PNAS LETTER

The limitations of machine learning models for predicting scientific replicability

M. J. Crockett[a,b,1], Xuechunzi Bai[a,c], Sayash Kapoor[d,e], Lisa Messeri[f], and Arvind Narayanan[d,e]

The past decade has witnessed substantial investments in evaluating and improving the replicability of scientific findings (1, 2). In *PNAS,* Youyou, Yang, and Uzzi claim that a machine learning model (MLM; 3) can predict the replicability of entire subfields of psychology based on individual papers' narrative text and reported statistics (4). Here, we highlight five serious limitations of replicability MLMs that ironically mimic several aspects of the psychology replication "crisis" (5; Table 1). Considering these limitations invites us to expand our modes of inquiry in conversations about replicability.

First, training MLMs to reliably predict complex phenomena requires massive datasets, but the available data for training a replicability MLM are limited to <500 existing replication studies in psychology. This training set is orders of magnitude smaller than those used to train MLMs for far simpler tasks than predicting replicability (Fig. 1*A*). Small training sets can result in wide CIs and thus inflate estimates of an MLM's accuracy, just as underpowered samples inflate false-positive findings (1).

Second, these training data disproportionately represent "classic papers by selected authors or specific subfields" (4) (Fig. 1*B*). Nonrepresentative training data seriously limit an MLM's generalizability to new data (5, 6), just as nonrepresentative samples reduce generalizability of psychology findings (1, 5).

Third, developers argue that MLMs could be used to efficiently allocate research funding by assigning "replication likelihood scores" to individual papers (3, 4). With reported error rates up to 30%, replicability MLMs risk falsely assigning low scores to individuals or entire subfields. If such errors are inequitably distributed (e.g., disproportionately stigmatizing subfields with more racial or gender diversity), replicability MLMs could exacerbate existing inequalities in science, joining a long list of past algorithmic injustices (7). And because algorithms are perceived to be more objective than humans (8), MLM-based replication likelihood scores could impose even more stigma on researchers or subfields than human expert predictions of replicability.

Fourth, MLMs predict replicability from superficial features of papers' narrative text, rather than deeper conceptual aspects of the underlying science. If MLMs are used for consequential decisions like allocating funding, this creates incentives for authors to change their paper's style without changing its scientific substance to improve their chances at funding. Such practices would hardly improve scientific replicability even though they might superficially appear to do so.

Fifth, MLMs cannot provide causal explanations for predictions of replicability (9). Explanations are seen as especially important for algorithms that make high-stakes decisions or distribute scarce resources (10). Without an explanation, researchers cannot effectively dispute a low replicability score or adjust their scientific practices to improve it.

Table 1. Replication MLMs ironically recreate several aspects of the replication crisis in psychology

Replicability MLM limitations	Psychology replication crisis
An insufficient number of training samples can result in wide CIs and lead to inflated estimates of classification accuracy (Fig. 1*A*).	Underpowered studies with insufficient sample sizes can inflate false positive findings.
Replicability MLMs are trained on replication studies not representative of all psychology studies, limiting their generalizability (Fig. 1*B*).	Most psychology study participants are not representative of the global population, limiting generalizability of study findings.
Replicability MLMs based on superficial text features are vulnerable to gaming, e.g., by changing text style to achieve a higher replicability score.	Researcher degrees of freedom make data analysis vulnerable to gaming, e.g., by running multiple analyses to achieve a lower *P*-value ("p-hacking")
Selective reporting of results (e.g., absence of error bars on AUC metrics) provides false confidence in replicability MLM accuracy.	Selective reporting of results (e.g., only reporting statistics consistent with a paper's hypotheses) gives false confidence in a paper's claims.
Replicability MLMs claim to provide "discipline-wide" predictions of replicability despite relying on data nonrepresentative of psychology as a whole.	Psychology studies claim to provide insights into "human nature" despite relying on data nonrepresentative of all humans.

For a more detailed discussion of these issues, see refs. 1 and 5.

Author affiliations: [a]Department of Psychology, Princeton University, Princeton, NJ 08540; [b]University Center for Human Values, Princeton University, Princeton, NJ 08544; [c]School of Public and International Affairs, Princeton University, Princeton, NJ 08540; [d]Department of Computer Science, Princeton University, Princeton, NJ 08540; [e]Center for Information Technology Policy, Princeton University, Princeton, NJ 08544; and [f]Department of Anthropology, Yale University, New Haven, CT 06511

Author contributions: X.B. and S.K. analyzed data; and M.J.C., X.B., S.K., L.M., and A.N. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: mj.crockett@princeton.edu.

Published August 7, 2023.

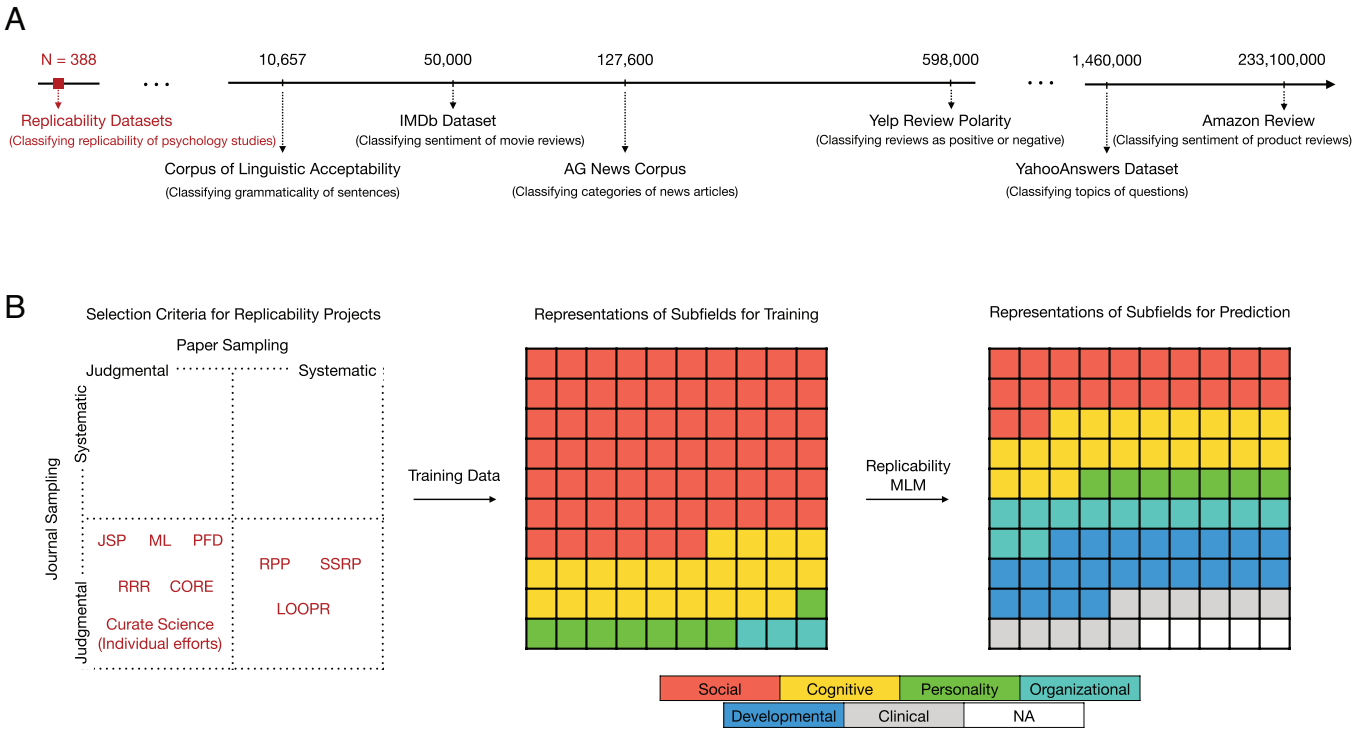

Fig. 1. Limitations of training datasets for replicability MLMs. (*A*) The training dataset for a psychology replicability MLM (4) consists of N = 388 replication studies. By contrast, vastly larger datasets are commonly used to train MLMs to perform a range of much simpler tasks than predicting replicability, illustrated here with examples from a popular python library (https://pytorch.org/text/stable/datasets.html). Each example dataset is annotated with its number of labeled instances and a prominent task it can be used to train. For the Amazon Review data, we used the most up-to-date dataset compiled in 2018. (*B*) The available data for training a replicability MLM is not representative of psychology research in several respects. The *Left* panel illustrates that the papers used to train the psychology replicability MLM (4) are dominated by judgmental sampling of journals and papers, a nonrandom sampling method that is more susceptible to bias than systematic sampling; notably, no psychology replication projects used to train the replicability MLM employed systematic sampling across both journals and papers. RPP: Reproducibility Project Psychology; RRR: Registered Replication Report; ML: Many Labs; JSP: Replications of Important Results in Social Psychology; SSRP: Social Sciences Replication Projects; LOOPR: The Life Outcomes of Personality Replication Project; CORE: Mass Replications and Extensions by the Collaborative Open-science REsearch team; Curate Science: Individual effort projects; PFD: PsychFileDrawer. The middle and right panels illustrate differences in the subfields represented in the training dataset compared with the subfields used to evaluate replicability in psychology as a whole.

Overall, these limitations mean that replicability MLMs cannot offer shortcuts to building a more credible psychological science. However, they helpfully nudge us to reconsider whether optimizing scientific tools for quantification and prediction always leads to a better understanding of psychology. A narrow focus on quantitative replication necessarily constrains what aspects of psychology can be known. Instead, conversations about replication need to broaden engagement with modes of scholarship that resist reducing psychology to that which can be predicted by algorithms.

1. B. A. Nosek *et al.*, Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **73**, 719–748 (2022).
2. A. Russell, "Systematizing confidence in open research and evidence (SCORE)" (Tech. Rep., Defense Advanced Research Projects Agency, Arlington, VA, 2019).
3. Y. Yang, W. Youyou, B. Uzzi, Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10762–10768 (2020).
4. W. Youyou, Y. Yang, B. Uzzi, A discipline-wide investigation of the replicability of Psychology papers over the past two decades. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2208863120 (2023).
5. J. Hullman, S. Kapoor, P. Nanayakkara, A. Gelman, A. Narayanan, "The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning" in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, New York, NY, 2022), pp. 335–348.
6. A. Paullada, I. D. Raji, E. M. Bender, E. Denton, A. Hanna, Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* **2**, 100336 (2021).
7. R. Benjamin, *Race after Technology: Abolitionist Tools for the New Jim Code* (Polity, 2019).
8. T. Gillespie, "The relevance of algorithms" in *Media Technologies: Essays on Communication, Materiality, and Society*, T. Gillespie, P. Boczkowski, K. Foot, Eds. (Oxford University Press, Oxford, United Kingdom, 2014), p. 167.
9. C. Barabas, M. Virza, K. Dinakar, J. Ito, J. Zittrain, "Interventions over predictions: Reframing the ethical debate for actuarial risk assessment" in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (PMLR, 2018), vol. 81, pp. 62–76.
10. A. M. Nussberger, L. Luo, L. E. Celis, M. J. Crockett, Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence. *Nat. Commun.* **13**, 5821 (2022).