Estimating the deep replicability of scientific findings using human and artificial intelligence

Yang Yang[a,b,1], Wu Youyou[a,b,1], and Brian Uzzi[a,b,2]

[a]Northwestern University Institute on Complex Systems, Evanston, IL 60208; and [b]Kellogg School of Management, Northwestern University, Evanston, IL 60208

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved March 26, 2020 (received for review May 31, 2019)

Replicability tests of scientific papers show that the majority of papers fail replication. Moreover, failed papers circulate through the literature as quickly as replicating papers. This dynamic weakens the literature, raises research costs, and demonstrates the need for new approaches for estimating a study's replicability. Here, we trained an artificial intelligence model to estimate a paper's replicability using ground truth data on studies that had passed or failed manual replication tests, and then tested the model's generalizability on an extensive set of out-of-sample studies. The model predicts replicability better than the base rate of reviewers and comparably as well as prediction markets, the best present-day method for predicting replicability. In out-of-sample tests on manually replicated papers from diverse disciplines and methods, the model had strong accuracy levels of 0.65 to 0.78. Exploring the reasons behind the model's predictions, we found no evidence for bias based on topics, journals, disciplines, base rates of failure, persuasion words, or novelty words like "remarkable" or "unexpected." We did find that the model's accuracy is higher when trained on a paper's text rather than its reported statistics and that n-grams, higher order word combinations that humans have difficulty processing, correlate with replication. We discuss how combining human and machine intelligence can raise confidence in research, provide research self-assessment techniques, and create methods that are scalable and efficient enough to review the ever-growing numbers of publications—a task that entails extensive human resources to accomplish with prediction markets and manual replication alone.

replicability | machine learning | computational social science

When 100 papers from top psychology journals were randomly selected and manually tested for their replicability using the same procedures as the original studies [the "Reproducibility Project: Psychology" (RPP)], 61 of 100 papers failed the test (1). Additional replication studies in psychology, medicine, and economics similarly determine that papers more often fail than pass replicability tests (2–7) and that, once published, these nonreplicable results permeate the literature as quickly as the results of replicating studies. In the RPP, the 61 nonreplicating psychology papers were cited as frequently as the 39 papers that passed replication, a pattern that continued even after the nonreplicating results were identified in manual replication tests (Fig. 1). Beyond tainting the scientific literature (8), nonreplicability has and continues to diminish public funding support (9, 10), engender pessimism toward the validity and value of scientific findings, and drive up costs. Half of scientists surveyed consider replicability a crisis, and a third distrust half the papers in their fields (11). The annual costs of replication failures are estimated to be $28 billion (12).

Conducting manual replication research is time-consuming and involves high opportunity costs for scientists. For example, individual replication studies in the RPP took an average of 314 d from the claim date to complete analysis. Consequently, research has turned to evaluating the accuracy of different methods for estimating a study's replicability. This approach is spearheaded by the Defense Advanced Research Projects Agency's (DARPA's) Systematizing Confidence in Open Research and Evidence (SCORE) program, which funds research that develops new approaches for prioritizing studies to be manually replicated (13).

Three methods to estimate a study's risk of passing or failing replication have been assessed: the statistics reported in the original study (e.g., *P* values and effect size, also known as "reviewer metrics"), prediction markets, and surveys (10). Currently, investigations of reviewer metrics report pairwise correlations of 0.044 to 0.277 between individual metrics and replicability (1, 7, 14, 15) (details are in *SI Appendix*, Section S1 and Table S1). In a prediction market, hundreds of researchers "bet" whether a published study will successfully replicate in a future manual replication test as well as record their subjective predictions about a study's replicability in surveys. A paper's final "market price" and survey responses reflect the crowd judgment of a paper's replicability (10). Markets and surveys produce varied, but generally high, levels of predictive accuracy in small samples. Prediction markets have examined a variety of samples that have included between 21 and 41 papers and reported accuracy levels between 0.71 to 0.85 across all samples (7, 10, 16, 17). Despite their high accuracy, prediction markets and surveys are costly to scale. They require the time commitment of many "judges/experts," take almost a year to run, and necessitate additional overhead specific to the effort (18).

We conducted an exploratory investigation into the use of machine learning to estimate a study's replicability. Our research aligns with the research goals of developing approaches for estimating a study's risk for replication failure and can be used to

Significance

After years of urgent concern about the failure of scientific papers to replicate, an accurate, scalable method for identifying findings at risk has yet to arrive. We present a method that combines machine intelligence and human acumen for estimating a study's likelihood of replication. Our model—trained and tested on hundreds of manually replicated studies and out-of-sample datasets —is comparable to the best current methods, yet reduces the strain on researchers' resources. In practice, our model can complement prediction market and survey replication methods, prioritize studies for expensive manual replication tests, and furnish independent feedback to researchers prior to submitting a study for review.

Author contributions: Y.Y. and B.U. designed research; Y.Y., W.Y., and B.U. performed research; Y.Y., W.Y., and B.U. analyzed data; and Y.Y., W.Y., and B.U. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]Y.Y. and W.Y. contributed equally to this work.

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

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1909046117/-/DCSupplemental.

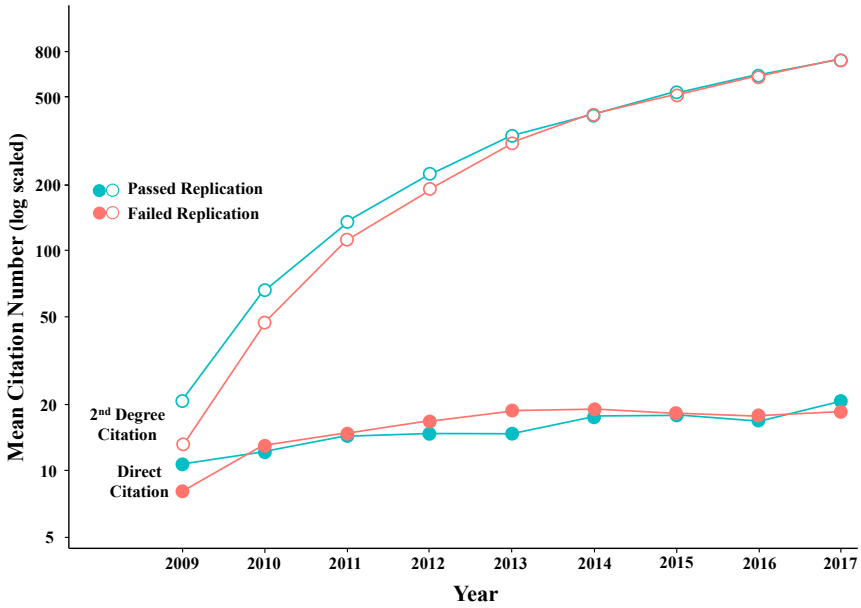

Fig. 1. Citation rates of replicating and nonreplicating studies are indistinguishable. We measured the direct citations and second-degree citations (citations to papers that have cited a nonreplicating study) of the 100 studies published in 2008 and reported in the RPP (1) in 2015. The 61 of 100 papers that failed to replicate are cited at the same yearly rate as papers that successfully replicated (per Google Scholar) for direct and second-degree citations, suggesting that papers with fluke results are incorporated into future research as much as papers with reproducible facts.

prioritize studies submitted to manual replication. If machine intelligence has predictive utility similar to other methods (like prediction markets) but can, in addition, scale to review larger samples of papers while also lowering costs (machine time, not people time) and making faster, comparably accurate predictions (minutes instead of months), it can potentially advance current programs of replication research.

Our model estimates differences in predictive power according to the information used to construct the model—the paper's narrative, papers reported statistics, or both. The first model uses only the study's narrative text under the assumption that machine intelligence can identify in text latent replicability information conveyed by the authors of the study (19, 20). The second model uses just a paper's reviewer metrics (i.e., no text). The third model uses both text and reviewer metrics. After calibrating the model with the training data, we test it on hundreds of out-of-sample, manually replicated studies. Table 1 presents the study's data sources, which include 2 million abstracts from scientific papers (21) and 413 manually replicated studies from 80 journals. These data, while meager by most machine learning study standards, nonetheless encompass the available majority of replication projects across disciplines, topics, journals, methods, and publication dates that include the definitive pass/fail information needed for training and testing models (see *SI Appendix*, Section S2 for details).

Analysis

We began with the narrative-only model, which was trained on the RPP's manually replicated papers. Our methodology involves three stages of development, and in stage 1 of this analysis, we obtained the manuscripts of all 96 studies and stripped each paper of all nontext content (e.g., authors, numbers, graphics, etc.; *SI Appendix*, Section S3.1).

In stage 2, a neural-network-based method—word2vec (29) with standard settings—was used to quantitatively represent a paper's narrative content by defining the quantitative relationship (co-occurrence) of each word with every other word in the corpus of words in the training set. First, to establish a reliable estimate of word co-occurrences, we used data from the Microsoft Academic Graph (MAG) to train our word2vec model on 2 million scientific article abstracts that were published between 2000 and 2017 (21). This training set has about 200 million tokens (words, letters, or symbols) and 18 million sentences. Second, we used word2vec to extract from the full matrix of word pairings a smaller matrix of underlying "factors" that more economically represent the interrelationships among all words in the corpus (30). As a result, each word is represented by a 200-dimension vector that defines its relationship with all other words in the corpus. We used 200 dimensions for two reasons. Prior analysis has used 200 dimensions in training word2vec models with large data (31), and retraining the word2vec model on 100 dimensions produced similar cross-validation results. Third, we multiplied the normalized frequency of each word in a paper by its corresponding word vector. These steps resulted in a final paper-level vector representing the unique linguistic information of each of the 96 papers (see *SI Appendix*, Section S3 for details).

In stage 3, we predicted a study's manually replicated outcome (pass or fail) from its paper-level vector using a simple ensemble model of bagging with random forests and bagging with logistic (32–34), which works well with small datasets (see *SI Appendix*, Section S3.2.2 for details). This simple ensemble model generates predictions of a study's likelihood of replicating [0.0, 1.0] using threefold repeated cross-validation. It is trained on a random draw of 67% of the papers to predict outcome for the remaining 33%, and the process is repeated 100 times with different random splits of training sets vs. test sets. Hence, each study has a unique distribution of 100 predictions between 0.0 and 1.0. The random forests and logit models produced similar results (see *SI Appendix*, Section S3.2 for details).

Results

Fig. 2 visualizes the raw prediction data for the narrative-only machine learning model. To evaluate the model's accuracy in predicting the true outcome of the manual replication, we calculated each paper's average prediction from its 100 rounds.

Table 1. Training and out-of-sample test datasets

	Project	No. of studies	Discipline	No. of journals	Original study year	Original study methodology	Replication method	Replication report publication status	Training vs. test
1	RPP (1)	96	Cog psych; social/ personality psych	3	2008	Experiments & correlational studies	Single-lab, same method	Published	Training set
2	RRR (22)	8	Cog psych; social psych	4	1988 to 2014	Experiments	Multi-lab, same method	Published	Test set I
3 to 5	ML1 (23), 2 (2), 3 (24)	42	Cog psych; social psych	22	1973 to 2013	Experiments	Multi-lab, same method	Published	
6	JSP (25)	16	Social psych Cog psych; social	7	1999 to 2012	Experiments	Multi-lab, same method	Published	
7	SSRP (16)	21	Social psych; economics	2	2011 to 2015	Experiments	Single lab, same method	Published	
8	Individual efforts (26)	33	Cog psych; social psych	8	1972 to 2013	Experiments	Single-lab/Multi-lab, same method	Published	
9	PFD (27)	57	Cog psych	20	2001 to 2017	Experiments & correlational studies	Single lab, same method	Unpublished, includes class projects	Test set II
10	EERP (7)	18	Economics	2	2011 to 2014	Experiments	Single-lab, same method	Published	Test set III
11	ERW (28)	122	Economics	45	1973 to 2015	Experiments, correlational studies, & modeling	*1*) Same data, same code *2*) New data, same methods *3*) Same data, new methods *4*) New methods, new data	Published	Test set IV
Total		413		80 unique	1972 to 2017				Test set total = 317

For the initial training of the word2vec model, we used 2 million abstracts from the MAG database (21) to estimate the relationships among words in scientific papers. This step established a reliable quantification of word co-occurrences in scientific papers based on 200 million tokens (words, letters, or symbols) and 18 million sentences that was then used to digitally represent the word content of papers used in the analysis (n = 413). The training data (n = 96 papers) and out-of-sample testing datasets (n = 317 papers) used in the analysis encompass available manual replication studies that have reported pass vs. fail information for each paper. These data cover diverse disciplines, journals, publication dates, research methods, replication methods, topics, and published and unpublished reports, as described in the table. The final sample of RPP replications excluded three studies and combined two replications of the same paper into one record for a final sample size of 96 studies (see *SI Appendix*, Section S2 for data details). Cog, cognitive; ML1, Many Labs 1; PFD, Psychfiledrawer; psych, psychology; RRR, registered replication reports; SSRP, Social Sciences Replication Project. JSP, Journal Social Psychology; EERP, Experimental Economics Replication Project.

A typical standard for evaluating accuracy, which is assessed relative to a threshold selected according to the evaluator's interest, is the base rate of failure in the ground-truth data (35). At the base rate threshold we chose, 59 of 96 studies failed manual replication (61%). We found that the average accuracy of the machine learning narrative-only model was 0.68 (SD = 0.034). In other words, on average, our model correctly predicted the true pass or fail outcome for 69% of the studies, a 13% increase over the performance of a dummy predictor with 0.61 accuracy (predict all as nonreplicated).

A second standard approach to assessing predictive accuracy is top-k precision, which measures the number of actual failures among the k lowest-ranked studies (35) based on a study's average prediction. When k is equal to the true failure base rate, the machine learning model's top-k precision is 0.74 (SD = 0.028). In both cases, a simple bag-of-words model did relatively poorly, with an accuracy of 0.60.

To compare the accuracy of the narrative-only model with conventionally used reviewer metrics, we designed a statistics-only model using the same procedure used to design the narrative-only model. Prior research reported only pairwise correlations between replicability and reviewer metrics. Here, we clustered all reviewer metrics into one model under the assumption that the effects of statistics on reviewers' judgments may offer more information as a group than as individual metrics. The reviewer metrics-only model achieved an average accuracy and top-k precision of 0.66 (SD = 0.034) and 0.72 (SD = 0.027), respectively. Though the sample of 96 papers was necessarily small, the distribution of both accuracy and top-k precision values were lower ($P < 0.001$) than the same values of the narrative-only model per the Kolmogorov–Smirnov (KS), Wilcoxon rank-sum, Cramer–von Mises, and Anderson–Darling tests (36, 37) (*SI Appendix*, Section S3.6).

To investigate whether combining the narrative-only and reviewer metrics-only models provides more explanatory power than either model alone, we trained a model on a paper's

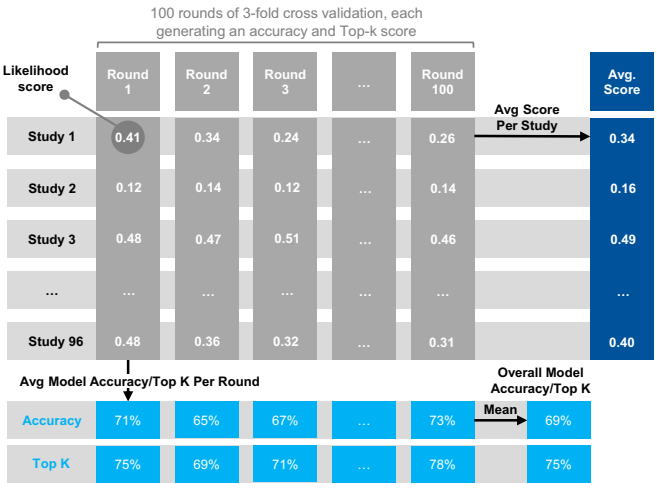

Fig. 2. Machine-predicted replication scores for the 96 RPP studies. Each row in a gray column represents the machine learning model's prediction [0.0, 1.00] of whether a study will replicate or not. There are 100 columns per study—one column for each round of cross-validation. Each cell in a dark blue column displays the average prediction score (and ranking) for each study across its 100 rounds of threefold cross-validation. Each cell in light blue shows the average accuracy and top-k precision for each one of the 100 rounds of cross-validation, with the last column in light blue showing the grand mean of the machine learning model's overall accuracy and precision scores. Avg, average.

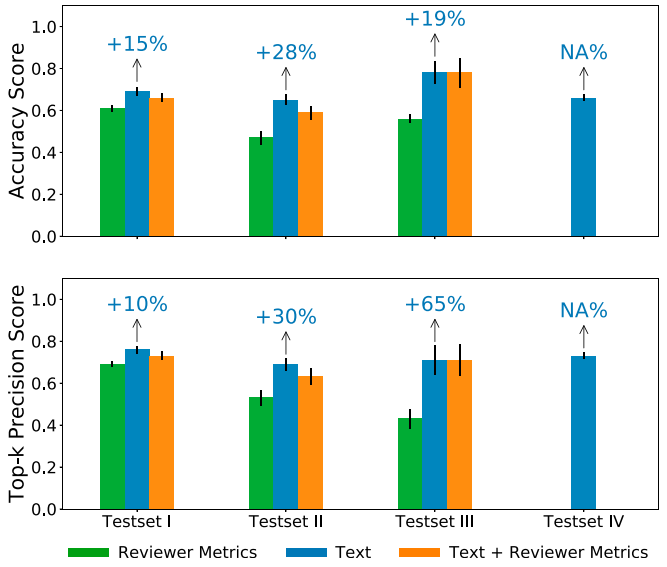

Fig. 3. Out-of-sample tests. From left to right, we present the accuracy and top-*k* precision results of four out-of-sample tests. Green, blue, and orange bars represent the percentage of correctly classified papers for the reviewer metrics-only model, narrative-only model, and combined narrative and reviewer metrics model. Numbers shown above narrative-only model bars indicate the percentage increase in accuracy/top-*k* relative to the reviewer metric-only model (reviewer metrics are unavailable for test set IV). Vertical bars are 95% CIs calculated by bootstrapping 100 random samples of 90% of the data with replacement. NA, not available.

narrative and reviewer metrics. The combined narrative and reviewer-metrics model achieved an average accuracy of 0.71 (SD = 0.031; KS, $P < 0.01$) and top-*k* precision of 0.76 (SD = 0.026; KS, $P < 0.01$). The combined model performed significantly better in terms of accuracy and top-*k* precision than either the narrative or reviewer metrics model alone, with an average increase of 27.8% (binomial test, $P < 0.01$; *SI Appendix*, Section S3.3). These tests suggest that the machine learning model based on the narrative of a study better predicts replicability than a reviewer metric-only model and that it may be advantageous in some circumstances to combine reviewer metric information into one model.

Out-of-Sample Tests We ran robustness tests of our machine learning model on five currently available out-of-sample datasets that report pass or fail outcomes (Table 1, rows 2 to 11). Fig. 3 summarizes the out-of-sample testing results for narrative (blue), reviewer metrics (green), and narrative-plus-reviewer metrics (orange) models. Test set I (described in Table 1, rows 2 to 8) consists of eight similarly conducted published psychology replication datasets ($n = 117$). The machine learning model generated an out-of-sample accuracy and top-*k* precision of 0.69 and 0.76, respectively.

Test set II (Table 1, row 9) consists of one set of 57 psychology replications done primarily by students as class projects, suggesting more noise in the "ground truth" data. Under these conditions of relatively high noise in the data, the machine learning model yielded an out-of-sample accuracy and top-*k* precision of 0.65 and 0.69, respectively.

Test sets III and IV are notable because they represent out-of-sample tests in the discipline of economics, a discipline that uses different jargon and studies different behavioral topics than does psychology—the discipline on which the model was trained (38). Test set III includes 18 economics experiments (Table 1, row 10). Test set IV includes 122 economics studies compiled by the Economics Replication Wiki (ERW) (Table 1, row 11). We tested these samples separately because the former consists of behavioral experiments, and the latter includes econometric modeling of archival data. The accuracy scores were 0.78 and 0.66, and the top-*k* precision scores were 0.71 and 0.73, respectively. To assess any training bias, we also conducted several cross-validation tests, which produced consistent results (*SI Appendix*, Section S3.7).

In another type of out-sample-test, we compared the machine learning model with prediction markets, the method with the highest prediction accuracy currently. Prediction markets provide pass/fail results and a level of confidence in each pass/fail prediction. The higher the confidence score, the more certain the market or survey participants were of their pass/fail predictions. Thus, we performed an out-of-sample test that focused on gauging the relative accuracy of machine learning and prediction markets from the point of view of correct classification and confidence. To construct our test, we collected the subsample of 100 papers from test sets I to IV that were included in prediction markets and ranked papers from least to most likely to replicate per the reported results of each prediction market and each associated survey. We then ranked the machine learning model's predictions of the same papers from least to most likely to replicate.

In comparing prediction markets, survey, and our machine learning model, we operated under the assumption that the most important papers to correctly identify for manual replication tests are the papers predicted to be least and most likely to replicate (DARPA SCORE objective; ref. 13). Fig. 4 shows that among the 10 most confident predictions of passing, the machine learning model correctly classified 90% of the studies, whereas the market or survey methods also correctly classified 90% of the studies. With respect to the 10 most confident failure predictions, the market or survey methods correctly classified 90% of the studies, and the machine learning model correctly classified

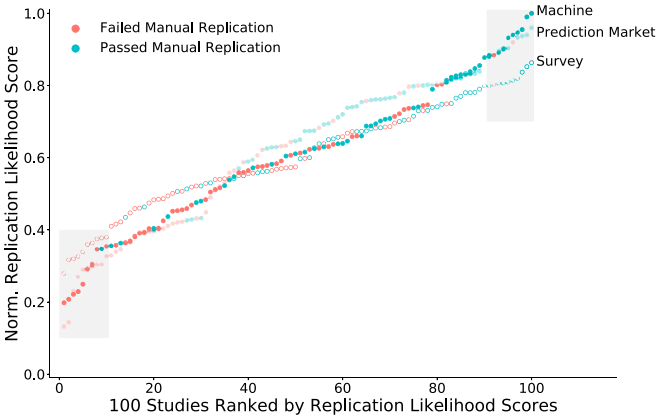

Fig. 4. Comparative performances of prediction market, survey, and machine learning models. One hundred psychology and economics studies from four different replication projects had likelihood scores assigned by prediction markets, surveys, and the machine learning model. The higher the likelihood score, the more certain the market or survey participants were of their pass predictions. In the figure, the likelihood scores are plotted from lowest to highest under the assumption that the chief papers to correctly identify for manual replication tests are the ones predicted to be least and most likely to replicate (13). With respect to the 10 most confident predictions of passing, the machine learning model predicts 90% of the studies correctly; the market or survey methods correctly classify 90% of the studies. Among the 10 most confident predictions of failing, the market or survey methods correctly classify 100% of the studies, and the machine learning model correctly classifies 90% of the studies. All three models have accuracy and top-*k* precision over 0.70. Norm., normalized.

Table 2. Tests of inherited bias in the machine learning model

Tests of inherited bias	Description of test and result
Prestigious authors and institutions are favored in the review process (39).	Adding citation information for first and senior authors and university affiliation prestige information to the RPP training data does not produce an accuracy significantly greater than the model without authors or affiliation information ($P = 0.36$).
Sex of authors has been associated with sex differences in the review process (40).	Adding the sex of authors to the RPP training data does not produce produce an accuracy statistically greater than the model without sex of author information ($P = 0.49$).
Journal-specific submission rates, reviewer norms, editors, or tastes can affect acceptance and rejection rates (41).	Adding journal-specific information to the RPP training data does not produce an accuracy significantly greater than the model without journal information ($P = 0.16$).
Disciplinary differences in replication failure rates (e.g., social psychology [72%] vs. cognitive psychology [47%]) or topics of analysis drive the machine learning model's predictions (42, 43).	Adding a disciplinary indicator to the RPP training data does not produce an accuracy statistically greater that the model without it ($P = 0.29$).
Scientific words, jargon, or native-English-speaking grammar is favored in the review process (44).	The frequency distributions of all words in replicating and nonreplicating papers showed that only one word differed: "experimenter" ($r = -0.26$, $P < 0.001$). Based on LIME (45), we also did not find any specific words highly associated with the reproducibility. Tests comparing the all-text model against the text model with all content words (nouns representing topics) removed demonstrated that the accuracy scores of the model containing content works are not significantly different from each other (KS test, $P < 0.05$). Tests comparing that all-text model against the text model with all function and stop words removed demonstrated that the accuracy scores of the model containing function and stop words are not significantly different from each other (KS test, $P = 0.91$).
Subjective probability language and hedging words/ persuasion phrases positively shape reviewers' interpretations of the findings (46).	The frequency distributions of subjective probability language hedging or persuasion phrases ("highly unlikely" or "little chance") produce no significant difference in replicating and nonreplicating papers ($t = 0.20$, $P = 0.84$).

We tested whether different lexical and nonlexical features of the training data resulted in bias in the model's predictions. The table summarizes the analyses and results. The text and *SI Appendix*, Section S4 describe the analyses in detail.

90%. Overall, the three models had accuracy and top-*k* precision over 0.70.

A machine learning model can inherit the predispositions found in the data used to create it (47). For example, human reviewers have been found to show partiality regarding author prestige and sex, institutional prestige, discipline, journal characteristics, word choice and grammar, English as a second language, and persuasive phrasing, leading to bias in reviewers' evaluations of a study (39, 42–44, 46, 48–51). We conducted tests to detect whether our machine learning framework was making classification decisions based on the preceding study characteristics. Our tests were done manually and with LIME, a popular and standardized algorithm for interpreting a machine's "reasoning" for its decisions (45). For example, LIME identifies if the machine used any particular word to make its classification decision. In addition, we did not detect statistical evidence of model bias regarding authorship prestige, sex of authors, discipline, journal, specific words, or subjective probabilities/persuasive language (*SI Appendix*, Section S4). Nascent forensic linguistic methods using n-grams—strings of neighboring words—have shown promise in identifying the authorship of anonymous texts, even though n-grams remain difficult to interpret (52). We found that the frequencies of two to five n-grams differed significantly in replicating and nonreplicating papers (Table 2 and *SI Appendix*, Section SI4).

Discussion

Machine learning appears to have the potential to aid the science of replication. Used alone, it offers accurate predictions at levels similar to prediction markets or surveys. In combination with prediction market or surveys predictions, accuracy scores are better than those from any other method on its own.

Though the findings should be taken as preliminary given the necessarily limited datasets with ground-truth data, our out-of-sample tests offer initial results that machine learning produces consistent predictions across studies having diverse methods, topics, disciplines, journals, replication procedures, and periods.

These research findings align with the DARPA SCORE program for developing new theories and approaches for replication. The SCORE program attempts to address the problem of replicability by assigning likelihood scores to published papers. The likelihood score is aimed at providing one piece of diagnostic information regarding the robustness of a study's findings and should be used in combination with a scientist's or user's own standards of review. The likelihood score can also be used by an author for self-reflection and diagnostics or by a researcher who, because of scarce resources, must prioritize the studies that should be replicated first. Our machine learning model can support SCORE by providing a scalable and rapid response approach to estimating a confidence score for all candidate studies of interest in a discipline. When scientists and users can set their own thresholds of confidence, a list of studies most and least likely to replicate can be determined generally or on a scientist-by-scientist basis. Manual replication-test allocation can then be prioritized for those studies with the lowest estimated confidence.

The machine learning model offers insights into how a scientific paper's narrative content hints at its replicability. Although reviewers generally estimate a paper's replicability from its statistics, we found that a model trained only on the narrative (text only) of a study achieves higher accuracy and top-*k* precision than reviewer metrics models, suggesting that the narrative-only method captures information that the reviewer metrics-only model does not. Moreover, no single reviewer metric predictor

was found to be a strong estimator of replicability in the literature. In our work, we tried to go beyond current literature on reviewer metrics and combine all metrics together to mimic the real-world use of a paper's reported statistics. When we combined reviewer metrics together, their predictive utility went up, but did not significantly surpass our text-only model. It may be that reviewers use reviewer metrics according to personal standards, which lowers their consistency. For example, *P* values are rules of thumb that can vary across reviewers. Nevertheless, if this conjecture is true, it may offer one explanation for why text provides a more consistent and higher level of predictability than reviewer metrics.

Our preliminary results indicate that the machine learning model capitalizes on differences in n-grams in replicating and nonreplicating papers. N-grams are strings of consecutive words that are colloquially referred to as writing "style" or phrasing. Forensic linguists have begun to research how to use n-grams to link anonymous or disguised authors with the documents they write. Consistent with their use of n-grams, we found that the frequency of n-grams differs in replicating and nonreplicating texts, with nonreplicating papers displaying a higher frequency of unusual n-grams and a lower frequency of common n-grams than do replicating papers. Nevertheless, because n-grams of three-, four-, and five-word strings are not yet directly interpretable by humans as style markers, n-grams require further testing to evaluate their usefulness for developing a theory of why text-based information reveals a paper's replicability (52).

In our investigative study, we marshalled nearly all available data sources of ground-truth manual replications to train and test our model. We found that tests of the predictive utility of machine learning are comparable to the best current methods, but also potentially have the distinct advantage of reducing burdens on researchers' scarce resources of time, money, and opportunity for conducting original research. Nevertheless, our samples are necessarily limited by being one-shot replication tests. Our research has been one instantiation of what we call deep-replicability research. Deep-replicability research aims to improve research and confidence in findings as data, methods, and scientific topics emerge by fruitfully combining human and machine intelligence in novel ways. Future research should continue to diversify and expand manual replication tests to all scientific disciplines with the aim of building a theory of replication and designing robust systems.

Materials and Methods

The manually replicated studies used to train and test our model came from seven projects in psychology and two projects in economics. The detailed information of all datasets used in our study can be found in Table 1. The data are available upon request.

ACKNOWLEDGMENTS. This work was supported by US Army Research Laboratory and US Army Research Office Grant W911NF-15-1-0577 QUANTA: Quantitative Network Based Models of Adaptive Teams; Air Force Office of Scientific Research Award FA9550-19-1-0354; and Northwestern Institute on Complex System.

1. O. S. Collaboration *et al.*, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
2. R. A. Klein *et al.*, Many Labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Practices Psychol. Sci.* **1**, 443–490 (2018).
3. F. Prinz, T. Schlange, K. Asadullah, Believe it or not: How much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10**, 712 (2011).
4. A. C. Chang, P. Li, "A preanalysis plan to replicate sixty economics research" in *American Economic Review: Papers & Proceedings 2017* (American Economic Association, 2017), pp. 60–64.
5. C. G. Begley, L. M. Ellis, Drug development: Raise standards for preclinical cancer research. *Nature* **483**, 531 (2012).
6. J. P. Ioannidis, Contradicted and initially stronger effects in highly cited clinical research. *Jama* **294**, 218–228 (2005).
7. C. F. Camerer *et al.*, Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
8. S. Fortunato *et al.*, Science of science. *Science* **359**, eaao0185 (2018).
9. D. B. Resnik, Scientific research and the public trust. *Sci. Eng. Ethics* **17**, 399–409 (2011).
10. A. Dreber *et al.*, Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15343–15347 (2015).
11. M. Baker, 1,500 scientists lift the lid on reproducibility. *Nat. News* **533**, 452–454(2016).
12. L. P. Freedman, I. M. Cockburn, T. S. Simcoe, The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, e1002165 (2015).
13. A. Russell, "Systematizing confidence in open research and evidence (SCORE)" (Tech. Rep., Defense Advanced Research Projects Agency, Arlington, VA, 2019).
14. U. Simonsohn, L. D. Nelson, J. P. Simmons, P-curve: A key to the file-drawer. *J. Exp. Psychol. Gen.* **143**, 534–547 (2014).
15. J. J. Van Bavel, P. Mende-Siedlecki, W. J. Brady, D. A. Reinero, Contextual sensitivity in scientific reproducibility. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6454–6459 (2016).
16. C. F. Camerer *et al.*, Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).
17. E. Forsell *et al.*, Predicting replication outcomes in the Many Labs 2 study. *J. Econ. Psychol.* **75**, 102117 (2019).
18. A. Altmejd *et al.*, Predicting the replicability of social science lab experiments. *PloS one* **14**, e0225826 (2019).
19. T. K. Landauer, S. T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997).
20. D. S. McNamara, Computational methods to extract meaning from text and advance theories of human cognition. *Top. Cognit. Sci.* **3**, 3–17 (2011).
21. K. Wang *et al.*, A review of Microsoft academic services for science of science studies. *Frontiers in Big Data* **2**, 45 (2019).
22. D. J. Simons, A. O. Holcombe, Registered replication reports. *APS Observer* **27**, (2014).
23. A. Klein Richard *et al.*, Investigating variation in replicability: A "Many Labs" replication project. *Soc. Psychol.* **45**, 142–152 (2014).
24. C. R. Ebersole *et al.*, Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).
25. B. A. Nosek, D. Lakens, Replications of important results in social psychology: Special issue of social psychology. *Soc. Psychol.* **44**, 59–60 (2013).
26. A. Aarts, E. LeBel, Curate science: A platform to gauge the replicability of psychological science. https://curatescience.org. Accessed 1 June 2017.
27. H. Pashler, B. Spellman, S. Kang, A. Holcombe, Psychfiledrawer: Archive of replication attempts in experimental psychology (2019). http://psychfiledrawer.org/view_article_list.php. Accessed 1 December 2017.
28. J. H. Höffler, Replicationwiki: Improving transparency in social sciences research. *D-Lib Mag.*, 10.1045/march2017-hoeffler (2017).
29. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality" in *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger, Eds. (Curran Associates, Red Hook, NY, 2013), vol. 2, pp. 3111–3119.
30. O. Levy, Y. Goldberg, "Neural word embedding as implicit matrix factorization" in *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger, Eds. (MIT Press, Cambridge, MA, 2014), vol. 2, pp. 2177–2185.
31. J. Pennington, R. Socher, C. D. Manning, "GloVe: Global vectors for word representation" in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, W. Daelemans, Eds. (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 1532–1543.
32. D. H. Wolpert, Stacked generalization. *Neural Networks* **5**, 241–259 (1992).
33. G. Forman, I. Cohen, "Learning from little: Comparison of classifiers given little training" in *Knowledge Discovery in Databases: PKDD 2004*, J. F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi, Eds. (Lecture Notes in Computer Science, Springer, Berlin, Germany, 2004), vol. 3202, pp. 161–172.
34. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, Berlin, 2006).
35. Y. Yang, R. N. Lichtenwalter, N. V. Chawla, Evaluating link prediction methods. *Knowl. Inf. Syst.* **45**, 751–782 (2015).
36. T. W. Anderson, D. A. Darling, Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Stat.* **23**, 193–212 (1952).
37. D. J. Steinskog, D. B. Tjøstheim, N. G. Kvamstø, A cautionary note on the use of the Kolmogorov–Smirnov test for normality. *Mon. Weather Rev.* **135**, 1151–1157 (2007).
38. R. Hertwig, A. Ortmann, Experimental practices in economics: A methodological challenge for psychologists? *Behav. Brain Sci.* **24**, 383–403 (2001).
39. A. Tomkins, M. Zhang, W. D. Heavlin, Reviewer bias in single-versus double-blind peer review. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12708–12713 (2017).
40. T. Tregenza, Gender bias in the refereeing process? *Trends Ecol. Evol.* **17**, 349–350 (2002).
41. J. R. Gilbert, E. S. Williams, G. D. Lundberg, Is there gender bias in JAMA's peer review process? *Jama* **272**, 139–142 (1994).
42. Y. Inbar, Association between contextual dependence and replicability in psychology may be spurious. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4933–E4934 (2016).

43. J. Gao, B. Ding, W. Fan, J. Han, S. Y. Philip, Classifying data streams with skewed class distributions and concept drifts. *IEEE Internet Computing* **12**, 37–49 (2008).
44. P. Martín, J. Rey-Rocha, S. Burgess, A. I. Moreno, Publishing research in English-language journals: Attitudes, strategies and difficulties of multilingual scholars of medicine. *J. Engl. Acad. Purp.* **16**, 57–67 (2014).
45. M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2016), pp. 1135–1144.
46. P. Iyer, M. Narasimha, *'Almost always' and 'sometime definitely' are not enough: Probabilistic quanifiers and probabilistic model-checking* (Tech. Rep. 996-16, North Carolina State University at Raleigh, Raleigh, NC, 1996).
47. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
48. C. Tan C., V. Niculae, C. Danescu-Niculescu-Mizil, L. Lee, "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions" in *WWW'16: Proceedings of the 25th International Conference on World Wide Web* (International World Wide Web Conferences Steering Committee, Geneva, Switzerland, 2016), pp. 613–624.
49. D. M. Romero, R. I. Swaab, B. Uzzi, A. D. Galinsky, Mimicry is presidential: Linguistic style matching in presidential debates and improved polling numbers. *Pers. Soc. Psychol. Bull.* **41**, 1311–1319 (2015).
50. L. L. Hargens, Variation in journal peer review systems: Possible causes and consequences. *Jama* **263**, 1348–1352 (1990).
51. C. J. Lee, C. R. Sugimoto, G. Zhang, B. Cronin, Bias in peer review. *J. Am. Soc. Inf. Sci. Technol.* **64**, 2–17 (2013).
52. D. Wright, Using word n-grams to identify authors and idiolects. *Int. J. Corpus Linguist.* **22**, 212–241 (2017).