Accepted manuscript

Inventory of Supporting Information

1. Supplementary Information:

A. PDF Files

Item	Present?	Filename	A brief, numerical description of file contents.
Supplementary Information	Yes	Grossmann_SI.pdf	Supplementary Methods, Supplementary Figures 1-15, Supplementary Tables 1-9, Supplementary Appendices 1-5
Reporting Summary	Yes	Grossmann_RS.pdf	
Peer Review Information	Yes	PRFile_Grossmann.pdf	

Title

Insights into accuracy of social scientists' forecasts of societal change

Author list

Igor Grossmann[1*], Amanda Rotella[2,1], Cendri A. Hutcherson[3], Konstantyn Sharpinskyi[1], Michael E. W. Varnum[4], Sebastian Achter[5], Mandeep K. Dhami[6], Xinqi Evie Guo[7], Mane Kara-Yakoubian[8], David R. Mandel[9,10], Louis Raes[11], Louis Tay[12], Aymeric Vie[13,14], Lisa Wagner[15], Matus Adamkovic[16,17], Arash Arami[18], Patrícia Arriaga[19], Kasun Bandara[20], Gabriel Baník[16], František Bartoš[21], Ernest Baskin[22], Christoph Bergmeir[23], Michał Białek[24], Caroline K. Børsting[25], Dillon T. Browne[1], Eugene M. Caruso[26], Rong Chen[27], Bin-Tzong Chie[28], William J. Chopik[29], Robert N. Collins[9], Chin W. Cong[30], Lucian G. Conway[31], Matthew Davis[32], Martin V. Day[33], Nathan A. Dhaliwal[34], Justin D. Durham[35], Martyna Dziekan[36], Christian T. Elbaek[25], Eric Shuman[37], Marharyta Fabrykant[38,39], Mustafa Firat[40], Geoffrey T. Fong[1,41], Jeremy A. Frimer[42], Jonathan M. Gallegos[43], Simon B. Goldberg[44], Anton Gollwitzer[45,46], Julia Goyal[47], Lorenz Graf-Vlachy[48,49], Scott D. Gronlund[35], Sebastian Hafenbrädl[50], Andree Hartanto[51], Matthew J. Hirshberg[52], Matthew J. Hornsey[53], Piers D. L. Howe[54], Anoosha Izadi[55], Bastian Jaeger[56], Pavol Kačmár[57], Yeun Joon Kim[58], Ruslan Krenzler[59,60], Daniel G. Lannin[61], Hung-Wen Lin[62], Nigel Mantou Lou[63,64], Verity Y. Q. Lua[51], Aaron W. Lukaszewski[65,66], Albert L. Ly[67], Christopher R. Madan[68], Maximilian Maier[69], Nadyanna M. Majeed[70], David S. March[71], Abigail A. Marsh[72], Michal Misiak[24,73], Kristian Ove R. Myrseth[74], Jaime M. Napan[67], Jonathan Nicholas[75], Konstantinos Nikolopoulos[76], Jiaqing O[77], Tobias Otterbring[78,79], Mariola Paruzel-Czachura[80,81], Shiva Pauer[21], John Protzko[82], Quentin Raffaelli[83], Ivan Ropovik[84,85], Robert M. Ross[86], Yefim Roth[87], Espen Røysamb[88], Landon Schnabel[89], Astrid Schütz[90], Matthias Seifert[91], A. T. Sevincer[92], Garrick T. Sherman[93], Otto Simonsson[94,95], Ming-Chien Sung[96], Chung-Ching Tai[96], Thomas Talhelm[97], Bethany A. Teachman[98], Philip E. Tetlock[99,100], Dimitrios Thomakos[101], Dwight C. K. Tse[102], Oliver J. Twardus[103], Joshua M. Tybur[56], Lyle Ungar[93], Daan Vandermeulen[104], Leighton Vaughan Williams[105], Hrag A. Vosgerichian[106], Qi Wang[107], Ke Wang[108], Mark E. Whiting[109,110], Conny E. Wollbrant[111], Tao Yang[112], Kumar Yogeeswaran[113], Sangsuk Yoon[114], Ventura r. Alves[115], Jessica R. Andrews-Hanna[83,116], Paul A. Bloom[75], Anthony Boyles[117], Loo Charis[118], Mingyeong Choi[119], Sean Darling-Hammond[120], Zoe E. Ferguson[121], Cheryl R. Kaiser[43], Simon T. Karg[122], Alberto López Ortega[123], Lori Mahoney[124], Melvin S. Marsh[125], Marcellin F. R. C. Martinie[54], Eli K. Michaels[126], Philip Millroth[127], Jeanean B. Naqvi[128], Weiting Ng[129], Robb B. Rutledge[130], Peter Slattery[131], Adam H. Smiley[43], Oliver Strijbis[132], Daniel Sznycer[133], Eli Tsukayama[134], Austin van Loon[135], Jan G. Voelkel[135], Margaux N. A. Wienk[75], Tom Wilkening[136], & The Forecasting Collaborative[**]

Affiliations

[1]Department of Psychology, University of Waterloo; Waterloo, Canada.

[2]Department of Psychology, Northumbria University, UK.

[3]Department of Psychology, University of Toronto Scarborough; Toronto, Canada.

[4]Department of Psychology, Arizona State University, Tempe, USA.

[5]Institute of Management Accounting and Simulation, Hamburg University of Technology; Hamburg, Germany.

[6]Department of Psychology, Middlesex University London; London, UK.

[7]Department of Experimental Psychology, University of California San Diego; San Diego, USA.

[8]Department of Psychology, Toronto Metropolitan University; Toronto, Canada.

[9]Defence Research and Development Canada; Ottawa, Canada.

[10]Department of Psychology, York University; Toronto, Canada.

[11]Department of Economics, Tilburg University, Tilburg, Netherlands.

[12]Department of Psychological Sciences, Purdue University; West Lafayette, USA.

[13]Mathematical Institute, University of Oxford; Oxford, UK.

[14]Institute of New Economic Thinking, University of Oxford, Oxford, UK.

[15]Jacobs Center for Productive Youth Development, University of Zurich; Zurich, Switzerland.

[16]Institute of Psychology, University of Prešov; Prešov, Slovakia.

[17]Institute of Social Sciences, CSPS, Slovak Academy of Sciences; Bratislava, Slovakia.

[18]Department of Mechanical and Mechatronics Engineering, University of Waterloo; Waterloo, Canada.

[19]Iscte-University Institute of Lisbon, CIS; Lisbon, Portugal.

[20]Melbourne Centre for Data Science, University of Melbourne; Melbourne, Australia.

[21]Faculty of Social and Behavioural Sciences, University of Amsterdam; Amsterdam, Netherlands.

[22]Department of Food Marketing, Haub School of Business, Saint Joseph's University; Philadelphia, USA.

[23]Department of Data Science and Artificial Intelligence, Monash University; Melbourne, Australia.

[24]Institute of Psychology, University of Wrocław; Wrocław, Poland.

[25]Department of Management, Aarhus University; Aarhus, Denmark.

[26]Anderson School of Management, UCLA; Los Angeles, USA.

[27]Department of Psychology, Dominican University of California; San Rafael, USA.

[28]Department of Industrial Economics, Tamkang University; New Taipei City, Taiwan.

[29]Department of Psychology, Michigan State University; East Lansing, USA.

[30]Department of Psychology and Counselling, Universiti Tunku Abdul Rahman; Kampar, Malaysia.

[31]Psychology Department, Grove City College; Grove City, USA.

[32]Department of Economics, Siena College; Loudonville, USA.

[33]Department of Psychology, Memorial University of Newfoundland; St. John's, Canada.

[34]UBC Sauder School of Business, University of British Columbia; Vancouver, Canada.

[35]Department of Psychology, University of Oklahoma; Norman, USA.

[36]Faculty of Psychology and Cognitive Science, Adam Mickiewicz University; Poznań, Poland.

[37]Department of Psychology, University of Groningen; Groningen, Netherlands.

[38]Laboratory for Comparative Studies in Mass Consciousness, Expert Institute, HSE University; Moscow, Russia.

[39]Faculty of Philosophy and Social Sciences, Belarusian State University; Minsk, Belarus.

[40]Department of Sociology, Radboud University; Nijmegen, Netherlands.

[41]Ontario Institute for Cancer Research; Toronto, Canada.

[42]Department of Psychology, University of Winnipeg; Winnipeg, Canada.

[43]Department of Psychology, University of Washington; Seattle, USA.

[44]Department of Counseling Psychology, University of Wisconsin - Madison; Madison, USA.

[45]Department of Leadership and Organizational Behaviour, BI Norwegian Business School; Oslo, Norway.

[46]Center for Adaptive Rationality, Max Planck Institute for Human Development; Berlin, Germany.

[47]School of Public Health Sciences, University of Waterloo; Waterloo, Canada.

[48]TU Dortmund University; Dortmund, Germany.

[49]ESCP Business School; Berlin, Germany.

[50]IESE Business School; Barcelona, Spain.

[51]School of Social Sciences, Singapore Management University; Singapore.

[52]Center for Healthy Minds, University of Wisconsin-Madison; Madison, USA.

[53]University of Queensland Business School; Brisbane, Australia.

[54]Melbourne School of Psychological Sciences, University of Melbourne; Melbourne, Australia.

[55]Department of Marketing, University of Massachusetts Dartmouth; Dartmouth, USA.

[56]Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam; Amsterdam, Netherlands.

[57]Department of Psychology, Faculty of Arts, Pavol Jozef Šafárik University; Košice, Slovakia.

[58]Cambridge Judge Business School, University of Cambridge; Cambridge, UK.

[59]Hermes Germany GmbH; Hamburg, Germany.

[60]University of Hamburg; Hamburg, Germany.

[61]Department of Psychology, Illinois State University; Normal, USA.

[62]Department of Marketing and Logistics Management, National Penghu University of Science and Technology; Magong City, Taiwan.

[63]Department of Psychology, University of Victoria; Victoria, Canada.

[64]Centre for Youth and Society, University of Victoria; Victoria, Canada.

[65]Department of Psychology, California State University - Fullerton; Fullerton, USA.

[66]Center for the Study of Human Nature, California State University - Fullerton; Fullerton, USA.

[67]Department of Psychology, Loma Linda University; Loma Linda, USA.

[68]University of Nottingham; Nottingham, UK.

[69]Department of Experimental Psychology, University College London; London, UK.

[70]Singapore Management University; Singapore.

[71]Department of Psychology, Florida State University; Tallahassee, USA.

[72]Department of Psychology, Georgetown University; Washington DC, USA.

[73]School of Anthropology & Museum Ethnography, University of Oxford; Oxford, UK.

[74]School for Business and Society, University of York; York, UK.

[75]Department of Psychology, Columbia University; New York City, USA.

[76]IHRR Forecasting Laboratory, Durham University; Durham, UK.

[77]Department of Psychology, Aberystwyth University; Aberystwyth, UK.

[78]School of Business and Law, Department of Management, University of Agder; Kristiansand, Norway.

[79]Institute of Retail Economics; Stockholm, Sweden.

[80]Institute of Psychology, University of Silesia in Katowice, Poland.

[81] Department of Neurology, Penn Center for Neuroaesthetics, University of Pennsylvania, United States.

[82]Central Connecticut State University; New Britain, USA.

[83]Department of Psychology, University of Arizona; Tucson, USA.

[84]Faculty of Education, Institute for Research and Development of Education, Charles University; Prague, Czech Republic.

[85]Faculty of Education; University of Prešov; Prešov, Slovakia.

[86]School of Psychology, Macquarie University; Sydney, Australia.

[87]Department of Human Service, University of Haifa; Haifa, Israel.

[88]Promenta Center, Department of Psychology, University of Oslo; Oslo, Norway.

[89]Department of Sociology, Cornell University; Ithaca, USA.

[90]Institute of Psychology, University of Bamberg; Bamberg, Germany.

[91]IE Business School, IE University; Madrid, Spain.

[92]Faculty of Psychology and Human Movement Science, University of Hamburg; Hamburg, Germany.

[93]Department of Computer and Information Science, University of Pennsylvania; Philadelphia, USA.

[94]Department of Clinical Neuroscience, Karolinska Institutet; Solna, Sweden.

[95]Department of Sociology, University of Oxford; Oxford, UK.

[96]Department of Decision Analytics and Risk, University of Southampton; Southampton, UK.

[97]University of Chicago Booth School of Business; Chicago, USA.

[98]Department of Psychology, University of Virginia; Charlottesville, USA.

[99]Psychology Department, University of Pennsylvania; Philadelphia, USA.

[100]Wharton School of Business, University of Pennsylvania; Philadelphia, USA.

[101]Department of Economics, National and Kapodistrian University of Athens; Athens, Greece.

[102]School of Psychological Sciences and Health, University of Strathclyde; Glasgow, UK.

[103]Department of Psychology, University of Guelph; Guelph, Canada.

[104]Psychology Department, Hebrew University of Jerusalem; Jerusalem, Israel.

[105]Department of Economics, Nottingham Trent University; Nottingham, UK.

[106]Department of Management and Organizations, Northwestern University; Evanston, USA.

[107]College of Human Ecology, Cornell University, Ithaca, USA.

[108]Harvard Kennedy School, Harvard University; Cambridge, USA.

[109]Computer and Information Science, University of Pennsylvania; Philadelphia, USA.

[110]Operations, Information, and Decisions Department, The Wharton School, University of Pennsylvania; Philadelphia, USA.

[111]School of Economics and Finance, University of St. Andrews; St. Andrews, UK.

[112]Department of Management, Cameron School of Business, University of North Carolina; Wilmington, USA.

[113]School of Psychology, Speech and Hearing, University of Canterbury; Christchurch, New Zealand, [114]Deparment of Marketing, University of Dayton; Dayton, USA.

[115]ISG Universidade Lusofona; Lisbon, Portugal.

[116]Cognitive Science, University of Arizona; Tucson, USA.

[117]Ephemer AI; Atlanta, USA.

[118]Questrom School of Business, Boston University; Boston, USA.

[119]Institute of Social Science Research, Pusan National University; Busan, South Korea, [120]Fielding School of Public Health, UCLA; Los Angeles, USA.

[121]Psychology Department, University of Washington; Seattle, USA.

[122]Department of Political Science, Aarhus University; Aarhus, USA.

[123]Faculty of Social Sciences, Vrije Universiteit Amsterdam; Amsterdam, Netherlands.

[124]College of Science and Mathematics, Wright State University; Fairborn, USA.

[125]Department of Psychology, Georgia Southern University; Statesboro, USA.

[126]Division of Epidemiology, School of Public Health, University of California, Berkeley; Berkeley, USA.

[127]Department of Psychology, Uppsala University; Uppsala, Sweden.

[128]Department of Psychology, Carnegie Mellon University; Pittsburgh, USA.

[129]School of Humanities & Behavioral Sciences, Singapore University of Social Sciences; Singapore.

[130]Department of Psychology, Yale University; New Haven, USA.

[131]BehaviourWorks Australia. Monash University; Melbourne, Australia.

[132]Institute of Political Science, University of Zurich; Zurich, Switzerland.

[133]Department of Psychology, Oklahoma State University; Stillwater, USA.

[134]Department of Business Administration, University of Hawaii - West Oahu; Kapolei, USA.

[135]Department of Sociology, Stanford University; Stanford, USA.

[136]Department of Economics, University of Melbourne; Melbourne, Australia.

*Corresponding author. Email: igrossma@uwaterloo.ca

**A list of authors and their affiliations appears at the end of the paper.

Abstract

How well can social scientists predict societal change, and what processes underlie their predictions? To answer these questions, we ran two forecasting tournaments testing accuracy of predictions of societal change in domains commonly studied in the social sciences: ideological preferences, political polarization, life satisfaction, sentiment on social media, and gender-career and racial bias. Following provision of historical trend data on the domain, social scientists submitted pre-registered monthly forecasts for a year (Tournament 1; *N*=86 teams/359 forecasts), with an opportunity to update forecasts based on new data six months later (Tournament 2; *N*=120 teams/546 forecasts). Benchmarking forecasting accuracy revealed that social scientists' forecasts were on average no more accurate than simple statistical models (historical means, random walk, or linear regressions) or the aggregate forecasts of a sample from the general public (*N*=802). However, scientists were more accurate if they had scientific expertise in a prediction domain, were interdisciplinary, used simpler models, and based predictions on prior data.

Main Text

Can social scientists predict societal change? Governments and the general public often rely on experts, based on a general belief that they make better judgments and predictions of the future in their domain of expertise. The media also seek out experts to render their judgments and opinions about what to expect in the future [1,2]. Yet research on predictions in many domains suggests that experts may not be better than purely stochastic models in predicting the future. For example, portfolio managers (who are paid for their expertise) do not outperform the stock market in their predictions [3]. Similarly, in the domain of geopolitics, experts often perform at chance levels when forecasting occurrences of specific political events [4]. Based on these insights, one might expect that experts would find it difficult to accurately predict societal change.

At the same time, social science researchers have developed rich, empirically-grounded models to explain social science phenomena. Examining sampled data, social scientists strive to develop theoretical models about causal mechanisms that, in ideal cases, reliably describe human behavior and societal processes [5]. Therefore, it is possible that explanatory models afford social science experts an advantage in predicting social phenomena in their domain of expertise. Here, we test these possibilities, examining the overall predictability of trends in social phenomena such as political polarization, racial bias, or well-being, and whether experts in social science are better able to predict those trends compared to non-experts.

Prior forecasting initiatives have not fully addressed this question for two reasons. First, forecasting initiatives with subject matter experts have focused on examining the probability of occurrence for specific one-time events [4,6] rather than the accuracy of *ex-ante* predictions of societal change over multiple units of time. In a sense, predicting events in the future (ex-ante) is the same as predicting events that have already happened, as long as the experts (research participants) don't know the outcome. Yet, there are reasons to think that future prediction is different in an important way. Consider stock prices: participants could predict stock returns for stocks in the past, except that they know many other things that have happened (conflicts, bubbles, Black Swans, economic trends, consumption trends, etc.). Post-hoc, those making those making predictions have access to the temporal variance/occurrence for each of these variables and hence are more likely to be successful in ex post predictions. Thus, predictions about past events end up being more about testing people's explanations, rather than their predictions per se. Moreover, all other things being equal, the likelihood of a prediction regarding a one-off event being accurate is by default higher than that of a prediction regarding societal change across an extended time period. Binary predictions for the one-off event do not require accuracy in estimating degree of change or the shape of the predicted time series, which are extra challenges in forecasting societal change.

Second, past research on forecasting has concentrated on predicting geopolitical [4] or economic events [7] rather than broader societal phenomena. Thus, in contrast to systematic studies concerning the replicability of in-sample explanations of social science phenomena [8], out-of-sample prediction accuracy in the social sciences remains understudied [9,10]. Similarly, little is known about the rationales and approaches social scientists use to make predictions for societal trends. For example, are social scientists more apt to rely on data-driven statistical methods or theory and intuitions when generating such predictions?

To address these unknowns, we performed a standardized evaluation of forecasting accuracy [9] among social scientists in well-studied domains for which systematic, cross-temporal data is available, namely subjective well-being, racial bias, ideological preferences, political polarization, and gender-career bias. With the onset of the COVID-19 pandemic as a backdrop, we selected these domains based on data availability and theoretical links to the pandemic. Prior research has suggested that each of these domains may be impacted by infectious disease [11–14] or pandemic-related social isolation [15]. To understand how scientists made predictions in these domains, we documented the rationales and processes they used to generate forecasts, then examined how different methodological choices were related to accuracy.

Research Overview

We present results from two forecasting tournaments conducted through the Forecasting Collaborative—a crowdsourced initiative among scientists interested in ex-ante testing of their theoretical or data-driven models. Examining performance across two tournaments allowed us to test the stability of forecasting accuracy in the context of unfolding societal events, and to investigate how social scientists recalibrate their models and incorporate new data when asked to update their forecasts.

The Forecasting Collaborative was open to behavioral, social, and data scientists from any field who wanted to participate in the tournament and were willing to provide forecasts over 12 months (May 2020 – April 2021) as part of the initial tournament and, upon receiving feedback on initial performance, again after 6 months for a follow-up tournament (recruitment details in Methods and demographic information in Supplementary Table 1). To ensure a "common task framework" [9,16,17], we provided all participating teams with the same time series data for the US for each of the 12 variables related to the phenomena of interest (i.e., life satisfaction, positive affect, negative affect, support for Democrats, support for Republicans, political polarization, explicit and implicit attitudes towards Asian Americans, explicit and implicit attitudes towards African Americans, and explicit and implicit associations between gender and specific careers.

Participating teams received historical data that spanned 39 months (January 2017 to March 2020) for Tournament 1 and data that spanned 45 months for Tournament 2 (January 2017 to September 2020), which they could use to inform their forecasts for the future values of the same time series. Teams could select up to 12 domains to forecast, including domains for which team members reported a track record of peer-reviewed publications as well as domains for which they did not possess relevant expertise (see Methods for multi-stage operationalization of expertise). By including social scientists with expertise in different subject matters, we could examine how such expertise may contribute to forecasting accuracy above and beyond general training in the social sciences. Teams were not constrained in terms of the methods used to generate time-point forecasts. They provided open-ended, free-text responses for the descriptions of the methods used, which were coded later. If they made use of data-driven methods, they also provided the model and any additional data used to generate their forecasts (see Methods). We also collected data on team size and composition, area of research specialization, subject domain and forecasting expertise, and prediction confidence.

We benchmarked forecasting accuracy against several alternatives. First, we evaluated whether social scientists' forecasts in Tournament 1 were better than the wisdom of the crowd (i.e., the average forecasts of a sample of lay participants recruited from Prolific). Second, we compared social scientists' performance in both tournaments to naïve random extrapolation algorithms (i.e., the average of historical data, random walks, and estimates based on linear trends). Finally, we systematically evaluated the accuracy of different forecasting strategies used by the social scientists in our tournaments, as well as the effect of expertise.

Results

Following the a priori outlined analytic plan (osf.io/7ekfm; details in the Supplementary Methods) to determine forecasting accuracy across domains, we examined the mean absolute scaled error (MASE) [18] across forecasted time-points for each domain. MASE is an asymptotically normal, scale-independent scoring rule that compares predicted values against predictions of a one-step random walk. Because it is scale-independent, it is an adequate measure when comparing accuracy across domains on different scales. A MASE of 1 reflects a forecast that is as good out-of-sample as the naive one-step random walk forecast is in-sample. A MASE below 1.76 is superior to median performance in prior large-scale data science competitions [7]. See Supplementary Materials for further details of the MASE method.

In addition to absolute accuracy, we also assessed the comparative accuracy of social scientists' forecasts using several benchmarks. First, during the period of the first tournament, we obtained forecasts from a non-expert crowdsourced sample of US residents ($N = 802$) via Prolific [19] who received the same data as tournament participants and filled out an identically structured survey to provide a wisdom-of-the-(lay)-crowd benchmark. Second, for both tournaments we simulated three different data-based naïve approaches to out-of-sample forecasting using the time series data provided to participants in the tournament, including, 1) the historical mean, calculated by randomly resampling the historical time series data; 2) a naïve random walk, calculated by randomly resampling historical change in the time series data with an autoregressive component; 3) extrapolation from linear regression, based on a randomly selected interval of historical time series data (see Supplementary Information for details). This latter approach captures the expected range of predictions that would have resulted from random, uninformed use of historical data to make out-of-sample predictions (as opposed to the naïve in-sample predictions that form the basis of MASE scores).

<u>How accurate were behavioral and social scientists at forecasting?</u>

Fig. 1 shows that in Tournament 1, social scientists' forecasts were, on average, inferior to in-sample random walks in nine domains. In seven domains, social scientists' forecasts were inferior to median performance in prior forecasting competitions (Supplementary Fig. 1 shows raw estimates; Supplementary Fig. 2 reports measures of uncertainty around estimates). In Tournament 2, forecasts were on average inferior to in-sample random walks in eight domains and inferior to median performance in prior forecasting competitions in five domains. Even winning teams were still less accurate than in-sample random walks for 8 of 12 domains in Tournament 1, and one domain (Republican support) in Tournament 2 (Supplementary Tables 1-2 and Supplementary Figs. 4-9). One should note that inferior performance to the in-sample random walk (MASE > 1) may not be too surprising; errors of the in-sample random walk in the denominator concern historical observations that occurred before the pandemic, whereas accuracy of scientific forecasts in the numerator is compared concerns the data for the first pandemic year. However, average forecasting accuracy did not generally beat more liberal benchmarks such as the median MASE in data science tournaments (1.76) [7] or the benchmark MASE for "good" forecasts in the tourism industry (see Supplement). Except for one team, top forecasters from Tournament 1 did not appear among the winners of Tournament 2 (Supplementary Tables 1-2).

We examined the accuracy of scientific and lay forecasts in a linear mixed effect model. To systematically compare results for different forecasted domains, we tested a full model with expertise (social scientist versus lay crowd), domain, and their interaction as predictors, and *log*(MASE) scores nested in participants. We observed no significant main effect difference between accuracy of social scientists and lay crowds, $F(11, 1747) = 0.88$, $P = .348$, *part* $R^2 < .001$. However, we observed a significant interaction between social science training and domain, $F(11, 1304) = 2.00$, $P = .026$. Simple effects show that social scientists were significantly more accurate than lay people when forecasting life satisfaction, polarization, as well as explicit and implicit gender-career bias. However, the scientific teams were no better than the lay sample in the remaining eight domains (Figure 1 and Table 1). Moreover, Bayesian analyses indicated that only for life satisfaction there is substantial evidence in favor of the difference, whereas for eight domains evidence was in favor of the null hypothesis. See supplementary information online for further details and interpretation of the multiverse analyses of domain-general accuracy.

Cross-validation of domain-general accuracy via forecast versus trend comparisons

The most elementary analysis of domain-general accuracy involves inspecting trends for each group and comparing them against the ground truth and historical time series in each domain. Fig. 2 allows us to inspect individual trends of social scientists and the naïve crowd per domain in Tournament 1, along with historical and ground truth markers for each domain. For social scientists, one can observe the diversity of forecasts from individual teams (light blue) along with a lowess regression and 95% confidence interval around the trend (blue). For the naïve crowd, one can see an equivalent lowess trend and the 95% CI around it (salmon). In half of the domains – explicit bias against African Americans, implicit bias against Asian-Americans, negative affect, life satisfaction, as well as support for Democrats and Republicans –lowess curves from both groups were overlapping, suggesting that the estimates from both social scientists and the naïve crowd were identical. Moreover, except for the domain of life satisfaction, forecasts of scientists and the naïve crowd were close to far off the mark vis-à-vis ground truth. In one further domain— explicit bias against African Americans—the naïve crowd estimate was in fact closer to the ground truth marker than the estimate from the lowess curve of the social scientists. In other five domains, which concerned explicit and implicit gender career bias, explicit bias against Asian-Americans, positive affect and political polarization, social scientists' forecasts were closer to the ground truth markers than the naïve crowd. We note, however, that these visual inspections may be somewhat misleading because the confidence intervals don't correct for multiple tests. This caveat aside, the overall message remains consistent with the results of the statistical tests above: For most domains social scientists' predictions were either similar to or worse than the naïve crowd's predictions.

Comparisons to naïve statistical benchmarks

Next, we compared scientific forecasts against three naïve statistical benchmarks by creating benchmark/forecast ratio scores (a ratio of 1 indicates that the social scientists' forecasts were equal in accuracy to the benchmarks, with ratios greater than 1 indicating greater accuracy). To account for interdependence of social scientists' forecasts, we examined estimated ratio scores for domains from linear mixed models, with responses nested in forecasting teams. To reduce the likelihood that social scientists' forecasts beat naïve benchmarks by chance, our main analyses focus on performance across all three benchmarks (see Supplement for rationale favoring this method over averaging across three benchmarks), and by adjusting confidence intervals of the ratio scores for simultaneous inference of 12 domains in each tournament by simulating a multivariate *t* distribution[20]. Figs. 1 and 3 and Supplementary Fig. 2 show that social scientists in Tournament 1 were significantly better than each of the three benchmarks in only one out of twelve domains, which concerned explicit gender-career bias, $1.53 < \text{ratio} \leq 1.60$, $1.16 < 95\%CI \leq 2.910$. In the remaining 11 domains, scientific predictions were either no different or worse than the benchmarks. The relative advantage of scientific forecasts over the historical mean and random walk benchmarks was somewhat larger in Tournament 2 (Supplementary Fig. 1). Scientific forecasts were significantly more accurate than the three naïve benchmarks in five out of twelve domains. These domains reflected explicit racial bias: African-American bias, 2.20 < ratio ≤ 2.86, 1.55 < 95%*CI* ≤ 4.05; Asian-American bias, 1.39 < ratio ≤ 3.14, 1.01 < 95%*CI* ≤ 4.40; and implicit racial and gender career biases: African-American bias, 1.35 < ratio ≤ 2.00, 1.35 < 95%*CI* ≤ 2.78; Asian-American bias, 1.36 < ratio ≤ 2.73, 1.001 < 95%*CI* ≤ 3.71; gender-career bias, 1.59 < ratio ≤ 3.22, 1.15 < 95%*CI* ≤ 4.46. In the remaining seven domains, forecasts were not significantly different from naïve benchmarks. Moreover, as Fig. 3 shows, for political polarization scientific forecasts in Tournament 2 were significantly *less* accurate than estimates from a naïve linear regression, ratio = 0.51, 95%*CI* [0.38, 0.68]. Fig. 3 also shows that in most domains at least one of the naïve forecasting methods produced errors that were comparable to or less than social scientists' forecasts (11 out of 12 in Tournament 1 and 8 out of 12 in Tournament 2).

To compare social scientists' forecasts against the average of three naïve benchmarks, we fit a linear mixed model with forecast/benchmark ratio scores nested in forecasting teams and examined estimated means for each domain. In Tournament 1, scientists performed better than the average of the naïve benchmarks in only three domains, which concerned political polarization, 95%*CI* [1.06; 1.63], explicit gender-career bias, 95%*CI* [1.23; 1.95], and implicit gender-career bias, 95%*CI* [1.17; 1.83]. In Tournament 2, social scientists performed better than the average of the naïve benchmarks in seven domains, 1.07 < 95%*CI*s ≤ 2.79, while they were statistically indistinguishable from the average of naïve benchmarks when forecasting the remaining five domains: ideological support for Democrats, 95%*CI* [0.76; 1.17], and for Republicans, 95%*CI* [0.98; 1.51], explicit gender-career bias, 95%*CI* [0.96; 1.52], and negative affect on social media, 95%*CI* [0.82; 1.25]. Moreover, in Tournament 2 social scientists' forecasts of political polarization were inferior to the average of naïve benchmarks, 95%*CI* [0.58; 0.89]. Overall, social scientists tended to do worse than the average of the three naïve statistical benchmarks in Tournament 1. While scientists did better than the average of naïve benchmarks in Tournament 2, this difference in overall performance was small, *M*(forecast /benchmark inaccuracy ratio) = 1.43, 95%*CI* [1.26; 1.62]. Moreover, in most domains at least one of the naïve benchmarks was on par if not more accurate to social scientists' forecasts.

<u>Which domains were harder to predict?</u>

Fig. 4 shows that some societal trends were significantly harder to forecast than others, Tournament 1: $F(11,295.69) = 41.88$, $P < .001$, $R^2 = .450$, Tournament 2: $F(11,469.49) = 26.87$, $P < .001$, $R^2 = .291$. Forecast accuracy was lowest in politics (underestimating Democratic and Republican support, and political polarization), well-being (underestimating life satisfaction and negative affect on social media), and racial bias against African Americans (overestimating; also see Supplementary Fig. 1). Differences in forecast accuracy across domains did not correspond to differences in quality of ground truth markers: Based on the sampling frequency and representativeness of the data, most reliable ground truth markers concerned societal change in political ideology, obtained via an aggregate of multiple nationally representative surveys by reputable pollsters, yet this domain was among most difficult to forecast. In contrast, some of the least representative markers concerned racial and gender-bias, which came from Project Implicit—a volunteer platform that is subject to self-selection bias—yet these domains were among the easiest to forecast. In a similar vein, both life satisfaction and positive affect on social media were estimated via texts on Twitter, even though forecasting errors between these domains varied. Though measurement imprecision undoubtedly presents a challenge for forecasting, it is unlikely to account for between-domain variability in forecasting errors (Figure 4).

Domain differences in forecasting accuracy corresponded to differences in the complexity of historical data: domains ranked more variable in terms of standard deviation (*SD*) and mean absolute difference (*MAD*) of historical data tended to have more forecasting error (as measured by the rank-order correlation between median inaccuracy scores across teams and variability scores for the same domain), Tournament 1: $\rho(SD) = .19$, $\rho(MAD) = .20$; Tournament 2: $\rho(SD) = .48$, $\rho(MAD) = .36$, and domain changes in variability of historical data across tournaments corresponded to changes in accuracy, $\rho(SD) = .27$, $\rho(MAD) = .28$.

<u>Comparison of accuracy across tournaments</u>

Forecasting error was higher in the first tournament than the second tournament (see Fig. 4), $F(1, 889.48) = 64.59$, $P < .001$, $R^2 = .063$. We explored several possible differences between the tournaments that may account for this effect. One possibility is that type of teams differed between tournaments (team size, gender, number of forecasted domains, field specialization and team diversity, number of PhDs on a team, prior experience with forecasting). However, the difference between the tournaments remained equally pronounced when running parallel analyses with team characteristics as covariates, $F(1, 847.79) = 90.45$, $P < .001$, $R^2 = .062$.

Another hypothesis is that forecasts for twelve months (Tournament 1) include further removed data points than forecasts for more immediate six months (Tournament 2), and it is the greater temporal distance between the tournament and the moment to forecast that results in greater inaccuracy at Tournament 1. To test this hypothesis, we zeroed in on Tournament 1 inaccuracy scores for the first and the last six months, while including domain type as a control dummy variable. By focusing on Tournament 1 data, we kept other characteristics such as team composition as a constant. Contrary to this seemingly straightforward hypothesis, error for the forecasts for the first six months was in fact significantly greater (MASE = 3.16, $SE = 0.21$, 95%*CI* [2.77, 3.60]) than for the last six months (MASE = 2.59, $SE = 0.17$, 95%*CI* [2.27, 2.95]), $F(1, 621.41) = 29.36$, $P < .001$, $R^2 = .012$. As Supplementary Fig. 1 shows, for many domains, social scientists underpredicted societal change in Tournament 1, and this difference between predicted and observed values was more pronounced in the first versus last six months. This suggests that for several domains social scientists anchored (*19*) their forecasts on the most recent historical data. Figure 2 further indicates that many domains showed unusual shifts (vis-à-vis prior historical data) in the first six months of the pandemic and started to return to the historical baseline in the following six months. For these domains, forecasts anchored on the most recent historical data were more inaccurate for the May-October 2020 forecasts compared to the November 2020-April 2021 forecasts.

Finally, we tested whether providing teams additional six months of historical trend capturing the on-set of the novel pandemic at Tournament 2 may have contributed to lower error compared to Tournament 1. To this end, we compared the inaccuracy of forecasts for the six months period of Nov 2020-April 2021 done in May 2020 (Tournament 1) and when provided with more data in October 2020 (Tournament 2). We focused only on participants who completed both tournaments to keep the number of participating teams and team characteristics constant. Indeed, Tournament 1 forecasts had significantly more error (MASE $M = 2.54$, $SE = 0.17$, 95%*CI* [2.23, 2.90]) than Tournament 2 forecasts (MASE $M = 1.99$, $SE = 0.13$, 95%*CI* [1.74, 2.27]), $F(1, 607.79) = 31.57$, $P < .001$, $R^2 = .017$. suggesting that it was availability of new (pandemic-specific) information rather than temporal distance that contributed to more accurate forecasts in the second compared to the first tournament.

Consistency in forecasting

Despite variability across scientific teams, domains, and tournaments, the accuracy of scientific predictions was highly systematic. Accuracy in one subset of predictions (ranking of model performance across odd months) was highly correlated with accuracy in the other subset (ranking of model performance across even months), first tournament: multilevel $r_{\text{across domains}}$ = .88, 95% CI [.85; .90], $t(357) = 34.80$, $P < .001$; domain-specific .55 < .rs ≤ .99; second tournament: multilevel $r_{\text{across domains}}$ = .72, 95% CI [.67; .75], $t(544) = 23.95$, $P < .001$; domain-specific .24 < .rs ≤ .96. Further, results of a linear mixed model with MASE scores in Tournament 1, domain, and their interaction predicting MASE in Tournament 2 showed that for eleven out of twelve domains accuracy in Tournament 1 was associated with greater accuracy in Tournament 2, Md(standardized β) = .26.

Moreover, the ranking of models based on performance in the initial 12-months tournament corresponds to the ranking of the updated models in the follow-up 6-months tournament (Fig. 4). Harder-to-predict domains in the initial tournament remained most inaccurate in the second tournament. Fig. 3 shows one notable exception. Bias against African Americans was easier to predict than other domains in the second tournament. Though speculative, this exception appears consistent with the idea that George Floyd's death catalyzed movements in racial awareness just after the first tournament (Supplementary Fig. 14 for a timeline of major historical events).

Which strategies and team characteristics promoted accuracy?

Finally, we examined forecasting approaches and individual characteristics of more accurate forecasters in the tournaments. In the main text, we focused on central tendencies across forecasting teams, whereas in the supplementary analyses we reviewed strategies of winning teams and characteristics of the top 5 performers in each domain (Supplementary Figs. 4-11). We compared forecasting approaches relying on 1) no data modeling (but possible consideration of theories); 2) pure data modeling (but no consideration of subject matter theories); 3) hybrid approaches. Roughly half of the teams relied on data-based modeling as a basis for their forecasts, whereas the other half of the teams in each tournament relied only on their intuitions or theoretical considerations (Fig. 5). This pattern was similar across domains (Supplementary Fig. 3).

In both tournaments, pre-registered linear mixed model analyses with approach as a factor, domain type as a control dummy variable, and MASE scores nested in forecasting teams as a dependent variable revealed that forecasting approaches significantly differed in accuracy, first tournament: $F(2, 149.10) = 5.47$, $P = .005$, $R^2 = .096$; second tournament: $F(2, 177.93) = 5.00$, $P = .008$, $R^2 = .091$ (Fig. 5). Forecasts that considered historical data as part of the forecast modeling were more accurate than models that did not, first tournament: $F(1, 56.29) = 20.38$, $P < .001$, $R^2 = .096$; second tournament: $F(1, 159.11) = 8.12$, $P = .005$, $R^2 = .084$. Model comparison effects were qualified by significant model type X domain interaction; first tournament: $F(11, 278.67) = 4.57$, $P < .001$, $R^2 = .045$; second tournament: $F(11, 462.08) = 3.38$, $P = .0002$, $R^2 = .028$. Post-hoc comparisons in Supplementary Table 4 revealed that data-inclusive (data-driven and hybrid) models were significantly more accurate than data-free models that did not include data in three domains (explicit and implicit racial bias against Asian-Americans and implicit gender-career bias) in Tournament 1 and two domains (life satisfaction, explicit gender-career bias) in Tournament 2. There were no domains where data-free models were more accurate than data-inclusive models. Analyses further demonstrated that, in the first tournament, data-free forecasts of social scientists were not significantly better than lay estimates, $t(577) = 0.87$, $P = .385$, whereas data-inclusive models tended to perform significantly better than lay estimates, $t(470) = 3.11$, $P = .006$, *Cohen's d* = 0.391.

To examine the incremental contribution of specific forecasting strategies and team characteristics to accuracy, we pooled data from both tournaments in a linear mixed model with inaccuracy (MASE) as a dependent variable. As Fig. 6 shows, we included predictors representing forecasting strategies, team characteristics, domain expertise (quantified via publications by team members on the topic) and forecasting expertise (quantified via prior experience with forecasting tournaments). We further included domain type as a control dummy variable, and nested responses in teams.

The full model fixed effects explained 31% of the variance in accuracy ($R^2 = .314$), though much of it was accounted for by differences in accuracy between domains (non-domain R^2 [partial] = .043). Consistent with prior research [22], model sophistication—i.e., considering a larger number of exogenous predictors, COVID-trajectory, or counterfactuals—did not significantly improve accuracy (Fig. 6 and Supplementary Table 5). In fact, forecasting models based on simpler procedures turned out to be significantly more accurate than complex models, $B = 0.14$, $SE = 0.06$, $t\,(220.82) = 2.33$, $P = .021$, R^2(partial) = .010.

On the one hand, experts' subjective confidence in their forecasts was not related to the accuracy of their estimates. On the other, people with expertise made more accurate forecasts. Teams were more accurate if they had members who published academic research on the forecasted domain, $B = -0.26$, $SE = 0.09$, $t\,(711.64) = 3.01$, $P = .003$, R^2(partial) = .007, and who took part in prior forecasting competitions, $B = -0.35$, $SE = 0.17$, $t\,(56.26) = 2.02$, $P = .049$, R^2(partial) = .010 (also see Supplementary Table 5). Critically, even though some of these effects were significant, only two factors – complexity of the statistical method and prior experience with forecasting tournaments – showed a non-negligible partial effect size (R^2 above .009). Additional testing whether inclusion of US-based scientists influenced forecasting accuracy did not yield significant effects, $F(1, 106.61) < 1$, *ns*.

In the second tournament, we provided teams with the opportunity to compare their original forecasts (Tournament 1, May 2020) to new data at a later point of time and to update their predictions (*22*) (Tournament 2, Nov 2020). Therefore, we tested whether updating improved people's predictive accuracy. Out of the initial 356 forecasts in the first tournament, 180 were updated in the second tournament (from 37% of teams for life satisfaction to 60% of teams for implicit Asian American bias). Updated forecasts in the second tournament (November) were significantly more accurate than the original forecasts in the first tournament (May), $t(94.5) = 6.04$, $P < .001$, *Cohen's d* = 0.804, but so were forecasts from the 34 new teams recruited in November, $t(75.9) = 6.30$, $P < .001$, *Cohen's d* = 0.816. Furthermore, updated forecasts were not significantly different from forecasts provided by new teams recruited in November, $t(77.8) < 0.10$, $P = .928$. This observation suggests that updating did not lead to more accurate forecasts (Supplementary Table 6 reports additional analyses probing different updating rationales).

Discussion

How accurate are social scientists' forecasts of societal change[23]? Results from two forecasting tournaments conducted during the first year of the COVID-19 pandemic show that for most domains social scientists' predictions were no better than those from a sample of the (non-specialist) general public. Further, apart from a few domains concerning racial and gender-career bias, scientists' original forecasts were typically not much better than naïve statistical benchmarks derived from historical averages, linear regressions, or random walks. Even when confining the analysis to the top 5 forecasts by social scientists per domain, a simple linear regression produced less error roughly half of the time (Supplementary Figs. 5 and 9).

Forecasting accuracy systematically varied across societal domains. In both tournaments, positive sentiment and gender-career stereotypes were easier to forecast than other phenomena, whereas negative sentiment and bias toward African Americans were most difficult to forecast. Domain differences in forecasting accuracy corresponded to historical volatility in the time series. Differences in the complexity of positive and negative affect are well-documented [25,26]. Moreover, racial attitudes showed more change than attitudes regarding gender during this period (perhaps due to movements like Black Lives Matter).

Which strategies and team characteristics were associated with more effective forecasts? One defining feature of more effective forecasters was that they relied on prior data rather than theory alone. This observation fits with prior studies on the performance of algorithmic versus intuitive human judgments[22]. Social scientists who relied on prior data also performed better than lay crowds and were overrepresented among the winning teams (Supplementary Figs. 4 and 8).

Forecasting experience and subject matter expertise on a forecasted topic also incrementally contributed to better performance in tournaments (R^2 [partial] = .010). This is in line with some prior research on the value of subject-matter expertise for geopolitical forecasts [6] and for prediction of success of behavioral science interventions [27]. Notably, we found that publication track record on a topic, rather than subjective confidence in domain expertise or confidence in the forecast, contributed to greater accuracy. It is possible that subjective confidence in domain expertise conflates expertise and overconfidence [28] (versus intellectual humility). There is some evidence that overconfident forecasters are less accurate [29]. These findings, along with a lack of domain-general effect of social science expertise on performance compared to the general public, invite consideration of whether what usually counts as expertise in social sciences translates into greater ability to predict future real-world trends.

The nature of our forecasting tournaments allowed social scientists to self-select any of the twelve forecasting domains, inspect three years of historical trends for each domain, and to update their predictions based on feedback on their initial performance in the first tournament. These features emulated typical forecasting platforms (e.g., metaculus.com). We argue that this approach enhances our ability to draw externally valid and generalizable inferences from a forecasting tournament. However, this approach also resulted in a complex, unbalanced design. Scholars interested in isolating psychological mechanisms fostering superior forecasts may benefit from a simpler design whereby all forecasting teams make forecasts for all requested domains.

Another issue in designing forecasting tournaments involves determination of domains one may want participants to forecast. In designing the present tournaments, we provided participants with at least three years of monthly historical data for each forecasting domain. An advantage of making the same historical data available for all forecasters is that it establishes a "common task framework" [9,16,17], ensuring main sources of information about the forecasting domains remain identical across all participants. However, this approach restricts the types of social issues participants can forecast. A simpler design without inclusion of historical data would have had the advantage of a greater flexibility in selecting forecasting domains.

Why were forecasts of societal change largely inaccurate, even though participants had data-based resources and ample time to deliberate? One possibility concerns self-selection. Perhaps participants in the Forecasting Collaborative were unusually bad at forecasting compared to social scientists as a whole. This possibility seems unlikely. We made efforts to recruit highly successful social scientists at different career stages and from different sub-disciplines (see Supplementary Materials). Indeed, many of our forecasters are well-established scholars. Thus, we do not expect members of the Forecasting Collaborative to be worse at forecasting than other members of the social science community. Nevertheless, albeit impractical, only a random sample of social scientists would have fully addressed the self-selection concern.

Second, it is possible that social scientists were not adequately incentivized to perform well in our tournaments. Though we provided reputational incentives by announcing winners and ranking of participating teams, like other big-team science projects [8,30] we did not provide performance-based monetary incentives [31], because they may not be key motivating factors for intrinsically motivated social scientists [32]. Indeed, the drop-out rate between Tournaments 1 and 2 was marginal, suggesting that participating teams were motivated to continue being part of the initiative. This reasoning aside, it is possible that stronger incentives for accurate forecasting (whether reputation-based or monetary) could have stimulated some scientists to perform better in our forecasting tournament, opening doors for future directions to address this question directly.

Third, social scientists often deal with phenomena with small effect sizes that are overestimated in the literature [8,30,33]. Additionally, social scientists frequently study social phenomena in conditions that maximize experimental control but may have little external validity, and it is argued that this not only limits the generalizability of findings but in fact reduces their internal validity. In the world beyond the laboratory, where more factors are at play, such effects may be smaller than social scientists might think based on their lab studies, and in fact, such effects may be spurious given the lack of external validity. Thus, social scientist may over(and mis)estimate the impact of the effects they study in the lab on real-world phenomena [34]

Fourth, social scientists tend to theorize about individuals and groups and conduct research at those scales. However, findings from such work may not scale up when predicting phenomena on a scale of entire societies [39]. Like other dynamical systems in economics, physics, or biology, societal level processes may also be genuinely stochastic rather than deterministic. If so, stochastic models will be hard to outperform.

Fifth, training in predictive modeling is not a requirement in many social sciences programs [10]. Social scientists often prioritize explanations over formal predictions [5]. For instance, statistical training in social sciences typically emphasizes unbiased estimation of model parameters in the sample over predictive out-of-sample accuracy [40]. Moreover, typical graduate curricula in many areas of social science, such as social or clinical psychology, do not require computational training in predictive modeling. The formal empirical study of societal change is relatively uncommon in these disciplines. Most social scientists approach individual- or group-level phenomena in an a-temporal fashion [39]. Scientists may favor post-hoc explanations of specific one-time events rather than the future trajectory of social phenomena. Although time is a key theoretical variable for foundational theories in many subfields of social sciences, such as field theory [41], it has remained an elusive concept.

Finally, perhaps it is unreasonable to expect theories and models developed during a relatively stable Post World War II period to accurately predict societal trends during a once-in-a-century health crisis. Precisely for this reason, we targeted predictions in domains possessing pandemic-relevant theoretical models (for instance, models about the impact of pathogen spread or social isolation). In this way, we sought to provide a "stress test" of ostensibly relevant theoretical models in a context (pandemic-induced crisis) where change was most likely to be both meaningful and measurable. Nevertheless, the present work suggests that social scientists may not be particularly accurate at forecasting societal trends in this context, though it remains possible that they would perform better during more “normal” times. Considerations above notwithstanding, future work should seek to address this question.

How can social scientists become better forecasters? Perhaps the first steps might involve probing the limits of social science theories by evaluating if a given theory is suitable for making societal predictions in the first place or if it is too narrow or too vague [5,42]. Relatedly, social scientists need to test their theories using representatively designed experiments. Moreover, social scientists may benefit from testing whether a societal trend is deterministic and hence can benefit from theory-driven components, or if it unfolds in a purely stochastic fashion. For instance, one can start by decomposing a time series into the trend, autoregressive, and seasonal components, examining each of them and their meaning for one’s theory and model. One can further perform a unit root test to examine whether the time series is non-stationary. Training in recognizing and modeling properties of time series and dynamical systems may need to become more firmly integrated into graduate curricula in the field. A classic insight in the time series literature is that the mean of the historical time series may be among the best multi-step ahead predictor for a stationary time series [43]. Using such insights to build predictions from the ground-up can afford greater accuracy. In turn, such training can open the door to more robust models of social phenomena and human behavior, with a promise of greater generalizability in the real-world.

Given the broad societal impact of phenomena like prejudice, political polarization or well-being, the ability to accurately predict trends in these variables would appear to be of crucial importance for policy makers and the experts guiding them. But despite common beliefs that social science experts are best equipped to accurately predict these trends compared to non-experts [1], the current findings suggest that social and behavioral scientists have a lot of room for growth [38]. The good news is that forecasting skills can be improved. Consider the growing accuracy in forecasting models in meteorology in the second part of the 20th century [44]. Greater consideration of representative experimental designs, temporal dynamics, better training in forecasting methods, and more practice with formal forecasting all may improve social scientists’ ability to accurately forecast societal trends going forward.

Methods

The study was approved by the Office of Research Ethics of the University of Waterloo under protocol # 42142.

Pre-registration and deviations. Forecasts of all participating teams along with their rationales were pre-registered on Open Science Framework (https://osf.io/6wgbj/registrations). Additionally, in an a priori specific document shared with the journal in April 2020, we outlined the operationalization of the key dependent variable (MASE), operationalization of covariates and benchmarks (i.e., use of naive forecasting methods), along with the key analytic procedures (linear mixed model and contrasts being different forecasting approaches; osf.io/7ekfm). We did not pre-register the use of a Prolific sample from the general public as an additional benchmark before their forecasting data was collected, though we did pre-register this benchmark in early September of 2020, prior to data pre-processing or analyses. Deviating from the pre-registration, to protect against inflating *p*-values, we performed a single analysis with all covariates in the same model rather than performing separate analyses for each set of covariates. Further, due to scale differences between domains, we chose not to feature analyses concerning absolute percentage errors of each time point in the main paper (but see corresponding analyses on the GitHub site of the project https://github.com/grossmania/Forecasting-Tournament, which replicate the key effects presented in the main manuscript).

Participants & recruitment. We initially aimed for a minimum sample of 40 forecasting teams in our tournament after prescreening to ensure that participants possess at minimum a bachelor's degree in behavioral, social, or computer sciences. To compare groups of scientists employing different forecasting strategies (e.g., data-free versus data-inclusive methods), we subsequently tripled the target size of the final sample (N = 120), the target we accomplished by the November phase of the tournament, to ensure sufficient sample for comparison of teams using different strategies (see Supplementary Table 1 for demographics).

The Forecasting Collaborative website we used for recruitment (https://predictions.uwaterloo.ca/faq) outlined guidelines for eligibility and approach for prospective participants. We incentivized participating teams in two ways. First, prospective participants had an opportunity for a co-authorship in a large-scale citizen science publication. Second, we incentivized accuracy by emphasizing throughout the recruitment that we will be announcing winners and will share the ranking of scientific teams in terms of performance in each tournament (per domain and in total).

As outlined in the recruitment materials, we considered data-driven (e.g., model-based) or expertise-based (e.g., general intuition, theory-based) forecasts from any field. As part of the survey, participants selected which method(s) they used to generate their forecasts. Next, they elaborated on how they generated their forecasts in an open-ended question. There are no restrictions, though all teams were encouraged to report their education, as well as areas of knowledge or expertise. Participants were recruited via large scale advertising on social media, mailing lists in the behavioral and social sciences, decision sciences, and data science, advertisement on academic social networks including ResearchGate, and through word of mouth. To ensure broad representation across the academic spectrum of relevant disciplines, we targeted groups of scientists working on computational modeling, social psychology, judgment and decision-making, and data science to join the Forecasting Collaborative.

The Forecasting Collaborative started by the end of April 2020, during which time the U.S. Institute for Health Metrics and Evaluation projected the initial peak of the COVID-19 pandemic in the US. The recruitment phase continued until mid-June 2020, to ensure at least 40 teams joined the initial tournament. We were able to recruit 86 teams for the initial 12-month tournament (M_{age} = 38.18; SD = 8.37; 73% of forecasts made by scientists with a Doctorate degree), each of which provided forecasts for at least one domain (M = 4.17; SD = 3.78). At the six-month mark after 2020 US Presidential Election, we provided the initial participants with an opportunity to update their forecasts (44% provided updates), while simultaneously opening the tournament to new participants. This strategy allowed us to compare new forecasts against the updated predictions of the original participants, resulting in 120 teams for this follow-up six-month tournament (M_{age} = 36.82; SD = 8.30; 67% of forecasts made by scientists with a Doctorate degree; $M_{\text{forecasted domains}}$ = 4.55; SD = 3.88). Supplementary analyses showed that updating likelihood did not significantly differ when comparing data-free and data-inclusive models, z = 0.50, P = .618.

Procedure. Information for this project was available on the designated website (predictions.uwaterloo.ca), which included objectives, instructions, and prior monthly data for each of the 12 domains they can use for modeling. Researchers who decided to partake in the tournament signed up via a Qualtrics survey, which asked them to upload their estimates for forecasting domains of their choice in a pre-programmed Excel sheet that presented the historical trend and automatically juxtaposed their point estimate forecasts against the historical trend on a plot (see Appendix S1) and answer a set of questions about their rationale and forecasting team composition. Once all data was received, de-identified responses were used to pre-register the forecasted values and models on the Open Science Framework (https://osf.io/6wgbj/).

At the half-way point (i.e., at six months), participants were provided with a comparison summary of their initial point estimates forecasts vs. actual data for the initial six months. Subsequently, they were provided with an option to update their forecasts, provide a detailed description of the updates, and answer an identical set of questions about their data model and rationale for their forecasts, as well as the consideration of possible exogenous variables and counterfactuals.

Materials

Forecasting Domains and Data Pre-Processing. Computational forecasting models require enough prior time series data for reliable modeling. Based on prior recommendations [45], in the first tournament we provided each team with 39 monthly estimates—from January 2017 to March 2020—for each of the domains participating teams chose to forecast. This approach enabled the teams to perform data-driven forecasting (should teams choose to do so) and to establish a baseline estimate prior to the U.S. peak of the pandemic. In the second tournament, conducted six months later, we provided the forecasting teams with 45 monthly timepoints—from January 2017 to September 2020.

Because of the requirement for rich standardized data for computational approaches to forecasting [9], we limited forecasting domains to issues of broad societal significance. Our domain selection was guided by the discussion of broad social consequences associated with these issues at the beginning of the pandemic [46,47], along with general theorizing about psychological and social effects of threats of infectious disease [48,49]. Additional pragmatic consideration concerning the availability of large-scale longitudinal monthly time series data for a given issue. The resulting domains include affective well-being and life satisfaction, political ideology and polarization, bias in explicit and implicit attitudes towards Asian Americans and African Americans, as well as stereotypes regarding gender and career vs. family. To establish the "common task framework"—a necessary step for the evaluation of predictions in data sciences [9,17], we standardized methods for obtaining relevant prior data for each of these domains, made the data publicly available, recruited competitor teams for a common task of inferring predictions from the data, and a priori announced how the project leaders will evaluate accuracy at the end of the tournament.

Further, each team had to 1) download and inspect the historical trends (visualized on an Excel plot; example in Appendix); 2) add their forecasts in the same document, which automatically visualized their forecasts against the historical trends; 3) confirm their forecasts; 4) answer prompts concerning their forecasting rationale, their theoretical assumptions, models, conditionals, and consideration of additional parameters in the model. This procedure ensured all teams, at the minimum, considered historical trends, juxtaposed them against their forecasted time series, and deliberated on their forecasting assumptions.

Affective Well-being and Life Satisfaction. We used monthly Twitter data to estimate markers of affective well-being (positive and negative affect) and life satisfaction over time. We rely on Twitter because no polling data for monthly well-being over the required time period exists, and because prior work suggests that national estimates obtained via social media language can reliably track subjective well-being [50]. For each month, we used previously validated predictive models of well-being, as measured by affective well-being and life satisfaction scales [51]. Affective well-being was calculated by applying a custom lexicon [52] to message unigrams. Life satisfaction was estimated using a ridge regression model trained on *latent Dirichlet allocation* topics, selected using univariate feature selection and dimensionally reduced using randomized principal component analysis, to predict Cantril ladder life satisfaction scores. Such twitter-based estimates closely follow nationally representative polls [53]. We applied the respective models to Twitter data from January 2017 to March 2020 to obtain estimates of affective well-being and life satisfaction via language on social media.

Ideological Preferences. We approximated monthly ideological preferences via aggregated weighted data from the Congressional Generic Ballot polls conducted between January 2017 and March 2020 (projects.fivethirtyeight.com/congress-generic-ballot-polls), which ask representative samples of Americans to indicate which party they would support in an election. We weighed polls based on FiveThirtyEight pollster ratings, poll sample size, and poll frequency. FiveThirtyEight pollster ratings are determined by their historical accuracy in forecasting elections since 1998, participation in professional initiatives that seek to increase disclosure and enforce industry best practices and inclusion of live-caller surveys to cellphones and landlines. Based on this data, we then estimated monthly averages for support of Democrat and Republican parties across pollsters (e.g., Marist College, NBC/Wall Street Journal, CNN, YouGov/Economist).

Political Polarization. We assessed political polarization by examining differences in presidential approval ratings by party identification from Gallup polls (https://news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx). We obtained a difference score in % of Republican versus Democrat approval ratings and estimated monthly averages for the time period of interest. The absolute value of the difference score will ensure possible changes following the 2020 Presidential election will not change the direction of the estimate.

Explicit and Implicit Bias. Given the natural history of the COVID-19 pandemic, we sought to examine forecasted bias in attitudes towards Asian American (vs. European-Americans). To further probe racial bias, we sought to examine forecasted racial bias in preferences for African American (versus European-American) people. Finally, we sought to examine gender bias in associations of female (vs. male) gender with family versus career. For each domain we sought to obtain both estimates of explicit attitudes [54] and estimates of implicit attitudes [55]. To this end, we obtained data from the Project Implicit website (http://implicit.harvard.edu), which has collected continuous data concerning explicit stereotypes and implicit associations from a heterogeneous pool of volunteers (50,000 - 60,000 unique tests on each of these categories per month). Further details about the website and test materials are publicly available at https://osf.io/t4bnj. Recent work suggests that Project Implicit data can provide reliable societal estimates of consequential outcomes [56,57] and when studying cross-temporal societal shifts in U.S. attitudes [58]. Despite the non-representative nature of the Project Implicit data, recent analyses suggest that bias scores captured by Project Implicit are highly correlated with nationally representative estimates of explicit bias, r = .75, indicating that group aggregates of the bias data from Project Implicit can reliably approximate group-level estimates [57]. To further correct possible non-representativeness, we applied stratified weighting to the estimates, as described below.

Implicit attitude scores were computed using the revised scoring algorithm of the implicit association test (IAT) [59]. The IAT is a computerized task comparing reaction times to categorize paired concepts (in this case, social groups, e.g., Asian American vs. European American) and attributes (in this case, valence categories, e.g., good vs. bad). Average response latencies in correct categorizations were compared across two paired blocks in which participants categorized concepts and attributes with the same response keys. Faster responses in the paired blocks are assumed to reflect a stronger association between those paired concepts and attributes. Implicit gender-career bias was measured using the IAT with category labels of "male" and "female" and attributes of "career" / "family"). In all tests, positive IAT *D* scores indicate a relative preference for the typically preferred group (European-Americans) or association (men-career).

Respondents whose scores fell outside of the conditions specified in the scoring algorithm did not have a complete IAT *D* score and were therefore excluded from analyses. Restricting the analyses to only complete IAT *D* scores resulted in an average retention of 92% of the complete sessions across tests. The sample was further restricted to include only respondents from the United States to increase shared cultural understanding of attitude categories. The sample was restricted to include only respondents with complete demographic information on age, gender, race/ethnicity, and political ideology.

For explicit attitude scores, participants provided ratings on feeling thermometers towards Asian-Americans and European Americans (to assess Asian-American bias), and White and Black Americans (to assess racial bias), on a 7-point scale ranging from -3 to +3. Explicit gender-career bias was measured using 7-point Likert-type scales assessing the degree to which an attribute was female/male, from strongly female (-3) to strongly male (+3). Two questions assessed explicit stereotypes for each attribute (e.g., career with female/male, and, separately, the association of family). To match the explicit bias scores with the relative nature of the IAT, relative explicit stereotype scores were created by subtracting the "incongruent" association from the "congruent" association (e.g., [male vs. female-career] - [male vs. female-family]). Thus, for racial bias, -6 reflects a strong explicit preference for the minority over the majority (European-American) group, and +6 reflects a strong explicit preference for the majority over the minority (Asian American / African American) group. Similarly, for gender-career bias, counter-stereotype association (e.g., male-arts/female-science), and +6 reflects a strong stereotypic association (e.g., female-arts/male-science). In both cases, the midpoint of 0 represented equal liking of both groups.

We used explicit and implicit bias data for January 2017 – March 2020 and created monthly estimates for each of the explicit and implicit bias domains. Because of possible selection bias among the Project Implicit participants, we adjusted population estimates by weighting the monthly scores based on their representativeness of the demographic frequencies in the U.S. population (age, race, gender, education; estimated biannually by the U.S. Census Bureau; https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html). Further, we adjusted weights based on political orientation (1 = "strongly conservative;" 2 = "moderately conservative;" 3 = "slightly conservative;" 4 = "neutral;" 5 = "slightly liberal;" 6 = "moderately liberal;" 7 = "strongly liberal"), using corresponding annual estimates from the General Social Survey. With the weighting values for each participant, we computed weighted monthly means for each attitude test. These procedures ensured that weighted monthly averages approximated the demographics in the U.S. population. We cross-validated this procedure by comparing weighted annual scores to nationally representative estimates for feeling thermometer for African American and Asian American estimates from the American National Election studies in 2017 and 2018.

An initial procedure was developed for computing post-stratification weights for African American, Asian and gender career bias (implicit and explicit) to ensure that the sample was representative of the US population at large as much as possible. Originally, we computed weights for the entire year, which were then applied to each month in the year. After receiving feedback from co-authors, a more optimal approach was adopted wherein weights were computed on monthly, as opposed to yearly basis. This was necessary as demographic characteristics varied from month to month each year. This meant that using yearly weights had the potential of amplifying as opposed to reducing bias. Consequently, our new procedure ensured that sample representativeness was maximized. This insight affected forecasts from seven teams who provided them before the change. The teams were informed, and four teams chose to provide updated estimates using newly weighted historical data.

For each of these domains, forecasters were provided with 39 monthly estimates in the initial tournament (45 estimates in the follow-up tournament), as well as detailed explanation about the origin and the calculation of respective indices. Thereby, we aim to standardize the data source for the purpose of the forecasting competition [9]. See Supplementary Appendix S1 for example worksheets provided to participants for submissions of their forecasts.

Forecasting Justifications. For each forecasting model submitted to the tournament, participants provided detailed descriptions. They described the type of model they computed (e.g., time series, game theoretic models, other algorithms), model parameters, additional variables they included in their predictions (e.g., COVID-19 trajectory, presidential election outcome), and underlying assumptions.

Confidence in forecast. Participants rated their confidence in their forecasted points for each forecast model they submitted on a 7-point scale from 1 (not at all) to 7 (extremely).

Confidence in expertise. Participants provided ratings of their teams' expertise for a particular domain by indicating extent of agreement with the statement "my team has strong expertise on the research topic of [field]," on a 7-point scale from 1 (Strongly Disagree) to 7 (Strongly Agree).

COVID-19 Conditional. We considered the COVID-19 pandemic as a conditional of interest given links between infectious disease and the target social issues selected for this tournament. In Tournament 1, participants reported if they used the past or predicted trajectory of the COVID-19 pandemic (as measured by number of deaths or prevalence of cases or new infections) as a conditional in their model, and if so, provided their forecasted estimates for the COVID-19 variable included in their model.

Counterfactuals. Counterfactuals are hypothetical alternative historic events that would be thought to affect the forecast outcomes if they were to occur. Participants described the key counterfactual events between December 2019 and April 2020 that they theorized would have led to different forecasts (e.g., U.S.-wide implementation of social distancing practices in February). Two independent coders evaluated the distinctiveness of counterfactuals (interrater κ = .80). When discrepancies arose, they discussed individual cases with other members of the forecasting collaborative to make the final evaluation. In primary analyses, we focus on the presence of counterfactuals (yes/no).

Team Expertise. Because expertise can mean many things [2,60], we used a telescopic approach and operationalized expertise in four ways of varying granularity. First, we examined broad, domain-general expertise in social sciences by comparing social scientists' forecasts to forecasts provided by the general public without the same training in social science theory and methods. Second, we operationalized the prevalence of graduate training on a team as a more specific marker of domain-general expertise in social sciences. To this end, we asked each participating team to report how many team members have a doctorate degree in social sciences and calculated the percentage of doctorates on a team. Moving to domain-specific expertise, we instructed participating teams to report if any of their members had previously researched or published on the topic of their forecasted variable, operationalizing domain-specific expertise through this measure. Finally, moving to the most subjective level, we asked each participating team to report their subjective confidence in teams' expertise in a given domain (see Supplementary Information)

General Public Benchmark. In parallel to the tournament with 86 teams, on June 2-3, 2020, we recruited a regionally, gender- and socio-economically stratified sample of US residents via the Prolific crowdworker platform (targeted $N = 1{,}050$ completed responses) and randomly assigned them to forecast societal change for a subset of domains used in the tournaments (a. wellbeing: life satisfaction, positive and negative sentiment on social media; b. politics: political polarization, ideological support for Democrats and Republicans; c. Asian American Bias: explicit and implicit trends; d. African American Bias: explicit and implicit trends; e. Gender-career Bias: explicit and implicit trends). During recruitment, participants were informed that in exchange for 3.65 GDP they have to be able to open and upload forecasts in an Excel worksheet.

We considered responses if they provided forecasts for 12 months in at least one domain and if predictions did not exceed the possible range for a given domain (e.g., polarization above 100%). Moreover, three coders (intercoder $\kappa = .70$ unweighted, $\kappa = .77$ weighted) reviewed all submitted rationales from lay people and excluded any submissions where participants either misunderstood the task or wrote bogus bot-like responses. Coder disagreements were resolved via a discussion. Finally, we excluded responses if participants spent under 50s making their forecasts, which included reading instructions, downloading the files, providing forecasts, and re-uploading their forecasts (final $N = 802$, 1,467 forecasts; *M*age = 30.39, *SD* = 10.56, 46.36% female; education: 8.57% high school/GED, 28.80% some college, 62.63% college or above; ethnicity: 59.52% white, 17.10% Asian American, 9.45% African American/Black, 7.43% Latinx, 6.50% mixed/other; *Md* annual income = \$50,000-\$75,000; residential area: 32.37% urban, 57.03% suburban, 10.60% rural).

Exclusions of the General Public Sample. Supplementary Table 7 outlines exclusions by category. In the initial step, we considered all submissions via the Qualtrics platform, including partial submissions without any forecasting data ($N = 1{,}891$). Upon removing incomplete responses without forecasting data, and removing duplicate submissions from the same Prolific IDs, we removed 59 outliers whose data exceeded the range of possible values in a given domain. Subsequently, we removed responses independent coders flagged as either misunderstood ($n = 6$) or bot-like bogus responses ($n = 26$). See Supplementary Appendix S2 for verbatim examples of each screening category and exact coding instructions. Finally, we removed responses where participants took less than 50 seconds to provide their forecasts (including reading instructions, downloading the Excel file, filling it out, re-uploading the Excel worksheet, and completing additional information on their reasoning about the forecast). Finally, one response was removed based on open-ended information where the participant indicated they made forecasts for a different country than the US.

Naïve Statistical Benchmarks. There is evidence from data science forecasting competitions that the dominant statistical benchmarks are the Theta method, ARIMA, and ETS [7]. Given the socio-cultural context of our study, to avoid loss of generality we decided to employ more traditional benchmarks like naïve/Random walk, historical average, as well as the basic linear regression model—i.e., a method that is used more than anything else in practice and science. In short, we selected three benchmarks based on their common application in the forecasting literature (historical mean and random walk are most basic forecasting benchmarks) or the behavioral / social science literature (linear regression is the most basic statistical approach to test inferences in sciences). Furthermore, these benchmarks target distinct features of performance (historical mean speaks to the base rate sensitivity, linear regression speaks to sensitivity to the overall trend, whereas random walk captures random fluctuations and sensitivity to dependencies across consecutive time points). Each of these benchmarks may perform better in some but not in other circumstances. Consequently, to test the limits in scientists' performance, we examine if social scientists' performance is better than each of the three benchmarks. To obtain metrics of uncertainty around the naïve statistical estimates, we chose to simulate these three naïve approaches for making forecasts: 1) random resampling of historical data; 2) a naïve out-of-sample random walk based on random resampling of historical *change*; 3) extrapolation from a naïve regression based on a randomly selected interval of historical data. We describe each approach in the Supplement.

Analytic Plan

Categorization of Forecasts. We categorized forecasts based on modeling approaches. Two independent research associates categorize forecasts for each domain based on provided justifications: i. purely based on (a) theoretical model(s); ii. purely based on data-driven model(s); iii. a combination of theoretical and data-driven models – i.e., computational model relies on specific theoretical assumptions. See Appendix S3 for exact coding instructions and description of the classification (interrater $\kappa = .81$ unweighted, $\kappa = .90$ weighted). We further examined modelling complexity of approaches that relied on the extrapolation of time series from the data we provided (e.g., ARIMA, moving average with lags; yes/no; see Appendix S4 for exact coding instructions). Disagreements between coders here (interrater $\kappa = .80$ unweighted, $\kappa = .87$ weighted) and each coding task below were resolved through joint discussion with the leading author of the project.

Categorization of Additional Variables. We tested how the presence and number of additional variables as parameters in the model impact forecasting accuracy. To this end, we ensured that additional variables are distinct from one another. Two independent coders evaluated the distinctiveness of each reported parameter (interrater $\kappa = .56$ unweighted, $\kappa = .83$ weighted).

Categorization of Teams. We next categorized teams based on compositions. First, we counted the number of team members per team. We also sorted teams based on disciplinary orientation, comparing behavioral and social scientists to teams from computer and data science. Finally, we used information teams provided concerning their objective and subjective expertise level for a given subject domain.

Forecasting Update Justifications. Given that participants received both new data and a summary of diverse theoretical positions they can use as a basis for their updates, two independent research associates scored participants' justifications for forecasting updates on three dummy-categories: i. new six months of data we provided; ii. new theoretical insights; iii. consideration of other external events (interrater $\kappa = .63$ unweighted/weighted). See Appendix S5 for exact coding instructions.

Statistical analyses. A priori (https://osf.io/6wgbj/) we specified a linear mixed model as a key analytical procedure, with MASE scores for different domains nested in participating teams as repeated measures. Prior to analyses, we inspected MASE scores to determine violations of linearity, which we corrected via log-transformation prior to performing analyses. All *P* values refer to two-sided *t*-tests. For simple effects by domain, we applied Benjamini-Hochberg false discovery rate corrections. For 95% confidence intervals by domain, we simulated a multivariate *t* distribution[20] to adjust scores for simultaneous inference of estimates for 12 domains in each tournament.

Data availability

All data used in the main text and supplementary analysis is accessible on GitHub (https://github.com/grossmania/Forecasting-Tournament). All prior data presented to forecasters are available on https://predictions.uwaterloo.ca/. Historical and ground truth markers are obtained from Project Five Thirty Eight (https://projects.fivethirtyeight.com/polls/generic-ballot), Gallup (https://news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx), Project Implicit (see Open Science Framework website https://osf.io/t4bnj), and U.S. Census Bureau (https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html).

Code availability

Our project page at https://github.com/grossmania/Forecasting-Tournament displays all code from this paper. See reporting summary for *R* packages and their versions.

Acknowledgments

This program of research was supported by Basic Research Program at the National Research University Higher School of Economics (M. Fabrykant), John Templeton Foundation grant 62260 (I.G. and P.T.), Kega 079UK-4/2021 (P.K.), National Center for Complementary & Integrative Health of the National Institutes of Health under Award Number K23AT010879 (Simon B. Goldberg), National Science Foundation RAPID Grant 2026854 (M.E.W.V.), PID2019-111512RB-I00 (M.S.), NPO Systemic Risk Institute (LX22NPO5101) (I.R.), Slovak Research and Development Agency under contract no. APVV-20-0319 (M.A.), Social Sciences and Humanities Research Council of Canada Insight Grant 435-2014-0685 (I.G.), Social Sciences and Humanities Research Council of Canada Connection Grant 611-2020-0190 (I.G.), Swiss National Science Foundation grant PP00P1_170463 (O. Strijbis). Funders played no role in the conceptualization, design, analysis, or decision to publish this research. We thank Jordan Axt for providing monthly estimates of Project Implicit data, and the members of the Forecasting Collaborative who chose to remain anonymous for their contribution to the tournaments.

Author Contributions Statement

Conceptualization: I.G., A.R., C.A.H., M.E.W.V., L.T., and P.E.T.
Data curation: I.G., K.S., G.T.S., and O.J.T.
Forecasting: S.A., M.K.D., X.E.G., M.J. Hirshberg, M.K.-Y., D.R.M., L.R., A.V., L.W., M.A., A.A., P.A., K.B., G.B., F.B., E.B., C.B., M.B., C.K.B., D.T.B., E.M.C., R.C., B.-T.C., W.J.C., C.W.C., L.G.C., M. Davis, M.V.D., N.A.D., J.D.D., M. Dziekan, C.T.E., S.E., M. Fabrykant, M. Firat, G.T.F., J.A.F., J.M.G., S.B.G., A.G., J.G., L.G.-V., S.D.G., S.H., A.H., M.J. Hornsey, P.D.L.H., A.I., B.J., P.K., Y.J.K., R.K., D.G.L., H.-W.L., N.M.L., V.Y.Q.L., A.W.L., A.L.L., C.R.M., M. Maier, N.M.M., D.S.M., A.A.M., M. Misiak, K.O.R.M., J.M.N., J.N., K.N., J.O., T.O., M.P.-C., S.P., J.P., Q.R., I.R., R.M.R., Y.R., E.R., L.S., A.S., M.S., A.T.S., O. Simonsson,

M.-C.S., C.-C.T., T.T., B.A.T., D.T., D.C.K.T., J.M.T., L.U., D.V., L.V.W., H.A.V., Q.W., K.W., M.E.W., C.E.W., T.Y., K.Y., S.Y., V.r.A., J.R.A.-H., P.A.B., A.B., L.C., M.C., S.D.-H., Z.E.F., C.R.K., S.T.K., A.L.O., L.M., M.S.M., M.F.R.C.M., E.K.M., P.M., J.B.N., W.N., R.B.R., P.S., A.H.S., O. Strijbis, D.S., E.T., A.v.L., J.G.V., M.N.A.W., and T.W.

Formal analysis: I.G. and C.A.H.

Funding acquisition: I.G.

Investigation: I.G., A.R., and C.A.H.

Methodology: I.G., A.R., C.A.H., K.S., M.E.W.V., S.A., D.R.M., L.R., L.T., A.V., R.N.C., L.U., and D.V.

Project administration: I.G., A.R., M.E.W.V., M.K.-Y., and O.J.T.

Resources: I.G., A.R., J.N., and G.T.S.

Supervision: I.G.

Validation: K.S., X.E.G., and L.W.

Visualization: I.G. and M.K.D.

Writing - original draft: I.G.

Writing - review & editing: I.G., A.R., C.A.H., K.S., M.E.W.V., S.A., M.K.D., X.E.G., M.J. Hirshberg, M.K.-Y., D.R.M., L.R., L.T., A.V., L.W., M.A., A.A., P.A., K.B., G.B., F.B., E.B., C.B., M.B., C.K.B., D.T.B., E.M.C., R.C., B.-T.C., W.J.C., R.N.C., C.W.C., L.G.C., M. Davis, M.V.D., N.A.D., J.D.D., M. Dziekan, C.T.E., S.E., M. Fabrykant, M. Firat, G.T.F., J.A.F., J.M.G., S.B.G., A.G., J.G., L.G.-V., S.D.G., S.H., A.H., M.J. Hornsey, P.D.L.H., A.I., B.J., P.K., Y.J.K., R.K., D.G.L., H.-W.L., N.M.L., V.Y.Q.L., A.W.L., A.L.L., C.R.M., M. Maier, N.M.M., D.S.M., A.A.M., M. Misiak, K.O.R.M., J.M.N., K.N., J.O., T.O., M.P.-C., S.P., J.P., Q.R., I.R., R.M.R., Y.R., E.R., L.S., A.S., M.S., A.T.S., O. Simonsson, M.-C.S., C.-C.T., T.T., B.A.T., P.E.T., D.T., D.C.K.T., J.M.T., L.V.W., H.A.V., Q.W., K.W., M.E.W., C.E.W., T.Y., K.Y., and S.Y.

Competing interests Statement

Authors declare that they have no competing interests.

Tables

Table 1. Contrasts of Mean-level Inaccuracy (MASE) among Lay Crowds and Social Scientists

Domain	*t-ratio*	*df*	*p-value*	*Cohen's d* [95% *CI*]	*Bayes Factor*	*Interpretation*
Life Satisfaction	4.321	1725	< .001	0.93 [0.32;1.55]	22.72	Substantial ev. for difference
Explicit Gender-career Bias	3.204	1731	.006	0.90 [0.10; 1.71]	1.37	Some evidence for difference
Implicit Gender-career Bias	3.161	1747	.006	0.88 [0.09; 1.67]	2.49	Some evidence for difference
Political Polarization	2.819	1802	.015	0.71 [-0.01; 1.42]	0.77	Not enough evidence
Positive Affect	2.128	1796	.080	0.54 [-0.18; 1.26]	0.12	Substantial ev. for no difference
Exp. Asian American Bias	1.998	1789	.092	0.53 [-0.23; 1.29]	0.11	Substantial ev. for no difference
Ideology Republicans	1.650	1794	.170	0.40 [-0.29; 1.08]	0.06	Substantial ev. for no difference
Ideology Democrats	1.456	1795	.204	0.35 [-0.34; 1.04]	0.04	Substantial ev. for no difference
Imp. Asian American Bias	1.430	1802	.204	0.36 [-0.36; 1.09]	0.11	Substantial ev. for no difference
Exp. African American Bias	0.939	1747	.218	0.26 [-0.53; 1.05]	0.04	Substantial ev. for no difference
Imp. African American Bias	0.536	1780	.646	0.14 [-0.63; 0.91]	0.02	Substantial ev. for no difference
Negative Affect	-0.271	1796	.787	0.07 [-0.79; 0.65]	0.02	Substantial ev. for no difference

Note. Scores > 1: greater accuracy of scientific forecasts. Scores < 1: greater accuracy of lay crowds. Pairwise contrasts were obtained via *emmeans* package in *R* [21], drawing on the restricted information maximum likelihood model with group (scientist or naïve crowd), domain, and their interaction as predictors of the *log*(MASE) scores, with responses nested in participants. To avoid skew, tests are performed on log-transformed scores. Degrees of freedom were obtained via Kenward-Roger approximation. *P*-values are adjusted for false discovery rate. *CI* = Confidence intervals of effect size *(Cohen's d*), which are adjusted for simultaneous inference of 12 domains by simulating a multivariate *t* distribution[20]. For Bayesian analyses we relied on weakly informative priors for our linear mixed model (see Supplement for more detail). Interpretation of Bayes factor is in the right column. Bayes factors greater than 3 are interpreted as substantial evidence of a difference, values between 3 and 1 suggest some evidence of a difference, values between ⅓ and 1 indicate that there is not enough evidence to interpret, and values < ⅓ indicate substantial evidence in favor of the null hypothesis (no difference between groups).

Figure Legends/Captions

Figure 1: Social scientists' average forecasting errors, compared against different benchmarks. We rank domains from least to most error in Tournament 1, assessing forecasting errors via mean absolute scaled error (MASE). Estimated means for Scientists and Naïve Crowd indicate the fixed effect coefficients of a linear mixed model with domain (k = 12) and group (in Tournament 1: $n_{\text{scientists}} = 86$, $n_{\text{naïve crowd}} = 802$; only scientists in Tournament 2: $n = 120$) as a predictor of forecasting error (MASE) scores nested in teams (Tournament 1 observations: $n_{\text{scientists}} = 359$, $n_{\text{naïve crowd}} = 1467$; Tournament 2 observations: $n = 546$), using restricted maximum likelihood estimation. To correct for right skew, we used log-transformed MASE scores, which are subsequently back-transformed when calculating estimated means and 95% confidence intervals. In each tournament, confidence intervals are adjusted for simultaneous inference of estimates for 12 domains in each tournament by simulating a multivariate *t* distribution[20]. Benchmarks represent the naïve crowd and best performing naïve statistical benchmark (either historic mean, average random walk with an autoregressive lag of one, or linear regression). Statistical benchmarks were obtained via simulations (k = 10,000) with resampling (see Supplement). Scores to the left of the dotted vertical line show better performance than naïve in-sample random walk. Scores to the left of the dashed vertical line show better performance than median performance in M4 tournaments [7].

Figure 2. Forecasts and ground truth – are forecasts anchoring on the last few historical data points? Historical time series (40 months before the Tournament 1) and ground truth series (12 months over the Tournament 1), along with forecasts of individual teams (light blue), lowess curve and 95% confidence interval across social scientists' forecasts (blue), and lowess curve and 95% confidence interval across naïve crowd's forecasts (salmon). For most domains, Tournament 1 forecasts of both scientists and the naïve crowd start near the last few historical data points they received prior to the tournament (January – March 2020). Note, April 2020 forecast was not provided to the participants.

Figure 3. Ratios of Benchmarks against Scientific Forecasts. Scores >1: greater accuracy of scientific forecasts. Scores <1: greater accuracy of naïve benchmarks. Domains are ranked from least to most error among scientific teams in Tournament 1. Estimated means indicate the fixed effect coefficient of a linear mixed model with domain (k = 12) in each tournament ($n_{\text{Tournament 1}} = 86$; $n_{\text{Tournament 2}} = 120$) as a predictor of benchmark-specific ratio scores nested in teams

(observations: $n_{\text{Tournament 1}} = 359$, $n_{\text{Tournament 2}} = 546$), using restricted maximum likelihood estimation. To correct for right skew, we used square-root or log-transformed MASE scores, which were subsequently back-transformed when calculating estimated means and 95% confidence intervals. Confidence intervals are adjusted for simultaneous inference of estimates for 12 domains in each tournament by simulating a multivariate *t* distribution[20].

Figure 4: Slope graph showing consistency in the ranking of domains in terms of estimated mean forecasting error across all teams in each tournament, assessed via mean absolute scaled error, from most to least inaccurate forecasts across both tournaments. Solid line = change in accuracy between tournaments is statistically significant ($P < .05$); dashed line = non-significant change. Significance is determined via pairwise comparisons of *log*(MASE)scores for each domain, drawing on the restricted information maximum likelihood model with Tournament (first or second), domain, and their interaction as predictors of the *log*(MASE) scores, with responses nested in scientific teams ($N_{\text{teams}} = 120$, $N_{\text{observations}} = 905$).

Figure 5: Forecasting errors by prediction approach. Estimated means and 95% confidence intervals are based on a restricted information maximum likelihood linear mixed effects model with model type (data-driven, hybrid or intuition/theory-based) as a fixed effects predictor of the *log*(MASE) scores, domain as a fixed effects covariate, and responses nested in participants. We ran separate models for each tournament (first: $N_{\text{groups}} = 86$; $N_{\text{observations}} = 359$; second: $N_{\text{groups}} = 120$; $N_{\text{observations}} = 546$). Scores below the dotted vertical line show better performance than naïve in-sample random walk. Scores below the dashed vertical line show better performance than median performance in M4 tournaments [7].

Figure 6: Contribution of specific forecasting strategies (*n* parameters, statistical model complexity, consideration of exogenous events and counterfactuals) and team characteristics for forecasting accuracy (reversed MASE scores), ranked in terms of magnitude. Scores to the right of the dashed vertical line contribute positively to accuracy, whereas estimates to the left of the dashed vertical line contribute negatively. Analyses control for domain type. All continuous predictors are mean-centered and scaled by 2 standard deviations, to afford comparability [24]. The reported standard errors are heteroskedasticity-robust. Thicker bands show 90% confidence interval, whereas thinner lines show at 95% confidence interval. Effects are statistically significant if the 95% confidence interval does not include zero (dashed vertical line).

References

1. Hutcherson, C. *et al.* On the accuracy, media representation, and public perception of psychological scientists' judgments of societal change. *American Psychologist* (in press).
2. Collins, H. & Evans, R. *Rethinking Expertise*. (University of Chicago Press, 2009).
3. Fama, E. F. Efficient Capital Markets: A Review of Theory and Empirical Work. *J. Finance* **25**, 383 (1970).
4. Tetlock, P. E. *Expert Political Judgement: How Good Is It?* (2005).
5. Hofman, J. M. *et al.* Integrating explanation and prediction in computational social science. *Nature* **595**, 181–188 (2021).
6. Mandel, D. R. & Barnes, A. Accuracy of forecasts in strategic intelligence. *Proc. Natl. Acad. Sci.* **111**, 10984–10989 (2014).
7. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. *Int. J. Forecast.* **36**, 54–74 (2020).
8. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science.* **349**, aac4716–aac4716 (2015).
9. Hofman, J. M., Sharma, A. & Watts, D. J. Prediction and explanation in social systems. *Science.* **355**, 486–488 (2017).
10. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).
11. Fincher, C. L. & Thornhill, R. Parasite-stress promotes in-group assortative sociality: The cases of strong family ties and heightened religiosity. *Behav. Brain Sci.* **35**, 61–79 (2012).
12. Varnum, M. E. W. & Grossmann, I. Pathogen prevalence is associated with cultural changes in gender equality. *Nat. Hum. Behav.* **1**, 3 (2016).
13. Schaller, M. & Murray, D. R. Pathogens, personality, and culture: Disease prevalence predicts worldwide variability in sociosexuality, extraversion, and openness to experience. *J. Pers. Soc. Psychol.* **95**, 212–221 (2008).
14. van Leeuwen, F., Park, J. H., Koenig, B. L. & Graham, J. Regional variation in pathogen prevalence predicts endorsement of group-focused moral concerns. *Evol. Hum. Behav.* **33**, 429–437 (2012).
15. Hawkley, L. C. & Cacioppo, J. T. Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms. *Ann. Behav. Med.* **40**, 218–227 (2010).
16. Salganik, M. J. *et al.* Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl. Acad. Sci.* **117**, 8398–8403 (2020).
17. Liberman, M. Reproducible Research and the Common Task Method. (2015).
18. Hyndman, R. J. & Koehler, A. B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **22**, 679–688 (2006).
19. Eyal, P., David, R., Andrew, G., Zak, E. & Ekaterina, D. Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* 1–20 (2021).

doi:10.3758/s13428-021-01694-3

20. Genz, A. & Bretz, F. *Computation of Multivariate Normal and t Probabilities*. (Springer-Verlag, 2009).

21. Lenth, R., Singmann, H., Love, J. & Maxime, H. emmeans: Estimated Marginal Means, aka Least-Squares Means. (2020).

22. Green, K. C. & Armstrong, J. S. Simple versus complex forecasting: The evidence. *J. Bus. Res.* **68**, 1678–1685 (2015).

23. Grossmann, I., Twardus, O., Varnum, M. E. W., Jayawickreme, E. & McLevey, J. Expert predictions of societal change: Insights from the world after COVID project. *American Psychologist* **77**, 276–290 (2022).

24. Gelman, A. Scaling regression inputs by dividing by two standard deviations. *Stat. Med.* **27**, 2865–2873 (2008).

25. Grossmann, I., Huynh, A. C. & Ellsworth, P. C. Emotional complexity: Clarifying definitions and cultural correlates. *J. Pers. Soc. Psychol.* **111**, 895–916 (2016).

26. Alves, H., Koch, A. & Unkelbach, C. Why Good Is More Alike Than Bad: Processing Implications. *Trends Cogn. Sci.* **21**, 69–79 (2017).

27. Dimant, E. *et al.* Politicizing mask-wearing: predicting the success of behavioral interventions among republicans and democrats in the U.S. *Sci. Rep.* **12**, 7575 (2022).

28. Dunning, D., Heath, C. & Suls, J. M. Flawed Self-Assessment. *Psychol. Sci. Public Interes.* **5**, 69–106 (2004).

29. Mellers, B., Tetlock, P. E. & Arkes, H. R. Forecasting tournaments, epistemic humility and attitude depolarization. *Cognition* **188**, 19–26 (2019).

30. Klein, R. A. *et al.* Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).

31. Voslinsky, A. & Azar, O. H. Incentives in experimental economics. *J. Behav. Exp. Econ.* **93**, 101706 (2021).

32. Cerasoli, C. P., Nicklin, J. M. & Ford, M. T. Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychol. Bull.* **140**, 980–1008 (2014).

33. Richard, F. D., Bond Jr., C. F. & Stokes-Zoota, J. J. One Hundred Years of Social Psychology Quantitatively Described. *Rev. Gen. Psychol.* **7**, 331–363 (2003).

34. Cesario, J. What Can Experimental Studies of Bias Tell Us About Real-World Group Disparities? *Behav. Brain Sci.* 1–80 (2021). doi:10.1017/S0140525X21000017

35. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).

36. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2022).

37. IJzerman, H. *et al.* Use caution when applying behavioural science to policy. *Nat. Hum. Behav.* **4**, 1092–1094 (2020).

38. Makridakis, S. & Taleb, N. Living in a world of low levels of predictability. *Int. J. Forecast.* **25**, 840–844 (2009).

39. Varnum, M. E. W. & Grossmann, I. Cultural Change: The How and the Why. *Perspect. Psychol. Sci.* **12**, 956–972 (2017).

40. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16**, (2001).

41. Lewin, K. Defining the 'field at a given time.' *Psychol. Rev.* **50**, 292–310 (1943).

42. Turchin, P., Currie, T. E., Turner, E. A. L. & Gavrilets, S. War, space, and the evolution of Old World complex societies. *Proc. Natl. Acad. Sci.* **110**, 16384–16389 (2013).

43. Brockwell, P. J. & Davis, R. A. *Introduction to Time Series and Forecasting*. (Springer International Publishing, 2016). doi:10.1007/978-3-319-29854-2

44. Hitchens, N. M., Brooks, H. E. & Kay, M. P. Objective Limits on Forecasting Skill of Rare Events. *Weather Forecast.* **28**, 525–534 (2013).

45. Jebb, A. T., Tay, L., Wang, W. & Huang, Q. Time series analysis for psychological research: examining and forecasting change. *Front. Psychol.* **6**, (2015).

46. Van Bavel, J. *et al.* Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **4**, 460–471 (2020).

47. Seitz, B. M. *et al.* The pandemic exposes human nature: 10 evolutionary insights. *Proc. Natl. Acad. Sci.* **117**, 27767–27776 (2020).

48. Schaller, M. & Park, J. H. The behavioral immune system (and why it matters). *Curr. Dir. Psychol. Sci.* (2011). doi:10.1177/0963721411402596

49. Wang, I. M., Michalak, N. M. & Ackerman, J. M. Threat of Infectious Disease. in *The SAGE Handbook of Personality and Individual Differences: Volume II: Origins of Personality and Individual Differences* (eds. Zeigler-Hill, V., Shackelford, T. K., Zeigler-Hill, V. & Shackelford, T. K.) 321–345 (2018). doi:10.4135/9781526451200.n18

50. Luhmann, M. Using Big Data to study subjective well-being. *Curr. Opin. Behav. Sci.* **18**, 28–33 (2017).

51. Schwartz, H. A. *et al.* Predicting individual well-being through the language of social media. in *Biocomputing 2016* 516–527 (2016). doi:10.1142/9789814749411_0047

52. Kiritchenko, S., Zhu, X. & Mohammad, S. M. Sentiment analysis of short informal texts. *J. Artif. Intell. Res.* **50**, 723–762 (2014).

53. Witters, D. & Harter, J. *In U.S., Life Ratings Plummet to 12-Year Low*. (2020).

54. Axt, J. R. The Best Way to Measure Explicit Racial Attitudes Is to Ask About Them. *Soc. Psychol. Personal. Sci.* **9**, 896–906 (2018).

55. Nosek, B. A. *et al.* Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur. Rev. Soc. Psychol.* **18**, 36–88 (2007).

56. Hehman, E., Flake, J. K. & Calanchini, J. Disproportionate Use of Lethal Force in Policing Is Associated with Regional Racial Biases of Residents. *Soc. Psychol. Personal. Sci.* **9**, 393–401 (2018).

57. Ofosu, E. K., Chambers, M. K., Chen, J. M. & Hehman, E. Same-sex marriage legalization associated with reduced implicit and explicit antigay bias. *Proc. Natl. Acad. Sci.* **116**, 8846–8851 (2019).

58. Charlesworth, T. E. S. & Banaji, M. R. Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability From 2007 to 2016. *Psychol. Sci.* **30**, 174–192 (2019).

59. Greenwald, A. G., Nosek, B. A. & Banaji, M. R. Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *J. Pers. Soc. Psychol.* **85**, 197–216 (2003).

60. Gobet, F. The Future of Expertise: The Need for a Multidisciplinary Approach. **1**, 7 (2018).

The Forecasting Collaborative

Igor Grossmann[1*], Amanda Rotella[2,1], Cendri A. Hutcherson[3], Konstantyn Sharpinskyi[1], Michael E. W. Varnum[4], Sebastian Achter[5], Mandeep K. Dhami[6], Xinqi Evie Guo[7], Mane Kara-Yakoubian[8], David R. Mandel[9,10], Louis Raes[11], Louis Tay[12], Aymeric Vie[13,14], Lisa Wagner[15], Matus Adamkovic[16,17], Arash Arami[18], Patrícia Arriaga[19], Kasun Bandara[20], Gabriel Baník[16], František Bartoš[21], Ernest Baskin[22], Christoph Bergmeir[23], Michał Białek[24], Caroline K. Børsting[25], Dillon T. Browne[1], Eugene M. Caruso[26], Rong Chen[27], Bin-Tzong Chie[28], William J. Chopik[29], Robert N. Collins[9], Chin W. Cong[30], Lucian G. Conway[31], Matthew Davis[32], Martin V. Day[33], Nathan A. Dhaliwal[34], Justin D. Durham[35], Martyna Dziekan[36], Christian T. Elbaek[25], Eric Shuman[37], Marharyta Fabrykant[38,39], Mustafa Firat[40], Geoffrey T. Fong[1,41], Jeremy A. Frimer[42], Jonathan M. Gallegos[43], Simon B. Goldberg[44], Anton Gollwitzer[45,46], Julia Goyal[47], Lorenz Graf-Vlachy[48,49], Scott D. Gronlund[35], Sebastian Hafenbrädl[50], Andree Hartanto[51], Matthew J. Hirshberg[52], Matthew J. Hornsey[53], Piers D. L. Howe[54], Anoosha Izadi[55], Bastian Jaeger[56], Pavol Kačmár[57], Yeun Joon Kim[58], Ruslan Krenzler[59,60], Daniel G. Lannin[61], Hung-Wen Lin[62], Nigel Mantou Lou[63,64], Verity Y. Q. Lua[51], Aaron W. Lukaszewski[65,66], Albert L. Ly[67], Christopher R. Madan[68], Maximilian Maier[69], Nadyanna M. Majeed[70], David S. March[71], Abigail A. Marsh[72], Michal Misiak[24,73], Kristian Ove R. Myrseth[74], Jaime M. Napan[67], Jonathan Nicholas[75], Konstantinos Nikolopoulos[76], Jiaqing O[77], Tobias Otterbring[78,79], Mariola Paruzel-Czachura[80,81], Shiva Pauer[21], John Protzko[82], Quentin Raffaelli[83], Ivan Ropovik[84,85], Robert M. Ross[86], Yefim Roth[87], Espen Røysamb[88], Landon Schnabel[89], Astrid Schütz[90], Matthias Seifert[91], A. T. Sevincer[92], Garrick T. Sherman[93], Otto Simonsson[94,95], Ming-Chien Sung[96], Chung-Ching Tai[96], Thomas Talhelm[97], Bethany A. Teachman[98], Philip E. Tetlock[99,100], Dimitrios Thomakos[101], Dwight C. K. Tse[102], Oliver J. Twardus[103], Joshua M. Tybur[56], Lyle Ungar[93], Daan Vandermeulen[104], Leighton Vaughan Williams[105], Hrag A. Vosgerichian[106], Qi Wang[107], Ke Wang[108], Mark E. Whiting[109,110], Conny E. Wollbrant[111], Tao Yang[112], Kumar Yogeeswaran[113], Sangsuk Yoon[114], Ventura r. Alves[115], Jessica R. Andrews-Hanna[83,116], Paul A. Bloom[75], Anthony Boyles[117], Loo Charis[118], Mingyeong Choi[119], Sean Darling-Hammond[120], Zoe E. Ferguson[121], Cheryl R. Kaiser[43], Simon T. Karg[122], Alberto López Ortega[123], Lori Mahoney[124], Melvin S. Marsh[125], Marcellin F. R. C. Martinie[54], Eli K. Michaels[126], Philip Millroth[127], Jeanean B. Naqvi[128], Weiting Ng[129], Robb B. Rutledge[130], Peter Slattery[131], Adam H. Smiley[43], Oliver Strijbis[132], Daniel Sznycer[133], Eli Tsukayama[134], Austin van Loon[135], Jan G. Voelkel[135], Margaux N. A. Wienk[75], Tom Wilkening[136]

[1]Department of Psychology, University of Waterloo; Waterloo, Canada.

[2]Department of Psychology, Northumbria University, UK.

[3]Department of Psychology, University of Toronto Scarborough; Toronto, Canada.

[4]Department of Psychology, Arizona State University, Tempe, USA.

[5]Institute of Management Accounting and Simulation, Hamburg University of Technology; Hamburg, Germany.

[6]Department of Psychology, Middlesex University London; London, UK.

[7]Department of Experimental Psychology, University of California San Diego; San Diego, USA.

[8]Department of Psychology, Toronto Metropolitan University; Toronto, Canada.

[9]Defence Research and Development Canada; Ottawa, Canada.

[10]Department of Psychology, York University; Toronto, Canada.

[11]Department of Economics, Tilburg University, Tilburg, Netherlands.

[12]Department of Psychological Sciences, Purdue University; West Lafayette, USA.

[13]Mathematical Institute, University of Oxford; Oxford, UK.

[14]Institute of New Economic Thinking, University of Oxford, Oxford, UK.

[15]Jacobs Center for Productive Youth Development, University of Zurich; Zurich, Switzerland.

[16]Institute of Psychology, University of Prešov; Prešov, Slovakia.

[17]Institute of Social Sciences, CSPS, Slovak Academy of Sciences; Bratislava, Slovakia.

[18]Department of Mechanical and Mechatronics Engineering, University of Waterloo; Waterloo, Canada.

[19]Iscte-University Institute of Lisbon, CIS; Lisbon, Portugal.

[20]Melbourne Centre for Data Science, University of Melbourne; Melbourne, Australia.

[21]Faculty of Social and Behavioural Sciences, University of Amsterdam; Amsterdam, Netherlands.

[22]Department of Food Marketing, Haub School of Business, Saint Joseph's University; Philadelphia, USA.

[23]Department of Data Science and Artificial Intelligence, Monash University; Melbourne, Australia.

[24]Institute of Psychology, University of Wrocław; Wrocław, Poland.

[25]Department of Management, Aarhus University; Aarhus, Denmark.

[26]Anderson School of Management, UCLA; Los Angeles, USA.

[27]Department of Psychology, Dominican University of California; San Rafael, USA.

[28]Department of Industrial Economics, Tamkang University; New Taipei City, Taiwan.

[29]Department of Psychology, Michigan State University; East Lansing, USA.

[30]Department of Psychology and Counselling, Universiti Tunku Abdul Rahman; Kampar, Malaysia.

[31]Psychology Department, Grove City College; Grove City, USA.

[32]Department of Economics, Siena College; Loudonville, USA.

[33]Department of Psychology, Memorial University of Newfoundland; St. John's, Canada.

[34]UBC Sauder School of Business, University of British Columbia; Vancouver, Canada.

[35]Department of Psychology, University of Oklahoma; Norman, USA.

[36]Faculty of Psychology and Cognitive Science, Adam Mickiewicz University; Poznań, Poland.

[37]Department of Psychology, University of Groningen; Groningen, Netherlands.

[38]Laboratory for Comparative Studies in Mass Consciousness, Expert Institute, HSE University; Moscow, Russia.

[39]Faculty of Philosophy and Social Sciences, Belarusian State University; Minsk, Belarus.

[40]Department of Sociology, Radboud University; Nijmegen, Netherlands.

[41]Ontario Institute for Cancer Research; Toronto, Canada.

[42]Department of Psychology, University of Winnipeg; Winnipeg, Canada.

[43]Department of Psychology, University of Washington; Seattle, USA.

[44]Department of Counseling Psychology, University of Wisconsin - Madison; Madison, USA.

[45]Department of Leadership and Organizational Behaviour, BI Norwegian Business School; Oslo, Norway.

[46]Center for Adaptive Rationality, Max Planck Institute for Human Development; Berlin, Germany.

[47]School of Public Health Sciences, University of Waterloo; Waterloo, Canada.

[48]TU Dortmund University; Dortmund, Germany.

[49]ESCP Business School; Berlin, Germany.

[50]IESE Business School; Barcelona, Spain.

[51]School of Social Sciences, Singapore Management University; Singapore.

[52]Center for Healthy Minds, University of Wisconsin-Madison; Madison, USA.

[53]University of Queensland Business School; Brisbane, Australia.

[54]Melbourne School of Psychological Sciences, University of Melbourne; Melbourne, Australia.

[55]Department of Marketing, University of Massachusetts Dartmouth; Dartmouth, USA.

[56]Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam; Amsterdam, Netherlands.

[57]Department of Psychology, Faculty of Arts, Pavol Jozef Šafárik University; Košice, Slovakia.

[58]Cambridge Judge Business School, University of Cambridge; Cambridge, UK.

[59]Hermes Germany GmbH; Hamburg, Germany.

[60]University of Hamburg; Hamburg, Germany.

[61]Department of Psychology, Illinois State University; Normal, USA.

[62]Department of Marketing and Logistics Management, National Penghu University of Science and Technology; Magong City, Taiwan.

[63]Department of Psychology, University of Victoria; Victoria, Canada.

[64]Centre for Youth and Society, University of Victoria; Victoria, Canada.

[65]Department of Psychology, California State University - Fullerton; Fullerton, USA.

[66]Center for the Study of Human Nature, California State University - Fullerton; Fullerton, USA.

[67]Department of Psychology, Loma Linda University; Loma Linda, USA.

[68]University of Nottingham; Nottingham, UK.

[69]Department of Experimental Psychology, University College London; London, UK.

[70]Singapore Management University; Singapore.

[71]Department of Psychology, Florida State University; Tallahassee, USA.

[72]Department of Psychology, Georgetown University; Washington DC, USA.

[73]School of Anthropology & Museum Ethnography, University of Oxford; Oxford, UK.

[74]School for Business and Society, University of York; York, UK.

[75]Department of Psychology, Columbia University; New York City, USA.

[76]IHRR Forecasting Laboratory, Durham University; Durham, UK.

[77]Department of Psychology, Aberystwyth University; Aberystwyth, UK.

[78]School of Business and Law, Department of Management, University of Agder; Kristiansand, Norway.

[79]Institute of Retail Economics; Stockholm, Sweden.

[80]Institute of Psychology, University of Silesia in Katowice, Poland.

[81] Department of Neurology, Penn Center for Neuroaesthetics, University of Pennsylvania, United States.

[82]Central Connecticut State University; New Britain, USA.

[83]Department of Psychology, University of Arizona; Tucson, USA.

[84]Faculty of Education, Institute for Research and Development of Education, Charles University; Prague, Czech Republic.

[85]Faculty of Education; University of Prešov; Prešov, Slovakia.

[86]School of Psychology, Macquarie University; Sydney, Australia.

[87]Department of Human Service, University of Haifa; Haifa, Israel.

[88]Promenta Center, Department of Psychology, University of Oslo; Oslo, Norway.

[89]Department of Sociology, Cornell University; Ithaca, USA.

[90]Institute of Psychology, University of Bamberg; Bamberg, Germany.

[91]IE Business School, IE University; Madrid, Spain.

[92]Faculty of Psychology and Human Movement Science, University of Hamburg; Hamburg, Germany.

[93]Department of Computer and Information Science, University of Pennsylvania; Philadelphia, USA.

[94]Department of Clinical Neuroscience, Karolinska Institutet; Solna, Sweden.

[95]Department of Sociology, University of Oxford; Oxford, UK.

[96]Department of Decision Analytics and Risk, University of Southampton; Southampton, UK.

[97]University of Chicago Booth School of Business; Chicago, USA.

[98]Department of Psychology, University of Virginia; Charlottesville, USA.

[99]Psychology Department, University of Pennsylvania; Philadelphia, USA.

[100]Wharton School of Business, University of Pennsylvania; Philadelphia, USA.

[101]Department of Economics, National and Kapodistrian University of Athens; Athens, Greece.

[102]School of Psychological Sciences and Health, University of Strathclyde; Glasgow, UK.

[103]Department of Psychology, University of Guelph; Guelph, Canada.

[104]Psychology Department, Hebrew University of Jerusalem; Jerusalem, Israel.

[105]Department of Economics, Nottingham Trent University; Nottingham, UK.

[106]Department of Management and Organizations, Northwestern University; Evanston, USA.

[107]College of Human Ecology, Cornell University, Ithaca, USA.

[108]Harvard Kennedy School, Harvard University; Cambridge, USA.

[109]Computer and Information Science, University of Pennsylvania; Philadelphia, USA.

[110]Operations, Information, and Decisions Department, The Wharton School, University of Pennsylvania; Philadelphia, USA.

[111]School of Economics and Finance, University of St. Andrews; St. Andrews, UK.

[112]Department of Management, Cameron School of Business, University of North Carolina; Wilmington, USA.

[113]School of Psychology, Speech and Hearing, University of Canterbury; Christchurch, New Zealand, [114]Deparment of Marketing, University of Dayton; Dayton, USA.

[115]ISG Universidade Lusofona; Lisbon, Portugal.

[116]Cognitive Science, University of Arizona; Tucson, USA.

[117]Ephemer AI; Atlanta, USA.

[118]Questrom School of Business, Boston University; Boston, USA.

[119]Institute of Social Science Research, Pusan National University; Busan, South Korea, [120]Fielding School of Public Health, UCLA; Los Angeles, USA.

[121]Psychology Department, University of Washington; Seattle, USA.

[122]Department of Political Science, Aarhus University; Aarhus, USA.

[123]Faculty of Social Sciences, Vrije Universiteit Amsterdam; Amsterdam, Netherlands.

[124]College of Science and Mathematics, Wright State University; Fairborn, USA.

[125]Department of Psychology, Georgia Southern University; Statesboro, USA.

[126]Division of Epidemiology, School of Public Health, University of California, Berkeley; Berkeley, USA.

[127]Department of Psychology, Uppsala University; Uppsala, Sweden.

[128]Department of Psychology, Carnegie Mellon University; Pittsburgh, USA.

[129]School of Humanities & Behavioral Sciences, Singapore University of Social Sciences; Singapore.

[130]Department of Psychology, Yale University; New Haven, USA.

[131]BehaviourWorks Australia. Monash University; Melbourne, Australia.

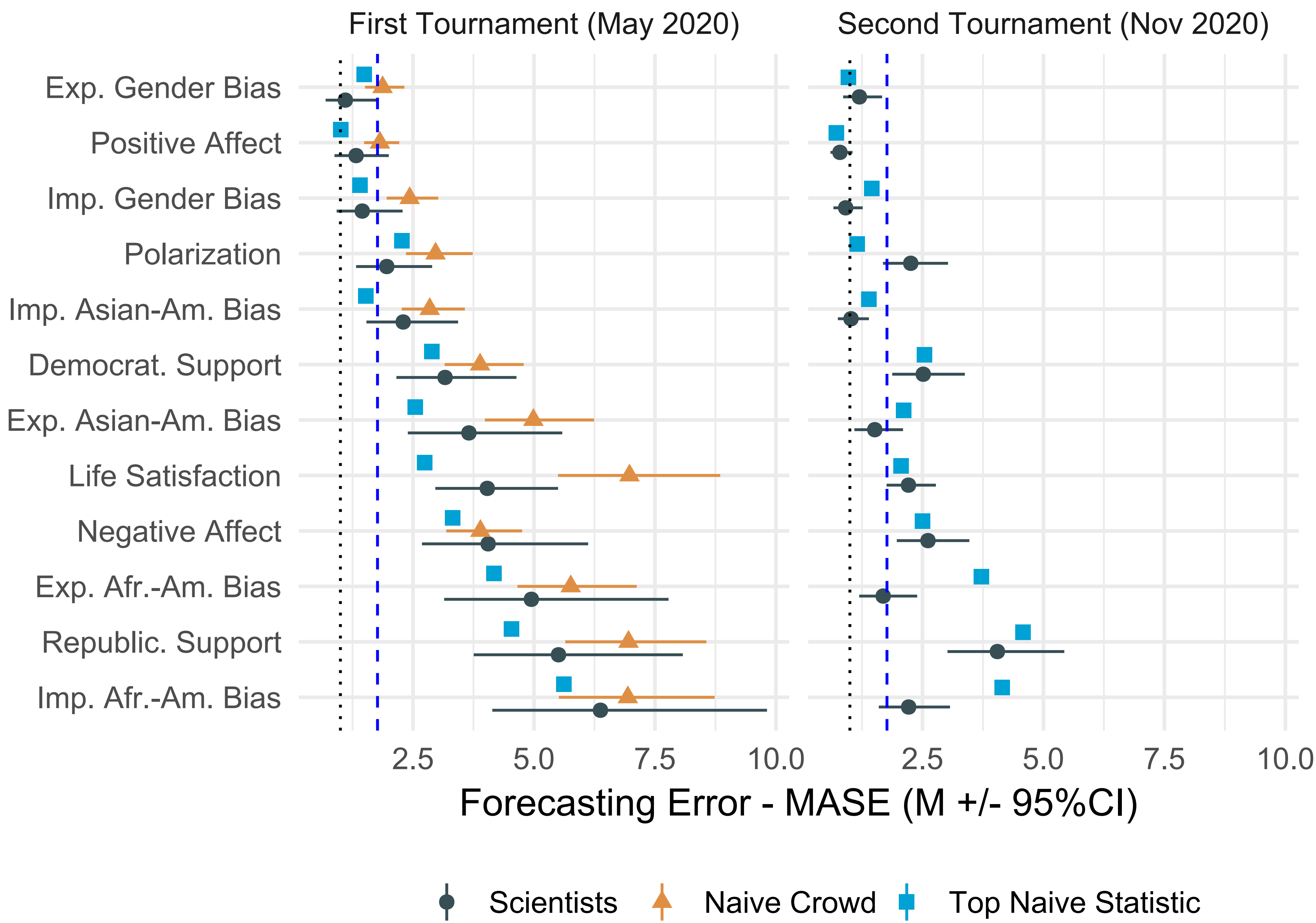
[132]Institute of Political Science, University of Zurich; Zurich, Switzerland.

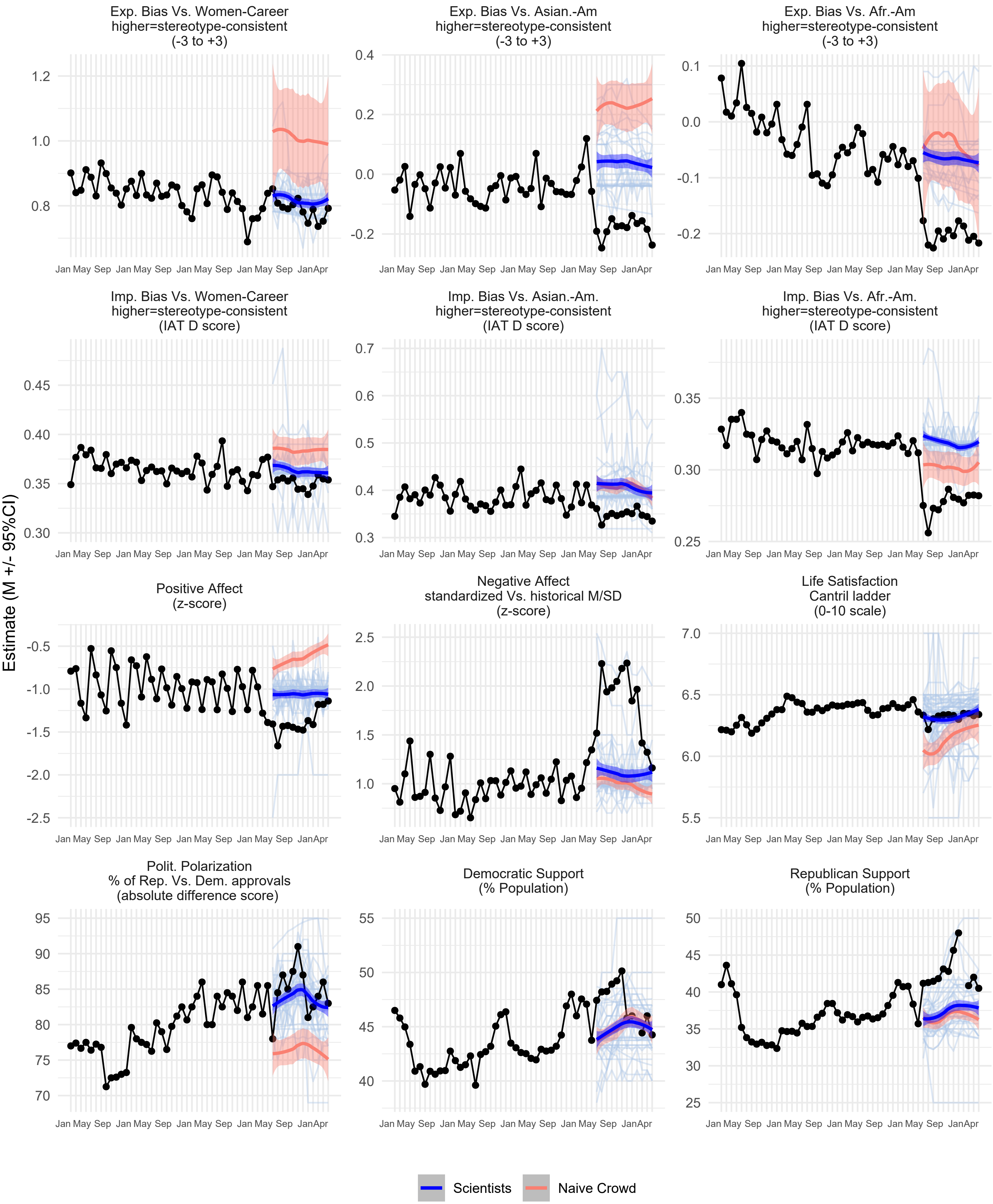
[133]Department of Psychology, Oklahoma State University; Stillwater, USA.

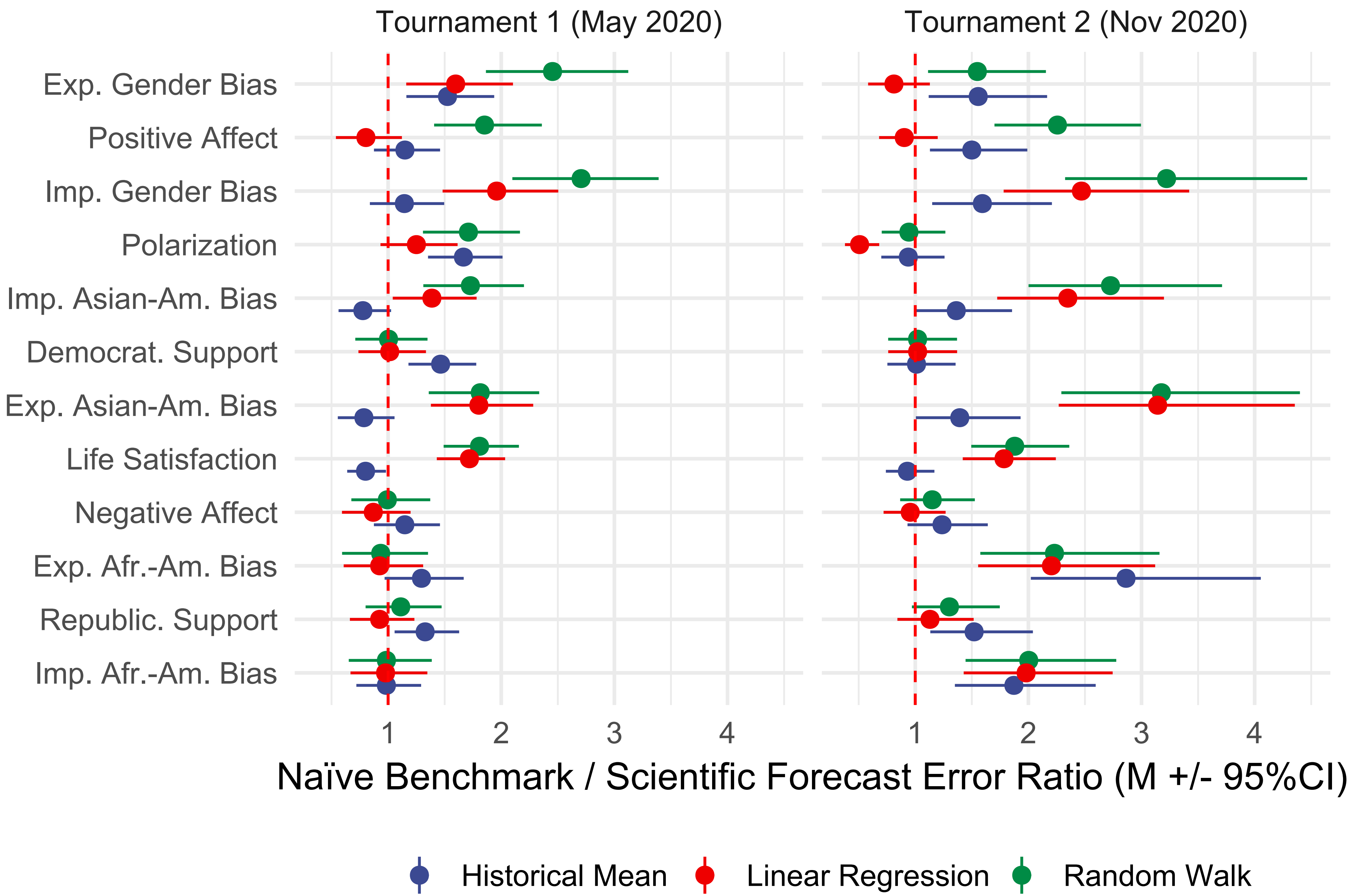
[134]Department of Business Administration, University of Hawaii - West Oahu; Kapolei, USA.

[135]Department of Sociology, Stanford University; Stanford, USA.

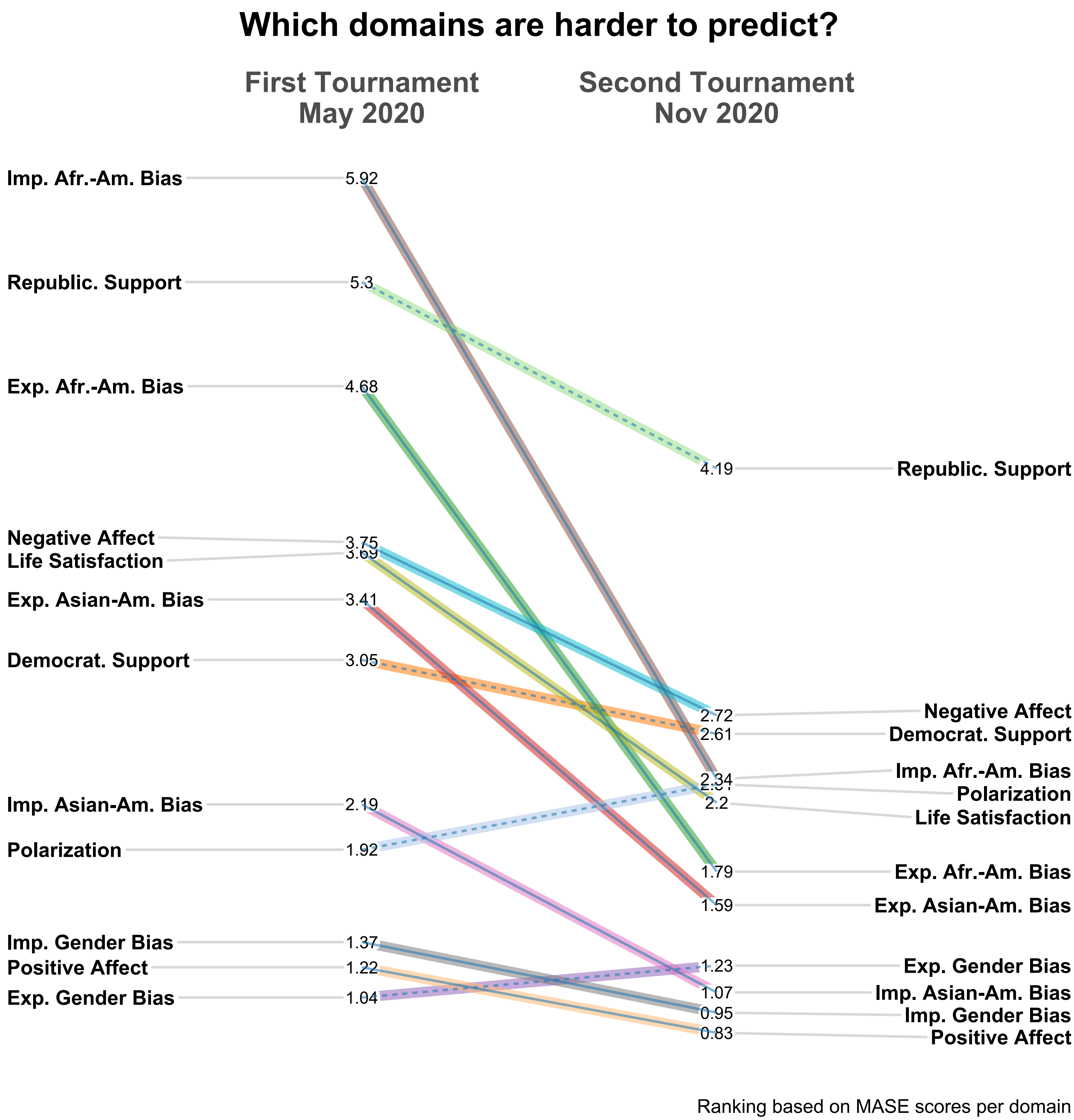
[136]Department of Economics, University of Melbourne; Melbourne, Australia.

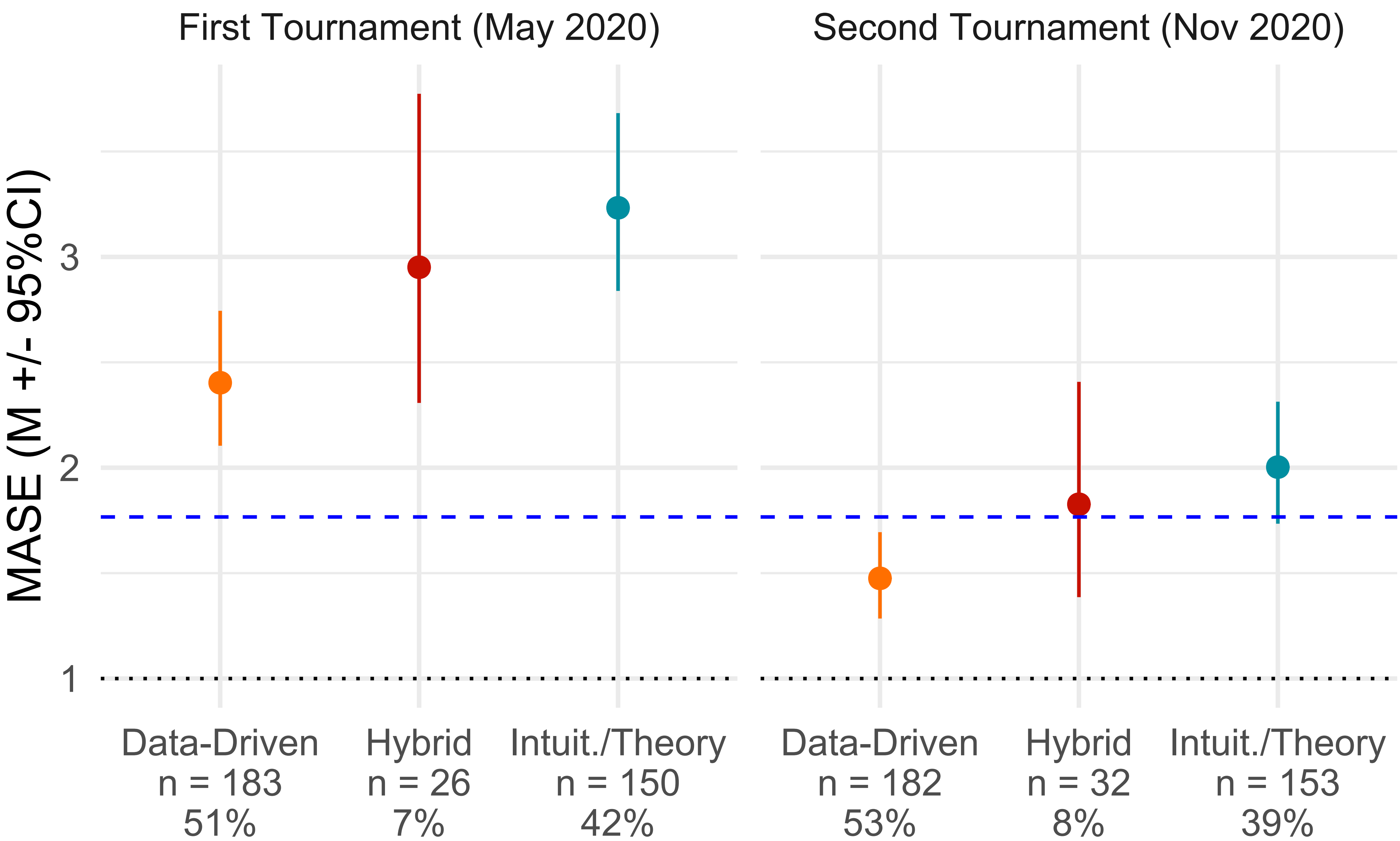

First Tournament (May 2020)
Second Tournament (Nov 2020)
Exp. Gender Bias
Positive Affect
Imp. Gender Bias
Polarization
Imp. Asian-Am. Bias
Democrat. Support
Exp. Asian-Am. Bias
Life Satisfaction
Negative Affect
Exp. Afr.-Am. Bias
Republic. Support
Imp. Afr.-Am. Bias
2.5
5.0
7.5
10.0
Forecasting Error - MASE (M +/- 95%CI)
Scientists
Naive Crowd
Top Naive Statistic

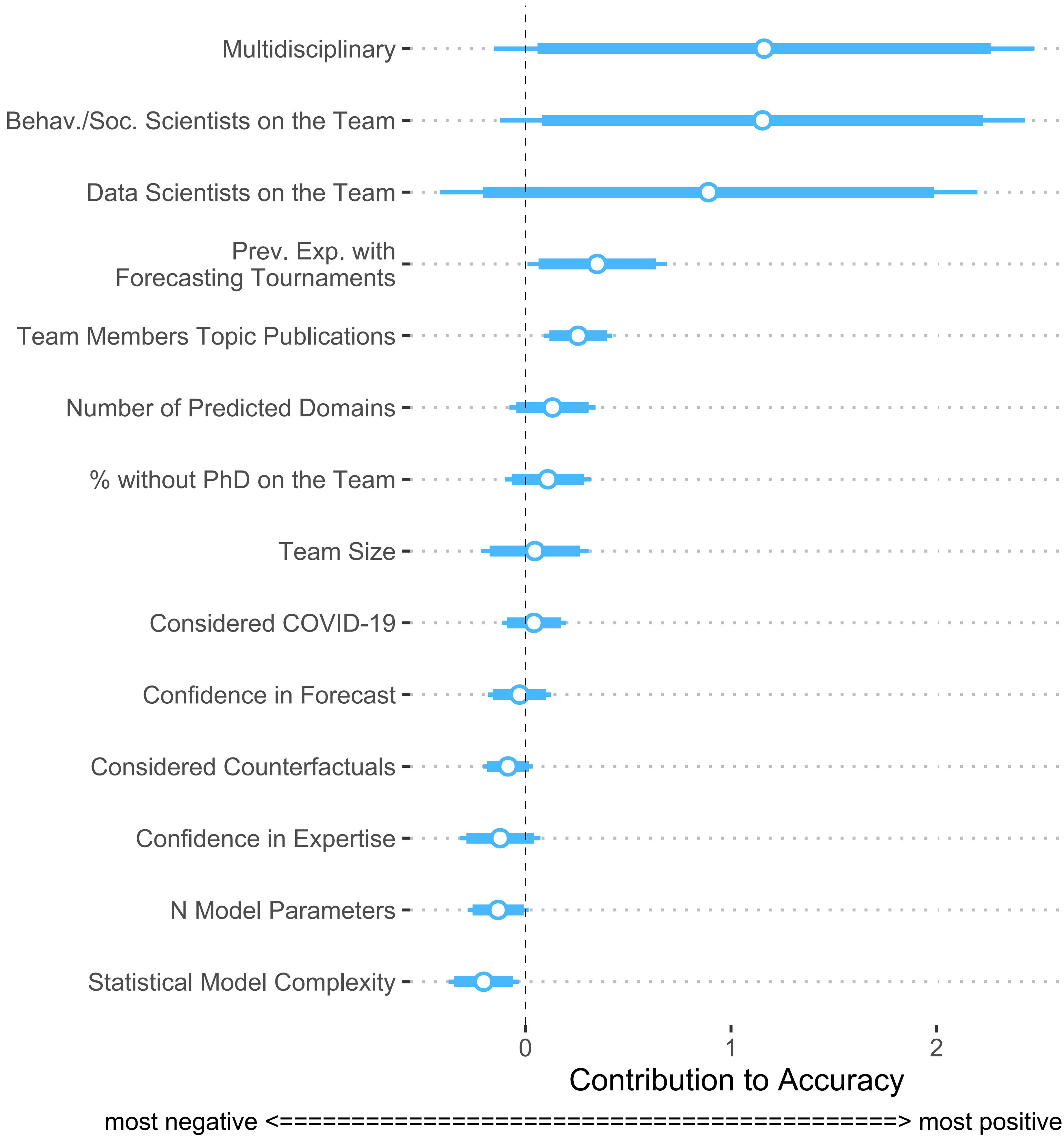

Exp. Bias Vs. Women-Career
higher=stereotype-consistent
(-3 to +3)
Exp. Bias Vs. Asian.-Am
higher=stereotype-consistent
(-3 to +3)
Exp. Bias Vs. Afr.-Am
higher=stereotype-consistent
(-3 to +3)
Imp. Bias Vs. Women-Career
higher=stereotype-consistent
(IAT D score)
Imp. Bias Vs. Asian.-Am.
higher=stereotype-consistent
(IAT D score)
Imp. Bias Vs. Afr.-Am.
higher=stereotype-consistent
(IAT D score)
Positive Affect
(z-score)
Negative Affect
standardized Vs. historical M/SD
(z-score)
Life Satisfaction
Cantril ladder
(0-10 scale)
Polit. Polarization
% of Rep. Vs. Dem. approvals
(absolute difference score)
Democratic Support
(% Population)
Republican Support
(% Population)
Estimate (M +/- 95%CI)
Jan May Sep Jan May Sep Jan May Sep Jan May Sep JanApr
Scientists
Naive Crowd



Tournament 1 (May 2020)
Tournament 2 (Nov 2020)
Exp. Gender Bias
Positive Affect
Imp. Gender Bias
Polarization
Imp. Asian-Am. Bias
Democrat. Support
Exp. Asian-Am. Bias
Life Satisfaction
Negative Affect
Exp. Afr.-Am. Bias
Republic. Support
Imp. Afr.-Am. Bias
1
2
3
4
Naïve Benchmark / Scientific Forecast Error Ratio (M +/- 95%CI)
Historical Mean
Linear Regression
Random Walk


Which domains are harder to predict?
First Tournament
May 2020
Second Tournament
Nov 2020
Imp. Afr.-Am. Bias
5.92
Republic. Support
5.3
Exp. Afr.-Am. Bias
4.68
Negative Affect
3.75
Life Satisfaction
3.69
Exp. Asian-Am. Bias
3.41
Democrat. Support
3.05
Imp. Asian-Am. Bias
2.19
Polarization
1.92
Imp. Gender Bias
1.37
Positive Affect
1.22
Exp. Gender Bias
1.04
4.19
Republic. Support
2.72
Negative Affect
2.61
Democrat. Support
2.34
Imp. Afr.-Am. Bias
Polarization
2.2
Life Satisfaction
1.79
Exp. Afr.-Am. Bias
1.59
Exp. Asian-Am. Bias
1.23
Exp. Gender Bias
1.07
Imp. Asian-Am. Bias
0.95
Imp. Gender Bias
0.83
Positive Affect
Ranking based on MASE scores per domain


First Tournament (May 2020)
Second Tournament (Nov 2020)
MASE (M +/- 95%CI)
3
2
1
Data-Driven
n = 183
51%
Hybrid
n = 26
7%
Intuit./Theory
n = 150
42%
Data-Driven
n = 182
53%
Hybrid
n = 32
8%
Intuit./Theory
n = 153
39%



Multidisciplinary
Behav./Soc. Scientists on the Team
Data Scientists on the Team
Prev. Exp. with Forecasting Tournaments
Team Members Topic Publications
Number of Predicted Domains
% without PhD on the Team
Team Size
Considered COVID-19
Confidence in Forecast
Considered Counterfactuals
Confidence in Expertise
N Model Parameters
Statistical Model Complexity
0
1
2
Contribution to Accuracy
most negative <==> most positive