Pedović, E. Pekala, M. R. Penner, S. Pessers, B. Petrović, T. Pfeiffer, D. Pieńkosz, E. Preti, D. Purić, T. Ramos, J. Ravid, T. S. Razza, K. Rentzsch, J. Richetin, S. C. Rife, A. D. Rosa, K. H. Rudy, J. Salamon, B. Saunders, P. Sawicki, K. Schmidt, K. Schuepfer, T. Schultze, S. Schulz-Hardt, A. Schütz, A. N. Shabazian, R. L. Shubella, A. Siegel, R. Silva, B. Sioma, L. Skorb, L. E. C. de Souza, S. Steegen, L. A. R. Stein, R. W. Sternglanz, D. Stojilović, D. Storage, G. B. Sullivan, B. Szaszi, P. Szecsi, O. Szöke, A. Szuts, M. Thomae, N. D. Tidwell, C. Tocco, A.-K. Torka, F. Tuerlinckx, W. Vanpaemel, L. A. Vaughn, M. Vianello, D. Viganola, M. Vlachou, R. J. Walker, S. C. Weissgerber, A. L. Wichman, B. J. Wiggins, D. Wolf, M. J. Wood, D. Zealley, I. Žeželj, M. Zrubka, and B. A. Nosek, "Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability," *Advances in Methods and Practices in Psychological Science*, vol. 3, no. 3, pp. 309–331, 2020. [Online]. Available: https://doi.org/10.1177/2515245920958687

[31] OECD, "Addressing societal challenges using transdisciplinary research," OECD, Tech. Rep. 88, 2020. [Online]. Available: https://doi.org/10.1787/0ca0ca45-en

[32] P. Ralph, N. b. Ali, S. Baltes, D. Bianculli, J. Diaz, Y. Dittrich, N. Ernst, M. Felderer, R. Feldt, A. Filieri, B. B. N. de França, C. A. Furia, G. Gay, N. Gold, D. Graziotin, P. He, R. Hoda, N. Juristo, B. Kitchenham, V. Lenarduzzi, J. Martínez, J. Melegati, D. Mendez, T. Menzies, J. Molleri, D. Pfahl, R. Robbes, D. Russo, N. Saarimäki, F. Sarro, D. Taibi, J. Siegmund, D. Spinellis, M. Staron, K. Stol, M.-A. Storey, D. Taibi, D. Tamburri, M. Torchiano, C. Treude, B. Turhan, X. Wang, and S. Vegas, "Empirical Standards for Software Engineering Research," 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2010.03525

[33] H. W. J. Rittel and M. M. Webber, "Dilemmas in a general theory of planning," *Policy Sciences*, vol. 4, no. 2, pp. 155–169, Jun 1973. [Online]. Available: https://doi.org/10.1007/BF01405730

[34] K. Levin, B. Cashore, S. Bernstein, and G. Auld, "Overcoming the tragedy of super wicked problems: constraining our future selves to ameliorate global climate change," *Policy Sciences*, vol. 45, no. 2, pp. 123–152, Jun 2012. [Online]. Available: https://doi.org/10.1007/s11077-012-9151-0

[35] C. Chang, "Motivated processing: How people perceive news covering novel or contradictory health research findings," *Science Communication*, vol. 37, no. 5, pp. 602–634, 2015. [Online]. Available: https://doi.org/10.1177/1075547015597914

[36] G. Reed, Y. Hendlin, A. Desikan, T. MacKinney, E. Berman, and G. T. Goldman, "The disinformation playbook: how industry manipulates the science-policy process—and how to restore scientific integrity," *Journal of Public Health Policy*, vol. 42, pp. 622—634, 2021. [Online]. Available: https://doi.org/10.1057/s41271-021-00318-6

[37] L. Antilla, "Climate of scepticism: Us newspaper coverage of the science of climate change," *Global Environmental Change*, pp. 338–352, 2005. [Online]. Available: https://doi.org/10.1016/j.gloenvcha.2005.08.003

[38] M. Abdalla and M. Abdalla, "The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 287–297. [Online]. Available: https://doi.org/10.1145/3461702.3462563

[39] B. A. Nosek and D. Lakens, "Registered reports: A method to increase the credibility of published results," *Social Psychology*, vol. 45, pp. 137–141, 2014. [Online]. Available: https://doi.org/10.1027/1864-9335/a000192

[40] M. K. McNutt, M. Bradford, J. M. Drazen, B. Hanson, B. Howard, K. H. Jamieson, V. Kiermer, E. Marcus, B. K. Pope, R. Schekman, S. Swaminathan, P. J. Stang, and I. M. Verma, "Transparency in authors' contributions and responsibilities to promote integrity in scientific publication," *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, pp. 2557–2560, 2018. [Online]. Available: https://doi.org/10.1073/pnas.1715374115

[41] R. Rosenthal, "The file drawer problem and tolerance for null results," *Psychological Bulletin*, vol. 86, no. 3, pp. 638—-641, 1979. [Online]. Available: https://doi.org/10.1037/0033-2909.86.3.638

[42] C. K. Soderberg, T. M. Errington, S. R. Schiavone, J. Bottesini, F. S. Thorn, S. Vazire, K. M. Esterling, and B. A. Nosek, "Initial evidence of research quality of registered reports compared with the standard publishing model," *Nature Human Behaviour*, vol. 5, pp. 990–997, 2021. [Online]. Available: https://doi.org/10.1038/s41562-021-01142-4

[43] J. Zittrain, "Intellectual Debt: With Great Power Comes Great Ignorance," 2019. [Online]. Available: https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37373276

[44] R. A. DeMillo, R. J. Lipton, and F. G. Sayward, "Hints on test data selection: Help for the practicing programmer," *IEEE Computer*, vol. 11, no. 4, pp. 34–41, 1978.

[45] C. S. Timperley, L. Herckis, C. Le Goues, and M. Hilton, "Understanding and improving artifact sharing in software engineering research," *Empirical Software Engineering*, vol. 26, no. 4, 2021. [Online]. Available: https://doi.org/10.1007/s10664-021-09973-5

[46] B. Hermann, "What has artifact evaluation ever done for us?" *IEEE Security & Privacy*, vol. 20, no. 5, pp. 96–99, 2022. [Online]. Available: https://doi.org/10.1109/MSEC.2022.3184234

[47] A. M. Demetriou and C. C. S. Liem, "Alexandria: a Proof-of-Concept Publication Platform that Treats Academic Outputs like Software Artifacts," 2022. [Online]. Available: https://osf.io/hd5nu/

[48] A. M. Demetriou, A. van der Meijden, J. Sloof, M. de Wit, E. Witting, A. Zlei, and C. C. S. Liem, "Alexandria: a Proof-of-Concept Publication Platform that Treats Academic Outputs like Software Artifacts," in *Scientific Progress – Individual and Collective seminar*, 2022.