PNAS LETTER

Reply to Crockett et al. and Mottelson and Kontogiorgos: Machine learning's scientific significance and future impact on replicability research

Wu Youyou[a], Yang Yang[b], and Brian Uzzi[c,d,1]

Albert Einstein observed, "We cannot solve our problems with the same thinking we used when we created them." Likewise, the replication crisis will benefit from diverse methods that explore its potential solutions (1–3). Our text-based machine learning model (MLM) (4, 5) complements manual, prediction market, and econometric approaches to replication prediction (2, 6). Moreover, it is uniquely scalable, affordable, and replicable in its own right. Future research, data, and methodological advances are bound to further improve it.

We thank *PNAS*, Crockett et al. (7), and Mottelson and Kontogiorgos (M&K) (8) for supporting scientific debate and for the opportunity to show why our MLM meets the standards of good science.

Crockett et al. (7) argued that our MLM model should be shunned because its relatively small training sample will artificially inflate the accuracy of its predictions. The link between sample size and accuracy is hypothetical, not inevitable. We tested for it and showed that the "proof is in the pudding" by empirically doing extensive out-of-sample tests. Our MLM, for example, has accuracy that is on par with the gold standard of human-based prediction methods, prediction markets (2, 5).

Crockett et al. claimed that we endorsed the use of our MLM for funding decisions. We never made this claim. Rather, we proposed, following Isager et al. (3) that MLMs be used with other replication methods to accelerate replication research by rank-ordering papers for manual replication tests under resource constraints.

Crockett et al. asserted that MLMs are likely to "stigmatize subfields with more racial or gender diversity" (a contention weirdly supported by citing a Jim Crow law). Our analysis does the opposite by using data and tests to address alleged relationships in the replication literature (9–12). For example, our MLM investigated claims about replication failure and academic elitism, methods, and psychology subfields in over 14,000 papers, the largest sample of papers, authors, institutions, and media mentions ever examined for replicability. We discovered that replication rates from faculty at elite and nonelite universities do not differ, a researcher's productivity and impact positively correlate with their paper's estimated replicability, and that a paper's media coverage and replicability may be linked.

M&K (8) raised the concern that our training sample of manually replicated papers was tainted by papers that were double-counted and had conflicting manual replication outcomes. M&K misread our procedures and data. Generally, manual replication studies replicate one, rather than all of a paper's separate experiments. The "double-counted" papers in our sample appear more than once because they had multiple experiments that underwent separate manual replications, some of which had experiments with divergent replication outcomes.

M&K's replication of our model showed a small decrease in performance compared to our reported numbers. The difference is not unexpected given that M&K's replication differed extensively from our model. Per Table 1, their replication used a different training sample, the full text of a paper, and different methods to quantify texts, conduct machine learning, and evaluate performance.

We welcome the improvement of our model by other researchers. We are pleased that M&K attempted to improve our model, but we reasonably disagree with M&K's choices for "improvements." First, M&K's "improved" model uses $P \approx$ 20,000 variables, which massively exceeds the number of data points ($N < 400$). This practice is discouraged because it can lead to overfitting (13). Second, M&K use a random forest model, whereas we used an ensemble method (Table 1), which is superior for small training samples (14–16) and averages the predictions of different methods. Nevertheless, despite the discrepancies, their findings are broadly consistent with our findings, indicating that our MLM is demonstratively robust to variations in input, throughput, and output criteria.

M&K reported that our MLM predictions wrongly correlate with a study's number of words and nouns. This is not the case. When we redo their analysis using our model, the replication scores in the prediction sample ($N = 14{,}126$) do not correlate with the number of words ($r = 0.01$) or with the number of nouns ($r = 0.02$). Relatedly, M&K claimed that our predictions correlated with over a third of linguistic features. Besides the fact that features that correlate with a model's outcome do not necessarily mean that they are the features that drive the model (if so, MLMs would be easily interpretable), the correlations appear to

Author affiliations: [a]Department of Psychology and Human Development, Institute of Education, University College London, London WC1H 0AL, United Kingdom; [b]Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556; [c]The Kellogg Graduate School of Management, Northwestern University, Evanston, IL 60208; and [d]The Northwestern University Institute on Complex Systems and Data Science, Northwestern University, Evanston, IL 60208

Author contributions: W.Y., Y.Y., and B.U. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: uzzi@northwestern.edu.

Published August 7, 2023.

Table 1. Comparing Youyou et al. and Mottelson & Kontogiorgos' attempted replication[*]

Differences in	Youyou et al. (4)	Mottelson & Kontogiorgos' (8) attempted replication
Data		
Training sample	Each record is a single study or a set of studies in a paper that was targeted for replication and has a single outcome; N = 388.	Each record is a paper, often containing multiple studies; N = 348
Text data	Only use text related to the target study that had been manually replicated, rather than the full paper's text.	Without justification, the full texts of a paper are used, including acknowledgments and footnotes.
Modeling		
Software	Java (Weka)	Python (Scikit-learn)
Quantifying text	Used both TF (term-frequency) document and term frequency–inverse document frequency (TF-IDF) document vectors, along with word vectors.	TF-IDF document vectors were used, along with word vectors.
Machine-learning model	An ensemble learning model combining the bagging with random forest and bagging with simple logistic regression.	Random forest
Performance[†]		
Performance evaluation	One hundred rounds of repeated threefold cross-validation and report aggregated performance.	One round of threefold cross-validation.
Mean area under the curve (AUC)	0.72 (*SD* = 0.02)	0.68
Mean accuracy	0.68 (*SD* = 0.02)	0.62
Aggregated AUC	0.74	
Aggregated accuracy	0.68	
Correlations with linguistic features (in the full prediction sample, N = 14,126)[‡]		
	Youyou et al. (4)	**Our replication of M&K's altered model, using random forest with TF-IDF document vectors only.**
Number of words	$r = 0.01$	$r = 0.13$
Number of nouns	$r = 0.02$	$r = 0.13$
% of linguistic features with $r \geq 0.1$ and $P <$ Bonferroni corrected alpha	9%	34%

[*] The summary reflects our understanding of M&K's implementation based on their code, which can be different from their description in the letter or their actual implementation.

[†] Youyou et al. (4), averaged the predicted replication scores for each paper from 100 rounds of cross-validations and reported the aggregated AUC and accuracy; Here, we also calculate the AUC and accuracy for each round of the 100 cross-validation and report the mean and SD of these metrics (i.e., mean AUC and mean accuracy).

[‡] M&K (8) conduct the correlational analysis on a subsample (n = 98). We instead analyze and report results from the full prediction sample (N = 14,126). Here, we also reestimate results for the full prediction sample using our replication of M&K's altered model. Since a large sample can easily produce statistically significant results, we consider effect sizes and impose the $r \geq 0.1$ rule for practical significance.

be related to their "proposed improved model" rather than our model. If they had used our model and done a Bonferroni correction for multiple comparisons, under 1/10th of linguistic features and our model's predictions correlate at $r > 0.1$.

A related point, hinted at by Crockett et al., is whether MLMs will be reversed engineered and gamed. We think this is unlikely. The issue comes down to trust. If deception is the aim, there are easier ways for a researcher to fake results (17, 18). We believe in the honesty of researchers and in their desire to improve science with tools that have the potential to do so.

In closing, the facts show that Crockett et al.'s and M&K's comments have produced a valuable exchange of ideas. We showed that their hypothetical criticisms were not borne out by the data and that changes in the data and methods used in their replications and remodeling explain why their results do not exactly match our published results. As we step back from this response, we agree that better data are welcomed and that new models should be experimented with and adopted judiciously. At the same time, while we wait for impeccable data, there are risks in letting the "perfect be the enemy of the good" when nonreplicating findings spread freely through the literature (Fig. 1).

We advocate not to throw the baby out with the bathwater. We welcome new theory, data, and methodological pairings of humans minds and machines and look forward to more intellectual exchanges around machine learning and its scientific benefits and possibilities. We trust that

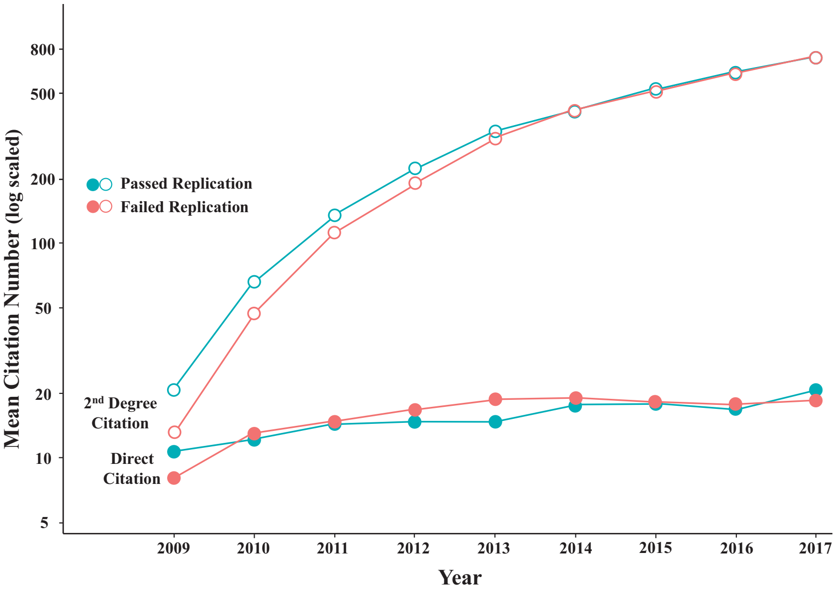

Fig. 1. Nonreplicating studies are cited as highly as replicating studies. We measured the direct citations and second-degree citations (citations to papers that have cited a nonreplicating study) of the 100 manually replicated studies published in 2008 (19). The plot indicates that the 61 papers that failed to replicate are cited at the same yearly rate as papers that successfully replicated for direct and second-degree citations (from ref. 5).

thoughtfully managed human and machine partnerships can improve science's understanding and prediction of replicability in psychology faster than either humans or machines can do on their own.

ACKNOWLEDGMENTS. This material is based upon work supported by the Air Force Office of Scientific Research under Minerva award number FA9550-19-1-0354, Northwestern Institute on Complex Systems, Northwestern University, and the Kellogg Graduate School of Management, Northwestern University.

1. J. L. Tackett, C. M. Brandes, K. M. King, K. E. Markon, Psychology's replication crisis and clinical psychological science. *Annu. Rev. Clin. Psychol.* **15**, 579–604 (2019).
2. B. A. Nosek *et al.*, Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **73**, 719–748 (2021).
3. P. M. Isager *et al.*, Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychol. Methods* **28**, 438–451 (2021).
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. 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).
6. O. E. Gundersen, The fundamental principles of reproducibility. *Philos. Trans. A Math. Phys. Eng. Sci.* **379**, 20200210 (2021).
7. M. J. Crockett, X. Bai, S. Kapoor, L. Messeri, A. Narayanan, The limitations of machine learning models for predicting scientific replicability. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2307596120 (2023).
8. A. Mottelson, D. Kontogiorgos, Replicating replicability modelling of Psychology. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2309496120 (2023).
9. B. A. Spellman, D. Kahneman, What the replication reformation wrought. *Behav. Brain Sci.* **41**, e149 (2018).
10. S. T. Fiske, A call to change science's culture of shaming. *APS Obs.* **29**, 5–6 (2016).
11. M. Baker, 1,500 scientists lift the lid on reproducibility. *Nat. News* **533**, 452 (2016).
12. M. Gordon *et al.*, Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R. Soc. Open Sci.* **7**, 200566 (2020).
13. R. O. Duda, P. E. Hart, *Pattern Classification* (John Wiley & Sons, 2006).
14. P. Kokol, M. Kokol, S. Zagoranski, Machine learning on small size samples: A synthetic knowledge synthesis. *Sci. Prog.* **105**, 368504211029777 (2022).
15. T. G. Dietterich, "Ensemble methods in machine learning" in *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1* (Springer, 2000), pp. 1–15.
16. M. Mohri, A. Rostamizadeh, A. Talwalkar, *Foundations of Machine Learning* (MIT press, 2018).
17. H. Ledford, R. Van Noorden, High-profile coronavirus retractions raise concerns about data oversight. *Nature* **582**, 160–161 (2020).
18. S. F. Lu, G. Zhe Jin, B. Uzzi, B. Jones, The retraction penalty: Evidence from the web of science. *Nat. Sci. Rep.* **3**, 3146 (2013).
19. B. A. Nosek *et al.*, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).