*HHAI 2023: Augmenting Human Intellect*
*P. Lukowicz et al. (Eds.)*

*doi:10.3233/FAIA230093*

# A Prototype Hybrid Prediction Market for Estimating Replicability of Published Work

Tatiana CHAKRAVORTI [a], Robert FRALEIGH [a], Timothy FRITTON [a], Michael MCLAUGHLIN [a], Vaibhav SINGH [a], Christopher GRIFFIN [a], Anthony KWASNICA [a], David PENNOCK [b], C. Lee GILES [a], Sarah RAJTMAJER [a,1]

[a] *The Pennsylvania State University*
[b] *Rutgers University*

**Abstract.** We present a prototype hybrid prediction market and demonstrate the avenue it represents for meaningful human-AI collaboration. We build on prior work proposing artificial prediction markets as a novel machine learning algorithm. In an artificial prediction market, trained AI agents (bot traders) buy and sell outcomes of future events. Classification decisions can be framed as outcomes of future events, and accordingly, the price of an asset corresponding to a given classification outcome can be taken as a proxy for the systems confidence in that decision. By embedding human participants in these markets alongside bot traders, we can bring together insights from both. In this paper, we detail pilot studies with prototype hybrid markets for the prediction of replication study outcomes. We highlight challenges and opportunities, share insights from semi-structured interviews with hybrid market participants, and outline a vision for ongoing and future work.

**Keywords.** Hybrid Prediction Market, Human-AI Collaboration, Reproducibility.

## 1. Introduction

A nascent literature is exploring artificial prediction markets – numerically simulated markets, populated by artificial agents (bot traders) for supervised learning of probability estimators [3]. Early work has demonstrated the plausibility of using a trained market as a supervised learning algorithm, achieving comparable performance to standard approaches on simple classification tasks [3,4,16,20]. We suggest the most promising opportunity afforded by artificial prediction markets is eventual human-AI collaboration – a market framework that supports human traders participating alongside agents to evaluate outcomes. In an initial study [7], we have outlined the theoretical foundation for such a *hybrid prediction market* and simulated simple human-like behaviors in this setting.

The hybrid markets we describe here aim to predict the outcomes of replication studies in the social and behavioral sciences. The study of reproducibility, replicability, and robustness of published scientific findings has gained widespread attention in the social sciences and beyond. A number of large-scale replication projects ((RPP = Replication

[1]Corresponding Author: smr48@psu.edu

Projection Psychology, SSRP = Social Science Replication Project, ML = Many Labs, ML2 = Many Labs 2)) [10,21,5,6,17,18,8] have reported successful replication rates anywhere between 36% and 78% and have sparked high-profile debate about the reliability of published findings [2,13,11,24]. The task of forecasting replication outcomes appears to be an ideal candidate for human-AI collaboration. Both human and machine-centered efforts have shown promise but neither has achieved good performance, e.g., [9,27]. Indeed, replication prediction appears to be a problem for which the scale and scope of machine-driven approaches is necessary but capturing the intangible wisdom of experts in the field remains elusive.

Our prior work [22] has developed and deployed synthetic prediction markets for the task of replication prediction. We further this work here to explore a hybrid market scenario. Our work in progress is scaffolded by three research questions, the answers to which will inform larger-scale development of human-AI hybrid prediction markets as a novel avenue for creative peer review.

**RQ1**: How does human participation in a synthetic prediction market impact market performance vs. the purely synthetic setting?

**RQ2**: In the context of replication prediction, what features matter most to human participants? How do participants formulate trading strategies?

**RQ3**: What outstanding challenges need to be addressed prior to large-scale deployment of hybrid prediction markets for replication prediction?

Overarchingly, these questions serve the broader aim of understanding how hybrid human-AI technologies can help us evaluate reproducibility, replicability, and robustness of published scientific findings. Following, we discuss insights from beta testing a hybrid market for replication prediction. We also discuss participants' perspectives based on follow-up interviews. We close with a vision for further work in this area.

## 2. Related Work

Given the resources required to run high-powered replication studies, researchers have sought other approaches to assess confidence in published claims and have looked to the creative assembly of expert judgment as one opportunity. Initial evidence has supported the promise of prediction markets in this context. Simple, binary option prediction markets have outperformed survey-based approaches in predicting outcomes for a number of high-profile replication projects [9,5,6,12,14,15]. In these markets, assets corresponding to future events can be bought and sold thereby manipulating underlying asset prices. These asset prices can be interpreted as probabilities [19,25] thereby providing a mechanism for event forecasting. This paper will present a human-AI collaboration in the prediction market for research reproducibility which is completely novel and has not been explored previously.

## 3. Data

Algorithmic agents were trained on outcomes of 400 replication studies and expert evaluations of published findings in the social and behavioral sciences, spearheaded by the Center for Open Science and the University of Melbourne for DARPA's Systematizing Confidence in Open Research and Evidence (SCORE)[2] program. Twelve additional

[2]See https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence.

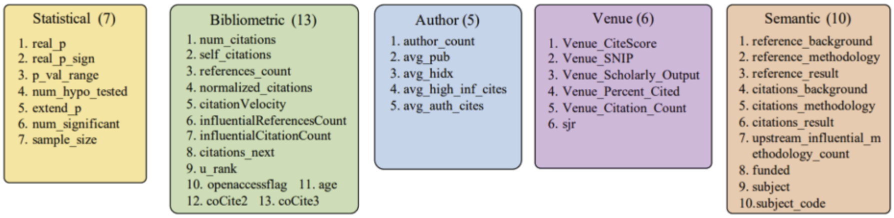


**Figure 1. Details of all five different features** (see [26]).

replication study outcomes served as test data for the hybrid market events.[3] Domains and journals from which findings were selected, as well as procedures for replications and expert evaluations are detailed in [1].

Full text of train and test papers were passed through a feature extraction pipeline to obtain semantic, bibliometric, and statistical information. Specifically, 41 features, e.g., reported p-values, author names, author count, venue, acknowledgment of funding, were extracted for each research claim in question. See [26] for further detail. In this paper, all these three types of features have been mentioned in detail which is also clearly mentioned in Figure 1.

## 4. Methodology

### *4.1. Hybrid Market Experimental Design*

The base model of this work is a simple artificial prediction market populated by algorithmic agents (bot traders) whose decisions to buy contracts are based on extracted features from the full text of published work and associated metadata and who are trained on ground-truth replication outcomes using a genetic algorithmic approach. Further detail on the mathematical formulation of the artificial market, bot training procedures, and feature extraction from scholarly manuscripts are detailed in prior work [20,26,22]. A schematic representation of the market and training processes is given in Figure 2.

In ongoing research, we have built a beta-tested a platform to enable bot interactions with real human participants, i.e., a hybrid market scenario, for replication prediction. The platform includes a web server that hosts a pre-trained artificial market and several API endpoints that: 1) enable artificial agents to buy assets; 2) enable human participants to buy and sell assets; 3) manage transaction bookkeeping and experiment statistics. The platform also includes an interactive web application for human participants to buy and sell assets intuitively (see Figure 3).

While the theoretical foundation of the artificial market is framed continuously, the deployed artificial and hybrid markets are discrete, iterating every one second. Market transactions were managed using a queuing system, first evaluating all participating agents and then any human transactions using a first-in-first-out rule. Both artificial agents and human participants were limited to single-share transactions. Upon each market transaction, the market price for all assets was updated using a logarithmic market

[3]Coordinated release of all data from the SCORE program is planned in late 2023. The subset of data used in our analyses will be linked at https://github.com/Tatianachakravorti as soon as available.

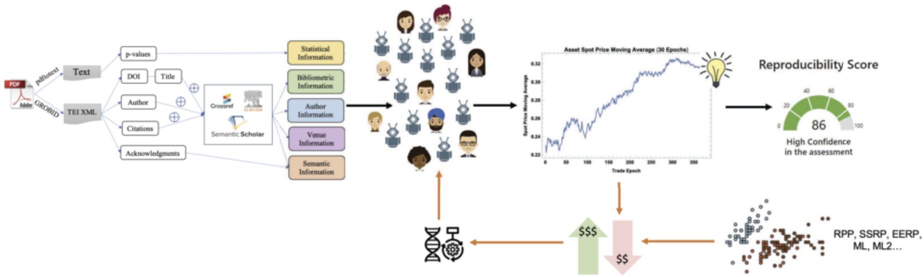


**Figure 2. Schematic representation of market for replication prediction.** Step 1: Features are extracted from full-text of a paper of interest (see [26]). Step 2: Extracted features are passed to algorithmic traders. Step 3: Algorithmic traders buy and sell contracts representing 'will replicate' or 'will not replicate' outcomes of a replication study associated with the primary claim of the paper of interest. *Note: in the hybrid market scenario, human traders participate alongside bots.* Trading manipulates the underlying asset prices via logarithmic market scoring rule (see [20]). (*training phase, orange arrows*) Step 3. At market close, the outcome of the replication study is revealed. Algorithmic traders profit or lose money based on the total value of the assets they hold. Traders who profit are allowed to reproduce, mutate, and remain in the market via genetic algorithms. (*test phase, black arrows*) Step 3. The price of a 'will replicate' asset at the time of market close is given as a proxy for the market's prediction.

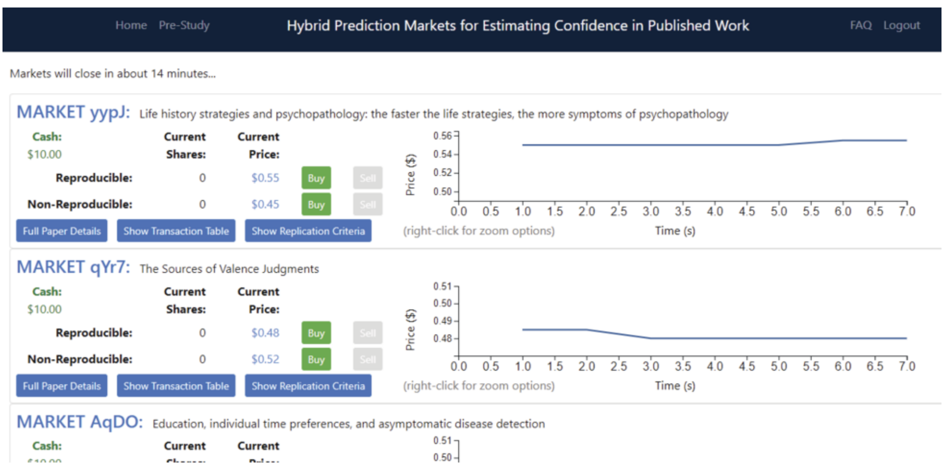


**Figure 3. Interactive web application.** Human participants are provided with information about the replication study and full text of the associated paper. We provide them with initial cash to invest. They may buy and sell contracts representing 'will replicate' and 'will not replicate' outcomes of the study via the app.

scoring rule (LMSR) [20]. During each market iteration, a stochastic sampling of available artificial agents were selected to participate; the sampling rate was reduced to 5% of the artificial agent population to align market price convergence with a 2-hour timeline (7200 market iterations) for the hybrid market scenario.

Hybrid markets were beta-tested in three separate events. Instiutional Review Board (IRB) approval was obtained prior to these events. Each of the three events involved four markets; that is, in each event, participants were given the opportunity to evaluate and buy/sell outcomes of four distinct replication studies. We recruited participants for each, respectively, with three research backgrounds: graduate students at Penn State's

Smeal School of Business (Event 1: April 2022, 9 participants), graduate students at Penn State's College of Information Science and Technology (Event 2: June 2022, 10 participants), and graduate students from the Department of Psychology at Penn State (Event 3: July 2022, 14 participants). Event 1 was conducted in-person at the Laboratory for Economics, Management & Auctions on campus and markets were open for one hour. Events 2 and 3 were held virtually and markets were open for two hours. Participants in all events were given $20 for their time and an additional $25 to invest in the markets. At the conclusion of each event, one market was randomly selected for payment. The selected market along with the replication outcome for that market was revealed to participants. Assuming a minimal activity requirement was satisfied, each participant was paid the value of their asset holdings in the selected market and any remaining (uninvested) cash. The average participant earning was $42.62. Table 1 provides a summary of hybrid market outcomes, alongside corresponding artificial market outcomes for each test paper. For each market, the final price (in dollars) is reported for an asset that pays $1 if the finding was successfully replicated and $0 if not, i.e., the 'will replicate' asset.[4] Thus, we classify a market prediction as 'correct' if the price is greater (less) than .5 and the paper was (not) reproducible. Absolute error (AE) is the absolute difference between the final price and the value of the asset (0 or 1). Markets with consistently lower AE are considered better predictors.

### *4.2. Participant Surveys and Interviews*

Hybrid market participants were asked to complete pre- and post-experimental surveys. The pre-market survey assessed participants' research background and familiarity with scholarly work on reproducibility and replicability. In addition, the pre-market survey asked participants to provide feedback on each paper they were about to evaluate during the hybrid market event. The post-market survey asked participants to describe their trading strategy, specific features of the studies, e.g., sample size, author reputation, that guided their predictions. We asked whether they were surprised by the market outcome, and whether they had changed their mind from their original assessments.[5]

We also conducted 30-minute semi-structured one-on-one interviews via Zoom teleconferencing with 8 hybrid market participants who responded to our request for additional feedback. Interviewees were active researchers with a PhD or currently enrolled in a PhD program.[6] Our interview protocol included questions about their experience in the market and some broader questions related to reproducibility. Participants received $20.

## 5. Results

### *5.1. RQ1: Hybrid market performance*

Table 1 provides a summary of hybrid market outcomes, alongside corresponding artificial market outcomes for each test paper. For each market, the final price (in dollars) is

[4] The market price of the 'will not replicate' asset is always 1 minus the price of this asset by construction.

[5] Survey instruments are shared at https://github.com/Tatianachakravorti.

[6] Six of the 8 interviewees were participants in hybrid market experiments run in October 2022. They were not part of the three market events reported in Table 1. October events followed the same 2-hour format and used the same platform as the market events in April, June and July 2022.

**Table 1. Experimental data summary.** R/NR: replicated/not replicated; C/NC: correct/not correct; price: final price; pred: prediction; AE: absolute error; and, – denotes no agent participation.

| Market | Outcome | Hybrid price | Hybrid pred | Hybrid AE | Artif price | Artif pred | Artif AE |
|---|---|---|---|---|---|---|---|
| Event 1 Market 1 (E1M1) | R | 0.66 | C | 0.34 | 0.41 | NC | 0.59 |
| Event 1 Market 2 (E1M2) | R | 0.36 | NC | 0.64 | 0.5 | – | 0.5 |
| Event 1 Market 3 (E1M3) | R | 0.64 | C | 0.36 | 0.52 | C | 0.48 |
| Event 1 Market 4 (E1M4) | NR | 0.72 | NC | 0.72 | 0.5 | – | 0.5 |
| Event 2 Market 1 (E2M1) | R | 0.38 | NC | 0.62 | 0.41 | NC | 0.59 |
| Event 2 Market 2 (E2M2) | R | 0.58 | C | 0.42 | 0.5 | – | 0.5 |
| Event 2 Market 3 (E2M3) | R | 0.8 | C | 0.2 | 0.52 | C | 0.48 |
| Event 2 Market 4 (E2M4) | NR | 0.47 | C | 0.47 | 0.5 | – | 0.5 |
| Event 3 Market 1 (E3M1) | R | 0.61 | C | 0.39 | 0.5 | – | 0.5 |
| Event 3 Market 2 (E3M2) | R | 0.47 | NC | 0.53 | 0.46 | NC | 0.54 |
| Event 3 Market 3 (E3M3) | NR | 0.76 | NC | 0.76 | 0.86 | NC | 0.86 |
| Event 3 Market 4 (E3M4) | R | 0.49 | NC | 0.51 | 0.42 | NC | 0.58 |

reported for an asset that pays $1 if the finding was successfully replicated and $0 if not, i.e., the 'will replicate' asset. [7] Thus, we classify a market prediction as 'correct' if the price is greater (less) than .5 and the paper was (not) reproducible. Absolute error (AE) is the absolute difference between final price and value of the asset (0 or 1). Markets with consistently lower AE are considered better predictors.

Hybrid market predictions were globally more accurate than predictions of the artificial markets (mean AE .497 vs .552). In 9 of 12 markets, AE was lower in the hybrid setting. A Wilcoxon signed ranks test fails to reject the null hypothesis that the distributions of errors between the hybrid and artificial markets are the same ($z = -1.373$, $p - value = 1.83$), likely due to the low power of the small sample. In no instance is the hybrid market price incorrect when the synthetic price was correct; in one instance the hybrid market flipped an incorrect prediction of the synthetic market to correct.

Artificial markets are vulnerable to lack of participation; agents will not participate if they have not seen a sufficiently similar training point (paper). In practice, this may leave some test points unevaluated. We have observed this in prior work [22] and that is the case here in five of twelve markets.[8] In the hybrid setting, human participation can support a prediction. And notably, we observed that *the presence of human traders often induced greater participation amongst algorithmic agents.* In four of the five markets inactive in the artificial setting, the presence of human traders induced agent trades; only in hybrid market E2M2 were agents completely inactive.

These small sample results indicate that small numbers of informed/expert human traders do not have an obviously deleterious effect on market performance and might even improve accuracy. Likewise, the trading activity of human traders has the potential benefit of triggering trading amongst algorithmic agents. In future, we have designed a plan where we can implement more human participants to test the model and its performance. We will target different platforms like LinkedIn and tweeter to recruit experts and also we have a list of universities that we will target to send out the participation emails.

[7] The market price of the 'will not replicate' asset is always 1 minus the price of this asset by construction.
[8] In the absence of trading, the market price for both assets is .5.

### 5.2. RQ2: Human participants' evaluative criteria and trading strategies

Findings in support of RQ2 primarily derive from participant interviews. We analyzed interview transcripts using an inductive approach guided by specific evaluation objectives. To analyze and code these transcripts we have used the Taguette software [23]. Following, we provide key themes and exemplar quotes.

#### 5.2.1. Study motivation, design, and reported outcomes

All eight interviewed participants mentioned methodology, sample size, p-values and soundness of research questions as prominent evaluative criteria.

> *I focused, specifically on the size of the sample, the diversity of the sample, and the complexity of the question, in order to predict how reliable it would be... so if the sample was small if the population wasn't diverse and the question was complex... This seems to have a lower likelihood of reproducing.* [Participant 2]

#### 5.2.2. Journal and author reputation

Six participants noted an impact of journal reputation on their prediction. However, most participants reported that they did not consider the papers' authors; one participant noted an exception for one paper.

> *I guess the journal itself... I know some of the journals have tended to have higher-quality kind of articles. So Im sure that that affected my thought process to some extent but really the main thing was, you know, sample size, methodology, and theoretical orientation.* [Participant 1].

#### 5.2.3. Trading strategy

Five participants out of eight reported having a probability in mind for the replicability of each finding prior to the markets opening, and made initial investments based on this probability. However, several mentioned that they later changed their mind when the market start trending in opposite direction to their initial prediction.

> *I have a probability in mind... so if I think that it should be seventy-five, and the price is sixty percent then I'm willing to buy because I think I'm making fifteen cents. But if it's trading at forty and I think that the probability of replication is twenty, I'm happy to buy because I think there I'm making twenty cents uh on the dollar or on the share or whatever.* [Participant 5].

#### 5.2.4. Impact of agent participation on decisions

All participants reported that the presence of agents in the market had no impact on their behavior because they had no information about the agents' training or behavior.

### 5.3. RQ3: Perspectives on human-AI technologies for replication prediction

#### 5.3.1. Concerns about reproducibility and replicability

All participants interviewed reported concern about reproducibility and replicability. Although we expect this is related to self-selection bias for our study. Participants reported variability in awareness around these issues in their field.

> *I guess maybe thirty percent of people in my field have pretty strong concerns about it.* [Participant 1]

Participants acknowledged lack of incentives for practices heralded by the open science movement (e.g., preregistration) and lack of venues for publishing replication studies.

> *I think the great challenge of replicability is that we still don't run nearly enough replications... We need to value replicability, and we need to create a niche for academics who work in replicability.* [Participant 5]

### 5.3.2. *Hybrid human-AI technologies to support reproducibility and replicability*

We queried interview participants to understand their perspectives on opportunities for technologies, and in particular, human-AI technologies to support and enhance reproducibility and replicability. Participants were hopeful for technological interventions although generally preferred hybrid solutions. Seven of the eight participants expressed incomplete trust in AI-driven solutions alone.

> *I would expect the hybrid model would be more reliable... just adding more predictors to a model tends to give you the better performance... and I'll be very keenly interested to see what the results of all those markets.* [Participant 5]

> *I believe in the markets... Probably my expectation would be the hybrid market performs most accurately.* [Participant 7].

### 5.3.3. *Hybrid prediction market experience*

Finally, we collected participants' inputs on their experience during the hybrid markets and solicited suggestions for improvements to the platform and/or methodology moving forward. Most participants were enthusiastic and reported the experience as fun. Several highlighted details of the UI that could be improved. All participants felt the market duration was too long. We had selected a 2-hour window to afford best flexibility to participants logging in from around the world. We will revisit this in next steps.

> *To me, it seems that the market lasted quite a while... It was really just slowly trending in one direction versus another direction... It was really a lot of fun. Actually, I quite enjoyed it.* [Participant 1]

## 6. Conclusion and Future work

We have described pilot studies with a prototype hybrid prediction market for replication prediction. This work in progress offers proof of the viability of collaborative human-AI technology for the evaluation of published scientific claims. Although we pilot this approach in the context of replication prediction, we suggest that the hybrid market offers a new avenue for hybrid human-AI applicable to a broad set of tasks for which neither human- or machine-driven approaches alone are sufficient. Post-market interviews with participants highlight opportunities and challenges for this work.

There are many things that can be explored in future work like there are many different other game theories where agents learn from each other and perform. In this study, agents are not communicating between themselves but this can be implemented in the future to make the study more interesting. At the same time how human participation can affect the market throughout can be explored, specifically between opening and closing of the market.

## References


[1] N. Alipourfard, B. Arendt, D. J. Benjamin, N. Benkler, M. M. Bishop, M. Burstein, M. Bush, J. Caverlee, Y. Chen, C. Clark, et al. Systematizing confidence in open research and evidence (score). Technical report, Center for Open Science, 2021.

[2] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452, 2016.

[3] A. Barbu and N. Lay. An introduction to artificial prediction markets for classification. *The Journal of Machine Learning Research*, 13(1):2177–2204, 2012.

[4] A. Barbu and N. Lay. Artificial prediction markets for lymph node detection. In *2013 E-Health and Bioengineering Conference (EHB)*, pages 1–7. IEEE, 2013.

[5] C. F. Camerer, A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, 2016.

[6] C. F. Camerer, A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018.

[7] T. Chakravorti, V. Singh, S. Rajtmajer, M. McLaughlin, R. Fraleigh, C. Griffin, A. Kwasnica, D. Pennock, and C. L. Giles. Artificial prediction markets present a novel opportunity for human-ai collaboration, 2022.

[8] F. Cova, B. Strickland, A. Abatista, A. Allard, J. Andow, M. Attie, J. Beebe, R. Berniūnas, J. Boudesseul, M. Colombo, et al. Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 12(1):9–44, 2021.

[9] A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347, 2015.

[10] T. M. Errington, E. Iorns, W. Gunn, F. E. Tan, J. Lomax, and B. A. Nosek. Science forum: An open investigation of the reproducibility of cancer biology research. *Elife*, 3:e04333, 2014.

[11] D. Fanelli. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11):2628–2631, 2018.

[12] E. Forsell, D. Viganola, T. Pfeiffer, J. Almenberg, B. Wilson, Y. Chen, B. A. Nosek, M. Johannesson, and A. Dreber. Predicting replication outcomes in the many labs 2 study. *Journal of Economic Psychology*, 75:102117, 2019.

[13] D. T. Gilbert, G. King, S. Pettigrew, and T. D. Wilson. Comment on estimating the reproducibility of psychological science. *Science*, 351(6277):1037–1037, 2016.

[14] M. Gordon, D. Viganola, M. Bishop, Y. Chen, A. Dreber, B. Goldfedder, F. Holzmeister, M. Johannesson, Y. Liu, C. Twardy, et al. Are replication rates the same across academic fields? community forecasts from the darpa score programme. *Royal Society open science*, 2020.

[15] M. Gordon, D. Viganola, A. Dreber, M. Johannesson, and T. Pfeiffer. Predicting replicabilityanalysis of survey and prediction market data from large-scale forecasting projects. *Plos one*, 16(4):e0248780, 2021.

[16] F. Jahedpari, J. Padget, M. De Vos, and B. Hirsch. Artificial prediction markets as a tool for syndromic surveillance. *Crowd Intelligence: Foundations, Methods and Practices*, 2014.

[17] R. A. Klein, K. A. Ratliff, M. Vianello, R. B. Adams Jr, Š. Bahník, M. J. Bernstein, K. Bocian, M. J. Brandt, B. Brooks, C. C. Brumbaugh, et al. Investigating variation in replicability. *Social psychology*, 2014.

[18] R. A. Klein, M. Vianello, F. Hasselman, B. G. Adams, R. B. Adams Jr, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Š. Bahník, et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

[19] C. F. Manski. Interpreting the predictions of prediction markets. *economics letters*, 91(3):425–429, 2006.

[20] N. Nakshatri, A. Menon, C. L. Giles, S. Rajtmajer, and C. Griffin. Design and analysis of a synthetic prediction market using dynamic convex sets. *arXiv preprint arXiv:2101.01787*, 2021.

[21] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

[22] S. Rajtmajer, C. Griffin, J. Wu, R. Fraleigh, L. Balaji, A. Squicciarini, A. Kwasnica, D. Pennock, M. McLaughlin, T. Fritton, et al. A synthetic prediction market for estimating confidence in published work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 13218–13220, 2022.

[23] R. Rampin and V. Rampin. Taguette: open-source qualitative data analysis. *Journal of Open Source Software*, 6(68):3522, 2021.

[24] A. D. Redish, E. Kummerfeld, R. L. Morris, and A. C. Love. Opinion: Reproducibility failures are essential to scientific inquiry. *Proceedings of the National Academy of Sciences*, 115(20):5042–5046, 2018.

[25] J. Wolfers and E. Zitzewitz. Interpreting prediction market prices as probabilities. Technical report, National Bureau of Economic Research, 2006.

[26] J. Wu, R. Nivargi, S. S. T. Lanka, A. M. Menon, S. A. Modukuri, N. Nakshatri, X. Wei, Z. Wang, J. Caverlee, S. M. Rajtmajer, et al. Predicting the reproducibility of social and behavioral science papers using supervised learning models. *arXiv preprint arXiv:2104.04580*, 2021.

[27] Y. Yang, W. Youyou, and B. Uzzi. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(20):10762–10768, 2020.