Journal of Quantitative Description: Digital Media 6(2026), 1–16 10.51685/jqd.2026.007

Commercial Determinants of Replication Infeasibility in Computational Social Science

J. NATHAN MATIAS

Cornell University Citizens and Technology Lab, USA

CASSIDY WALDRIP

DAVID LAZER

Northeastern University, USA

Computational social science has expanded the capacity of scientists to study connected human behavior at previously unprecedented scales. Yet from its beginning, scientists expressed concern that its reliance on private companies might produce a body of work that cannot be critiqued or replicated. Such commercial determinants of science have been observed in other fields including public health where science has implications for corporate liability. In this meta-scientific report, we analyze the population of computational social science articles about technology platforms published in three general scientific journals to investigate commercial determinants of scientific replicability in computational social science. We find that only 26% of those papers can be replicated today, and that 34% of computational social science studies published in leading general scientific journals rely on special arrangements with corporations that are impossible to replicate without special permission. We find that articles relying on API access, scraping or access to nonprofit platforms have much higher potential for replication. These findings are consistent with broader literature on the commercial forces that determine the direction and reliability of science.

J. Nathan Matias: nathan.matias@cornell.edu
Cassidy Waldrip: waldrip.c@northeastern.edu
David Lazer: d.lazer@northeastern.edu
Date submitted: 2025-11-16

Keywords: computational social science, replication, commercial determinants, API, online platforms, Wikipedia, technology platforms, scraping, data access

Introduction

One of the foundational papers announcing the arrival of a computational social science speculated two possible futures with respect to data access: "Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated" (Lazer et al., 2009). This matters for two reasons: such limited access threatens to dramatically set back scientific progress generally; and, in particular, eliminate progress in any domains where scientific findings might run counter to the very powerful interests that control data access. Given the global importance of the internet and the commercial giants that exist on top of the internet, the crucial question is: is it at all possible to speak truth to power in computational social science at our present time?

The independence of science from the technology industry has grown in importance as the public has turned to scientists to be arbiters of questions involving the harms and benefits of digital technologies ranging from social media to artificial intelligence (Young et al., 2022; Matias and Price, 2025; Dörr et al., 2025). When courts, policymakers, and public health authorities place expectations upon scientists that they provide guidance, it usually comes with a demand for generalizable knowledge across multiple studies, (Galea and Buckley, 2024; of the Surgeon General et al., 2023). Yet this work is rendered impossible when companies hinder, threaten, or fail to enable such replications (Orben and Matias, 2025).

Scholars who study the relationship between science and policy have observed similar patterns in multiple fields. In the fields of public health, toxicology, and climate science, companies have leveraged tools including funding, legal threats, lobbying, and public relations to shape the supply and interpretation of of available science in their favor (Oreskes and Conway, 2011; Gilmore et al., 2023). In recent years, scientists have raised concerns that this

pattern of influence and obstruction also affects computational social science (Tromble, 2021; Dommett and Tromble, 2022; de Vreese and Tromble, 2023; Freelon, 2018; Lazer et al., 2009; Bak-Coleman et al., 2025).

For something to be science it needs to be repeatable within a reasonable margin of similarity (Shapin and Schaffer, 1985; Buzbas and Devezer, 2024). Not all forms of inquiry promise to discover enduring explanations of social and behavioral phenomena. Yet scholars who aspire to scientific knowledge in social and behavioral fields including psychology (Zwaan et al., 2018; Gantman et al., 2018), sociology (Freese and Peterson, 2017), economics, and political science (Brodeur et al., 2026) generally agree that the pursuit of science depends on the ability to independently repeat a finding with new data over multiple attempts, what scholars call "replicability" (Miske et al., 2026; Buzbas et al., 2023). Such claims of generalizable laws with respect to matters of human society must be viewed with skepticism and constantly evaluated against the constantly re-imagined societies of the present. Since 21st century humans exist in a world where social arrangements are sometimes reconstituted literally with the touch of a button, social science must wrestle with the issue of temporal validity, that findings yesterday may not hold today (Matias, 2023; Bak-Coleman et al., 2021; Munger, 2019).

While digital technologies are not necessarily constrained to behave in predictable ways, the makers of these technologies face numerous economic and regulatory pressures to create systems whose technical and socio-technical outputs can be predicted and explained (Matias, 2023; Orben and Matias, 2025). To support this knowledge, technology firms have addressed the problem of repeatable knowledge by amassing an unprecedented capacity to conduct experiments and observe the behavior of billions of people in real-time (Kohavi et al., 2009). Where technology platforms have released datasets of their routine experiments, scientists have been able to find evidence of robust, reproducible findings over time (Matias et al., 2021; Robertson et al., 2023; Shulman et al., 2024; Aubin Le Quéré and Matias, 2025). In the earliest years of socio-technical platforms, such data was generously available to scientists, as companies grappled with their business models and society grappled with the privacy implications of that availability. As it became evident that data was central to firms' profitability (initially for ad targeting; and today for feeding large language models), organizations have rolled back data availability (Freelon, 2018). While these restrictions have affected the capacity to ask new questions, they also undermine the repeatability of previously-published results.

How much of internet-focused computational social science that has been published in the leading general social science journals may be repeated with new data from platforms today? To investigate the role of commercial factors in the replicability of computational social science, we compared the replication feasibility of research conducted with commercial and non-commercial platforms across a twenty year period. We queried three leading general scientific journals for all papers relying on data from communication technology platforms from 2004 into 2025. We then organized two analysts to code the papers for details related to the methods the papers relied on and whether replication would still be possible today. Within this population, we report the overall percentage of studies that could be replicated today. We also report the percentages within subgroups including data collection method and whether the paper included Wikipedia, the primary non-commercial platform that appeared in these journals.

Methods

To support this analysis of general scientific journals, we collected articles from Nature, Science, and the Proceedings of the National Academies of Science (PNAS). To identify articles, we queried the search feature of each journal within the period from January 1st 2004, just before the launch of Facebook, until the present day. For each journal, we queried a list of prominent social platforms, using the query "Facebook OR Twitter OR Reddit OR TikTok OR Instagram OR LinkedIn OR SnapChat OR Pinterest OR Discord OR Slack OR YouTube OR WhatsApp OR Telegram OR Tumblr OR Twitch OR Quora OR Mastodon OR WeChat OR Douyin OR QQ OR Weibo OR Wikipedia." While Nature and Science returned exact matches, PNAS returned 101 pages of 20 results each, sorted by relevance. For PNAS, we recorded all pages of results until observing 5 pages of results that were not relevant to the social sciences and technology platforms. The resulting result list was 32 pages long. For each set of search results, we then removed duplicates to arrive at the final list of papers.

We conducted a content analysis of papers by organizing two coders to review each of the 1,330 papers in the population. Viewing the title and abstract, we determined if the article was about a social or behavioral topic, and whether it involved a technology platform. If unsure, we read the article to confirm. Within the 187 that met these criteria, we then considered further categories by consulting the author list, the text of the article, the supplementary information, and as necessary, published descriptions of any datasets they relied on.

After initial classification, the pair of coders met for an extensive conversation about articles where there was disagreement on any item. In two articles for example, we collectively concluded that the article was about a technology platform but not social/behavioral (Table 2). This stage was essential, since assessing the replicability of a paper published up to twenty years ago can require substantial information of what has transpired across that time. For example, several papers in the population rely on techniques that were outlawed by Facebook as part of an agreement with the United States government after the Cambridge Analytica scandal. Other papers relied on application programming interfaces (APIs) which are no longer available or which have become prohibitively expensive. Other papers could only have been permitted by privacy laws in their jurisdiction if they resulted from negotiated agreements with governments or companies, agreements that were not acknowledged in the paper themselves. After considering such factors, the coders recorded a final agreement.

Classifying Articles

Coders assessed whether articles were social or behavioral, whether they focused on a technology platform, whether the study required special cooperation from the platform company, whether the study relied on scraping or APIs, whether it included Wikipedia, and whether data collection was still available in September and October 2025.

Coders determined an article to be social or behavioral if it involved a quantitative inquiry about human social or behavioral science. The study must also be empirical in nature, so a review article would not be included in the analysis. For example, a project analyzing user social media posts for misinformation would be included, but a review of papers that study misinformation patterns on social media would not.

To determine eligibility, coders assessed whether a study focused on a technological platform in pursuit of those social and behavioral questions. We define a technological platform as a system on which people have user accounts and use the platform to communicate with other humans in some way. These systems include but are not limited to traditional social media platforms; we included studies that involve hospital communication systems and mobile phone data, for example. However, we did not include studies that simply used a technology platform's advertising services to recruit survey participants, or studies that involved technologies that

did not have an interpersonal communication aspect.

Coders also determined whether the project required special cooperation from the platform company. Studies were classified as such for several reasons: disclosed involvement of employees of the company or disclosed access to privileged data or functions of the platform. Studies were also included if they used methods known to the coder to be unavailable to the average researcher at the time of the study, implying a special arrangement with the company. Using an API, whether free or paid, was not labeled as company cooperation. For example, these studies include research conducted by scientists who were employees of Twitter (now X) and who had access to full, real-time data from the company. In another example, it also includes Social Science One, a partnership between academics and Meta/Facebook that has since ended.

Coders also considered whether the researchers used an API or scrape public information for any part of the study. To be included, scraping access or the API must have been accessible to the average researcher at the time of the study. For example, some studies queried a public record of Wikipedia data from the Wikipedia website or through an API that the Wikimedia Foundation makes available. We also applied this label to studies that relied on data that had been scraped by third parties.

Articles also received a label for whether they included data from Wikipedia. Studies fall under this category if the authors report any use of Wikipedia data. For example, some of the included studies analyzed the network structure of links between Wikipedia articles, based on publicly-scraped data. Articles that include multiple sources of data, including negotiated data from other platforms would also be labeled as involving Wikipedia if just one of several datasets was from Wikipedia.

Finally, we considered whether data collection was possible in September and October 2025, when coders analyzed the papers. Studies meet this criterion if the study could be reproduced today with *new* data. To be classified in this way, all relevant APIs or data access infrastructures would need to be still available for further study. For example, studies that relied on Twitter data collected from the Twitter API do not fit this description as the Twitter/X API is no longer accessible to the average researcher due to increased restrictions and fees in recent

years. If web scraping had been used previously, we consider data collection to be possible so long as the website has not since implemented anti-scraping policies or legal threats. Finally, if a project relies on several different data sources and one or more of those sources is no longer available, coders determined that data collection is not still available.

A number of studies used publicly-available datasets that were collected by other researchers. When coding these studies, we applied a transitive principle by reviewing the methods used to collect data for that dataset. If any one of the datasets in a study had a characteristic, we assigned that characteristic to the study. For example, more than one study drew from Wikipedia data and also from data about social media accounts that social network platforms no longer make available to researchers. In those cases, we applied the Wikipedia label and noted that the paper used an API or scraping. We then labeled such papers to indicate that data collection is no longer possible since at least one dataset is now restricted, even though the Wikipedia data is still available.

Selecting the Final Sample

We report Cohen's Kappa for every variable in our coding scheme in Table 2. After conducting the coding process, we identified 14 eligible articles. The full querying and filtering process is reported in the PRISMA flow diagram in Figure 1.

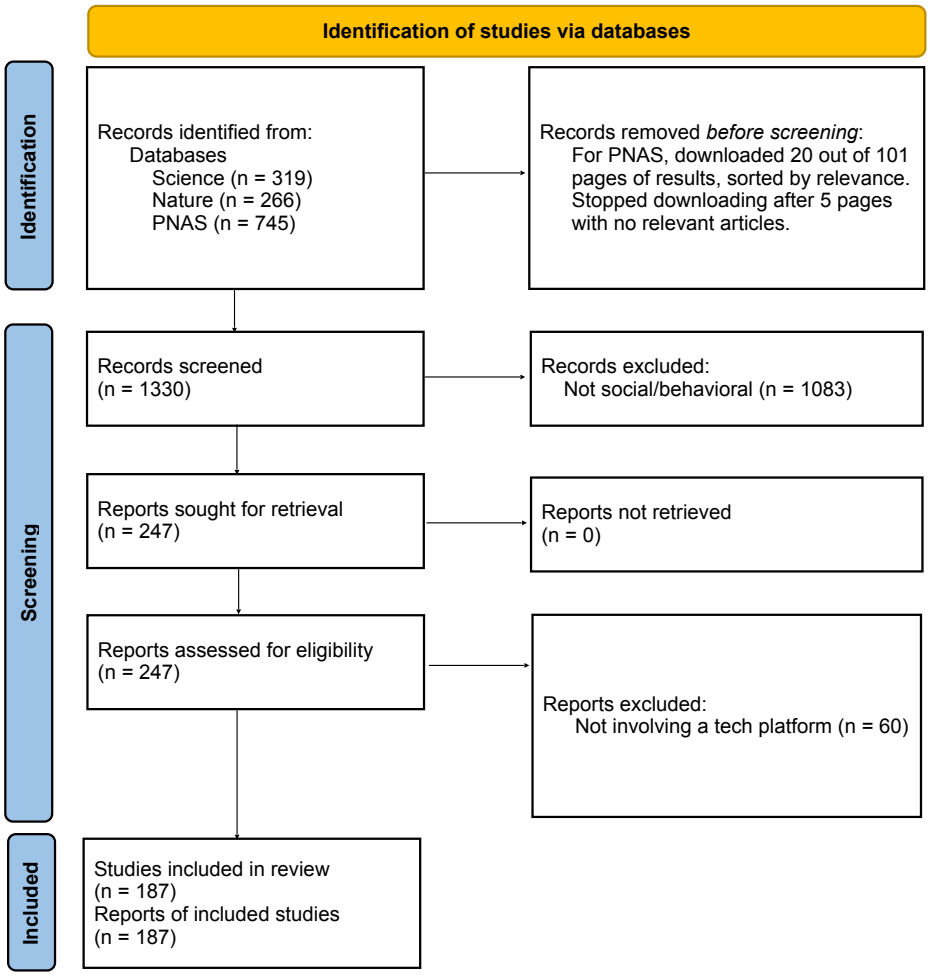

Figure 1. PRISMA (2000) diagram for the collection and analysis of social/behavioral studies about technology platforms published by Nature, Science, and PNAS from January 2004 through August 2025.

Table 1: Results of the coding process for two coders, among the population of all 1,330 articles published between January 2004 and August 2025 returned by Nature, Science, and PNAS search results.

Variable	**Cohen's Kappa**	**Count**	**% of Total**
Social / Behavioral	0.73	247	19%
Technology Platform	0.69	189	14%
Total eligible	NA	187	14%
Variable	**Cohen's Kappa**	**Count**	**% of Eligible**
Wikipedia	0.57	20	11%
Required Platform Cooperation	0.48	63	34%
Scraping or API	0.32	120	64%
Data Collection Still Possible	0.28	52	28%

Table 2: Counts and percentages of articles that were replicable in the population of all social and behavioral articles involving technology platforms published in Nature, Science, and PNAS between 2004 and 2025.

Category		**Total**	**Replicable**	**Not Replicable**
2*__All__	Count	187	49	138
	Row %		**26%**	74%
2*__Required Platform Cooperation__	Count	63	0	63
	Row %		**0%**	100%
2*__Scraping or API__	Count	120	38	82
	Row %		**32%**	68%
2*__Wikipedia__	Count	20	12	8
	Row %		**60%**	40%

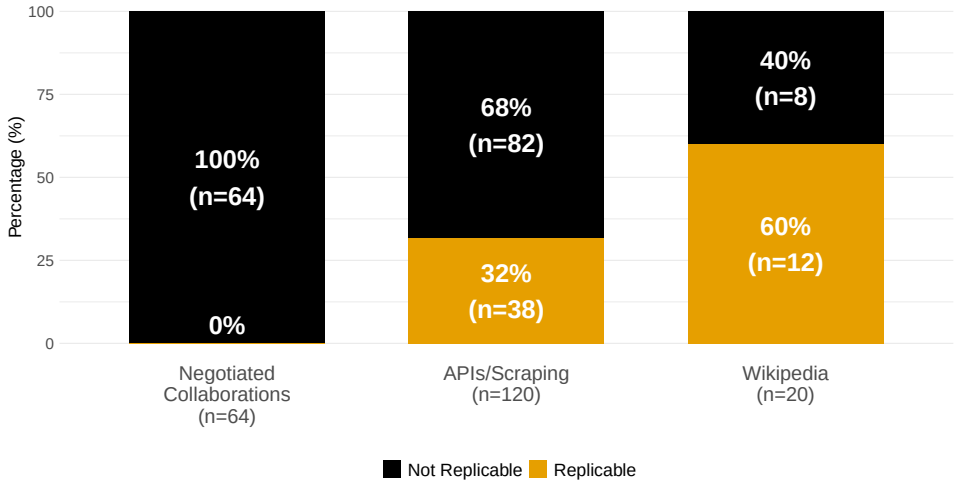

Figure 2. Counts and percentages of articles that were replicable within each article type, for social and behavioral articles involving technology platforms published in Nature, Science, and PNAS between 2004 and 2025.

Results

Among the population of 187 general science journal articles in the social and behavioral sciences involving technology platforms, only one in four could be replicated today without a special arrangement with a technology company (Table 2). Negotiated collaborations with technology firms account for 46% of papers that could not be replicated. Among articles that relied on widely-available APIs or scraping techniques for data collection, 68% could not be replicated today, largely due to actions by companies to restrict access or legally-threaten scraping operations. In contrast, 60% of articles that use noncommercial, Wikipedia data could be replicated today. Among these articles that use Wikipedia data, the cause of replication difficulties tends to derive from a reliance on data from commercial platforms in addition to Wikipedia.

Discussion

In this paper, we observe a serious replicability problem that scholars have warned about for over fifteen years (Lazer et al., 2009; Freelon, 2018; Lazer, 2015; Munger, 2019). As we observe, the ecosystem of computational social science, as visible in the most prominent and publishers

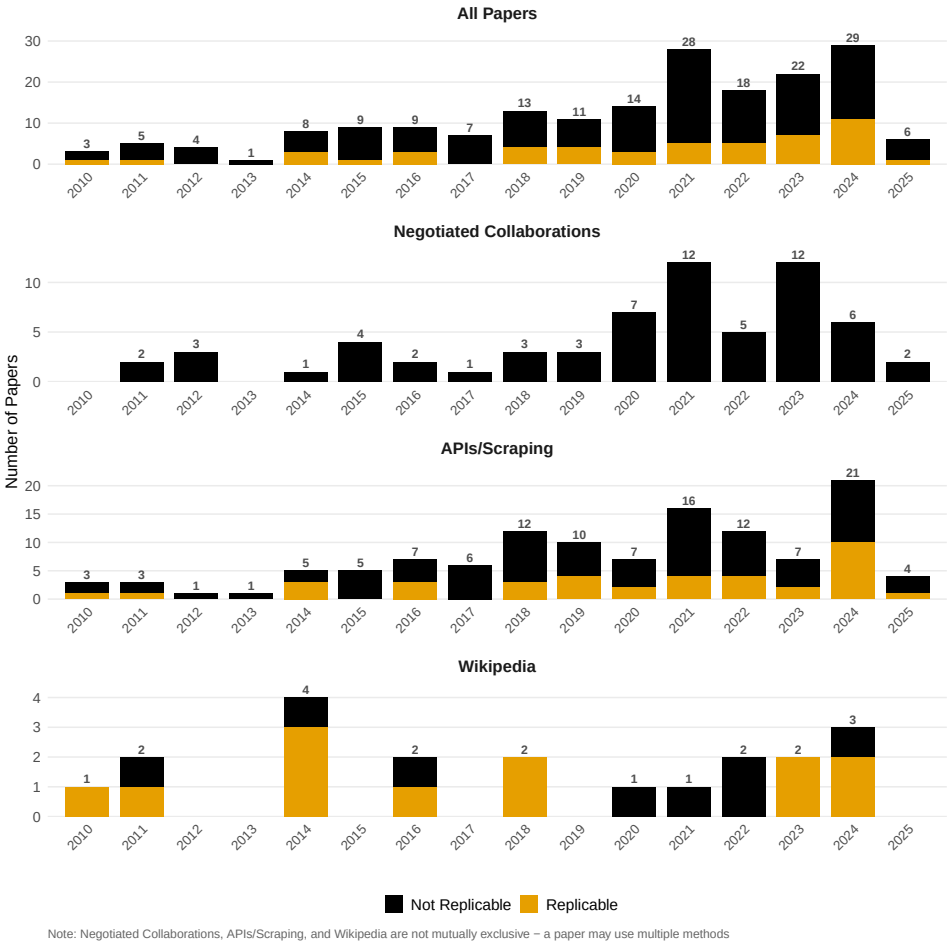

Figure 3. Yearly counts and percentages of articles that were replicable within each article type, for social and behavioral articles involving technology platforms published in Nature, Science, and PNAS between 2004 and 2025.

of such work, is adding to scientific uncertainty rather than scientific knowledge by publishing work that can never be practicably repeated. These results suggest that we are indeed close to a scientific dystopia, where most computational social science is proprietary and occurring within companies; and the most prominent public facing scholarship cannot be replicated (Figure 2, Figure 3).

We note multiple limitations from this study's focus on a small number of prominent general scientific journals. This is a very distinct, and tiny, subset of the relevant literature. However, it is a disproportionately important subset, with a much higher level of visibility both within and outside of the academy than even the most prominent of disciplinary journals. We also note that the comparison of commercial and non-commercial platforms should be exercised with care. This study cannot answer the counter-factual of what would have happened differently if people had conducted the same study on an identical non-commercial platform. It is not possible to "control" for the questions being asked, for example; and by their very nature, commercial platforms will offer very different foci for study than noncommercial. Further, we note that while commercial interests almost certainly play a role in limiting data access on commercial platforms, other factors likely play a role as well (such as protecting user privacy). These factors are not independent from each other, with privacy restrictions directly resulting from abuses of data use by commercial interests (Freelon, 2018). A more comprehensive investigation would consider a wider range of journals and account for the interrelations of multiple factors.

While this purely quantitative observational analysis cannot explain the full reasons for this state of uncertain evidence, we do observe factors that strongly correlate with un-replicable research. The first problem is corporate control over what research can be allowed. None of the questions that previously required specially-negotiated corporate access can be repeated again without similar permission from companies. The second problem is corporate control over data through prohibitive costs, restrictions on scraping, or legal threats. While 32% of studies that involved scraping or APIs are still replicable today, the vast majority of such studies are no longer possible. The third problem involves the pervasiveness of commercial technology platforms that are incentivized to restrict researcher access. Research based on the non-commercial platform Wikipedia had a replicability rate that was twice as high as studies that involved scraping or API access.

What can researchers do about this replication problem? One finding from this study is that public goods such as Wikipedia that have a public mission to advance knowledge have more reliably contributed to repeatable science over time than commercial platforms. Beyond Wikipedia, researchers can develop and maintain research software infrastructures that support platform-independent research (Matias and Mou, 2018; Feal et al., 2024). In 2026 as multiple governments consider domestic technology platforms as alternatives to U.S. corporations, researchers have a unique opportunity to support a form of digital sovereignty that advances scientific knowledge as a public good (Killock, 2026). In cases where studying commercial technology platforms is important, researchers can support policies that require companies to publicly release the results of randomized trials (Bengani, 2024) or require companies to comply with transparency requests that support scientific research (Sekwenz and Gsenger, 2025).

Overall, we show that computational social science has become another area of research where commercial forces have partially hindered the progress of science despite early hopes that they might be mutually beneficial. We hope this finding serves as a call to action for scientists considering how our individual and collective efforts can best advance replicable knowledge.

Acknowledgments

This study was supported by funds from the Heising-Simons Foundation, the John D. and Catherine T. MacArthur Foundation, the Ford Foundation, and the John S. and James L. Knight Foundation. We are also grateful to Vivian Pan for technical support on access to the PNAS archives.

References

Aubin Le Quéré, M. and Matias, J. N. (2025). When curiosity gaps backfire: effects of headline concreteness on information selection decisions. *Scientific Reports*, 15(1):994.

Bak-Coleman, J., O'Connor, C., Bergstrom, C., and West, J. (2025). The risks of industry influence in tech research. *arXiv preprint arXiv:2510.19894*.

Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., Donges, J. F., Galesic, M., Gersick, A. S., Jacquet, J., et al. (2021). Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, 118(27):e2025764118.

Bengani, Jonathan Stray, P. L. T. (2024). Experiments are the Best Kind of Transparency. *Tech Policy Press*.

Brodeur, A., Mikola, D., Cook, N., Fiala, L., Brailey, T., Briggs, R., De Gendre, A., Dupraz, Y., Gabani, J., Gauriot, R., et al. (2026). Reproducibility and robustness of economics and political science research. *Nature*, 652(8108):151–156.

Buzbas, E. O. and Devezer, B. (2024). Statistics in service of metascience: Measuring replication distance with reproducibility rate. *Entropy*, 26(10):842.

Buzbas, E. O., Devezer, B., and Baumgaertner, B. (2023). The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science*, 10(3):221042.

de Vreese, C. and Tromble, R. (2023). The data abyss: How lack of data access leaves research and society in the dark. *Political Communication*, 40(3):356–360.

Dommett, K. and Tromble, R. (2022). Advocating for platform data access: Challenges and opportunities for academics seeking policy change. *Politics and Governance*, 10(1):220–229.

Dörr, T., Nagpal, T., Watts, D., and Bail, C. (2025). A research agenda for encouraging prosocial behaviour on social media. *Nature Human Behaviour*, pages 1–9.

Feal, A., Gleason, J., Goel, P., Radford, J., Yang, K.-C., Basl, J., Meyer, M., Choffnes, D., Wilson, C., and Lazer, D. (2024). Introduction to national internet observatory. In *Workshop Proceedings of the 18th International AAAI Conference on Web and Social Media*, page 73.

Freelon, D. (2018). Computational research in the post-api age. *Political Communication*, 35(4):665–668.

Freese, J. and Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, 43:147–165.

Galea, S. and Buckley, G. J. (2024). Social media and adolescent mental health: A consensus report of the national academies of sciences, engineering, and medicine.

Gantman, A., Gomila, R., Martinez, J. E., Matias, J. N., Paluck, E. L., Starck, J., Wu, S., and Yaffe, N. (2018). A pragmatist philosophy of psychological science and its implications for replication. *Behavioral and Brain Sciences*.

Gilmore, A. B., Fabbri, A., Baum, F., Bertscher, A., Bondy, K., Chang, H.-J., Demaio, S., Erzse, A., Freudenberg, N., Friel, S., et al. (2023). Defining and conceptualising the commercial determinants of health. *The Lancet*, 401(10383):1194–1213.

Killock, J. (2026). The case for Digital Sovereignty and the Digital Commons.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181.

Lazer, D. (2015). The rise of the social algorithm. *Science*, 348(6239):1090–1091.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Computational social science. *Science*, 323(5915):721–723.

Matias, J. N. (2023). Humans and algorithms work together—so study them together. *Nature*, 617(7960):248–251.

Matias, J. N. and Mou, M. (2018). Civilservant: Community-led experiments in platform governance. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.

Matias, J. N., Munger, K., Le Quere, M. A., and Ebersole, C. (2021). The upworthy research archive, a time series of 32,487 experiments in us media. *Scientific Data*, 8(1):195.

Matias, J. N. and Price, M. (2025). How public involvement can improve the science of artificial intelligence. *Proceedings of the National Academy of Science*.

Miske, O., Abatayo, A. L., Daley, M., Dirzo, M., Fox, N., Haber, N., Hahn, K. M., Struhl, M. K., Mawhinney, B., Silverstein, P., et al. (2026). Investigating the reproducibility of the social and behavioural sciences. *Nature*, 652(8108):126–134.

Munger, K. (2019). The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media+ Society*, 5(3):2056305119859294.

of the Surgeon General, O. et al. (2023). Social media and youth mental health: The us surgeon general's advisory [internet].

Orben, A. and Matias, J. N. (2025). Fixing the science of digital technology harms. *Science*, 388(6743):152–155.

Oreskes, N. and Conway, E. M. (2011). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA.

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., and Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature human behaviour*, 7(5):812–822.

Sekwenz, M.-T. and Gsenger, R. (2025). The digital services act: Online risks, transparency and data access. *Digital Decade*, pages 115–140.

Shapin, S. and Schaffer, S. (1985). Seeing and believing: The experimental production of pneumatic facts. *Leviathan and the airpump: Hobbes, Boyle, and the experimental life*, pages 22–79.

Shulman, H. C., Markowitz, D. M., and Rogers, T. (2024). Reading dies in complexity: Online news consumers prefer simple writing. *Science Advances*, 10(23):eadn2555.

Tromble, R. (2021). Where have all the data gone? a critical reflection on academic digital research in the post-api age. *Social Media+ Society*, 7(1):2056305121988929.

Young, M., Katell, M., and Krafft, P. (2022). Confronting power and corporate capture at the facct conference. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1375–1386.

Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41:e120.