

Insights into the accuracy of social scientists' forecasts of societal change

Received: 26 June 2022

Accepted: 19 December 2022

Published online: 09 February 2023

 Check for updates

The Forecasting Collaborative* 

How well can social scientists predict societal change, and what processes underlie their predictions? To answer these questions, we ran two forecasting tournaments testing the 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. After we provided them with historical trend data on the relevant domain, social scientists submitted pre-registered monthly forecasts for a year (Tournament 1; $N = 86$ teams and 359 forecasts), with an opportunity to update forecasts on the basis of new data six months later (Tournament 2; $N = 120$ teams and 546 forecasts). Benchmarking forecasting accuracy revealed that social scientists' forecasts were on average no more accurate than those of simple statistical models (historical means, random walks 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.

Can social scientists predict societal change? Governments and the general public often rely on experts, on the basis of a general belief that they make better judgements and predictions of the future in their domain of expertise. The media also seek out experts to render their judgements 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 predictions³. Similarly, in the domain of geopolitics, experts often perform at chance levels when forecasting occurrences of specific political events⁴. On the basis of 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. By examining sampled data, social scientists strive to develop theoretical models about causal mechanisms that, in ideal cases, reliably describe human behaviour and societal processes⁵. 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 than 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 (the 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 and so on). Post hoc, those making predictions have access to the temporal variance or occurrence for each of these variables and hence are more

*A list of authors and their affiliations appears at the end of the paper.  e-mail: igrossma@uwaterloo.ca

likely to be successful in ex post predictions. Predictions about past events thus 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 period. Binary predictions for the one-off event do not require accuracy in estimating the degree of change or the shape of the predicted time series, which are extra challenges in forecasting societal change.

The second reason is that past research on forecasting has concentrated on predicting geopolitical⁴ or economic events⁷ rather than broader societal phenomena. Thus, in contrast to systematic studies concerning the replicability of in-sample explanations of social science phenomena⁸, out-of-sample prediction accuracy in the social sciences remains understudied^{9,10}. Similarly, little is known about the rationales and approaches that social scientists use to make predictions for societal trends. For example, are social scientists more apt to rely on data-driven statistical methods or on theory and intuitions when generating such predictions?

To address these unknowns, we performed a standardized evaluation of forecasting accuracy⁹ among social scientists in well-studied domains for which systematic, cross-temporal data are 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 on the basis of 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 isolation¹⁵. To understand how scientists made predictions in these domains, we documented the rationales and processes they used to generate forecasts, and we 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 behavioural, 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 to April 2021) as part of the initial tournament and, upon receiving feedback on initial performance, again after 6 months for a follow-up tournament (the recruitment details are in the Methods, and the demographic information is 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 United States for each of the 12 variables related to the phenomena of interest (that is, 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).

The 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 the Methods for the 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. The 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 used data-driven methods, they also provided the model and any additional data used to generate their forecasts (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 (that is, the average forecasts of a sample of lay participants recruited from Prolific). Second, we compared social scientists' performance in both tournaments with naive random extrapolation algorithms (that is, 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 (<https://osf.io/7ekfm>; the details are 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. The MASE is an asymptotically normal, scale-independent scoring rule that compares predicted values against the 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 competitions⁷. See the Supplementary Information for further details of the MASE method.

In addition to absolute accuracy, we assessed the comparative accuracy of social scientists' forecasts using several benchmarks. First, during Tournament 1, we obtained forecasts from a non-expert crowdsourced sample of US residents ($N = 802$) via Prolific¹⁹ who received the same data as the 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 naive approaches to out-of-sample forecasting using the time series data provided to the tournament participants: (1) the historical mean, calculated by randomly resampling the historical time series data; (2) a naive random walk, calculated by randomly resampling historical change in the time series data with an autoregressive component; and (3) extrapolation from linear regression, based on a randomly selected interval of the historical time series data (see the Supplementary Information for the 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 naive in-sample predictions that form the basis of MASE scores).

How accurate were behavioural and social scientists at forecasting?

Figure 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 the raw estimates; Supplementary Fig. 2 reports measures of uncertainty around the estimates). In Tournament 2, the forecasts were on average inferior to in-sample random walks in eight domains and inferior to median performance in prior forecasting competitions in

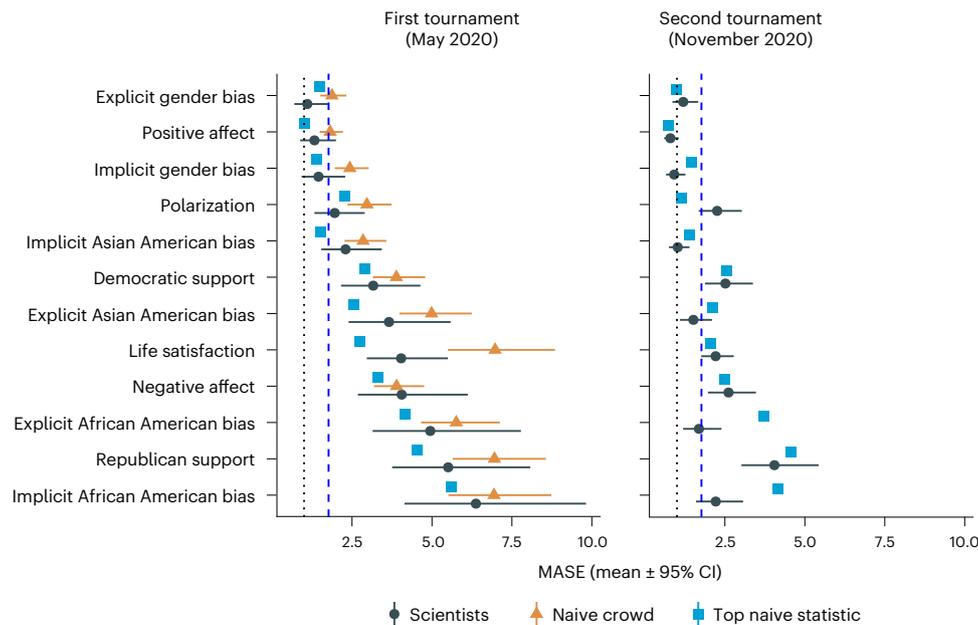


Fig. 1 | Social scientists' average forecasting errors, compared against different benchmarks. We ranked the domains from least to most error in Tournament 1, assessing forecasting errors via the MASE. The estimated means for the scientists and the naive crowd indicate the fixed-effect coefficients of a linear mixed model with domain ($k = 12$) and group (in Tournament 1: $N_{\text{scientists}} = 86$, $N_{\text{naive crowd}} = 802$; only scientists in Tournament 2: $N = 120$) as predictors of forecasting error (MASE) scores nested in teams (Tournament 1 observations: $N_{\text{scientists}} = 359$, $N_{\text{naive crowd}} = 1,467$; Tournament 2 observations: $N = 546$), using restricted maximum likelihood estimation. To correct for right skew, we used log-transformed MASE scores, which were subsequently back-

transformed when calculating estimated means and 95% CIs. In each tournament, the CIs were adjusted for simultaneous inference of estimates for 12 domains in each tournament by simulating a multivariate t distribution²⁰. The benchmarks represent the naive crowd and the best-performing naive statistical benchmark (historical mean, average random walk with an autoregressive lag of one or linear regression). Statistical benchmarks were obtained via simulations ($k = 10,000$) with resampling (Supplementary Information). Scores to the left of the dotted vertical line show better performance than a naive in-sample random walk. Scores to the left of the dashed vertical line show better performance than the median performance in M4 tournaments⁷.

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 and 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 the accuracy of scientific forecasts in the numerator 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 (Supplementary Information). Except for one team, the top forecasters from Tournament 1 did not appear among the winners of Tournament 2 (Supplementary Tables 1 and 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 the accuracy of social scientists and that of lay crowds ($F(11, 1,747) = 0.88$, $P = 0.348$, partial $R^2 < 0.001$). However, we observed a significant interaction between social science training and domain ($F(11, 1,304) = 2.00$, $P = 0.026$). Simple effects show that social scientists were significantly more accurate than lay people when forecasting life satisfaction, polarization, and explicit and implicit gender-career bias. However, the scientific teams were no better than the lay sample in the remaining eight domains (Fig. 1 and Table 1). Moreover, Bayesian analyses indicated that only for life satisfaction is there substantial evidence in favour of the difference, whereas for eight domains the evidence was in favour of the null hypothesis. See the

Supplementary Information for further details and the 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. Figure 2 allows us to inspect individual trends of social scientists and the naive 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 (CI) around the trend (blue). For the naive 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, and support for Democrats and Republicans—lowess curves from both groups were overlapping, suggesting that the estimates from both social scientists and the naive crowd were identical. Moreover, except for the domain of life satisfaction, the forecasts of scientists and the naive crowd were close to far off the mark vis-à-vis ground truth. In one further domain—explicit bias against African Americans—the naive crowd estimate was in fact closer to the ground truth marker than the estimate from the lowess curve of the social scientists. In the 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 those of the naive crowd. We note, however, that these visual inspections may be somewhat misleading because the CIs don't correct for multiple tests. This caveat

Table 1 | Contrasts of mean-level inaccuracy (MASE) among lay crowds and social scientists

Domain	t-ratio	d.f.	P	Cohen's d (95% CI)	Bayes factor	Interpretation
Life satisfaction	4.321	1,725	<0.001	0.93 (0.32;1.55)	22.72	Substantial evidence for difference
Explicit gender-career bias	3.204	1,731	0.006	0.90 (0.10; 1.71)	1.37	Some evidence for difference
Implicit gender-career bias	3.161	1,747	0.006	0.88 (0.09; 1.67)	2.49	Some evidence for difference
Political polarization	2.819	1,802	0.015	0.71 (-0.01; 1.42)	0.77	Not enough evidence
Positive affect	2.128	1,796	0.080	0.54 (-0.18; 1.26)	0.12	Substantial evidence for no difference
Explicit Asian American bias	1.998	1,789	0.092	0.53 (-0.23; 1.29)	0.11	Substantial evidence for no difference
Ideology Republicans	1.650	1,794	0.170	0.40 (-0.29; 1.08)	0.06	Substantial evidence for no difference
Ideology Democrats	1.456	1,795	0.204	0.35 (-0.34; 1.04)	0.04	Substantial evidence for no difference
Implicit Asian American bias	1.430	1,802	0.204	0.36 (-0.36; 1.09)	0.11	Substantial evidence for no difference
Explicit African American bias	0.939	1,747	0.218	0.26 (-0.53; 1.05)	0.04	Substantial evidence for no difference
Implicit African American bias	0.536	1,780	0.646	0.14 (-0.63; 0.91)	0.02	Substantial evidence for no difference
Negative affect	-0.271	1,796	0.787	0.07 (-0.79; 0.65)	0.02	Substantial evidence for no difference

Scores greater than 1 indicate greater accuracy of scientific forecasts. Scores less than 1 indicate greater accuracy of lay crowds. Pairwise contrasts were obtained via the emmeans package⁶² in R (version 4.2.2)⁶³, drawing on the restricted information maximum likelihood model with group (scientist or naive crowd), domain and their interaction as predictors of the log(MASE) scores, with responses nested in participants. To avoid skew, the tests were performed on log-transformed scores. Degrees of freedom were obtained via Kenward-Roger approximation. The *P* values are adjusted for false discovery rate. The CIs of effect size (Cohen's *d*) are adjusted for simultaneous inference of 12 domains by simulating a multivariate *t* distribution²⁰. For the Bayesian analyses, we relied on weakly informative priors for our linear mixed model (see the Supplementary Information for more details). The interpretation of the 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 1/3 and 1 indicate that there is not enough evidence to interpret, and values less than 1/3 indicate substantial evidence in favour of the null hypothesis (no difference between groups).

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 naive crowd's predictions.

Comparisons with naive statistical benchmarks

Next, we compared scientific forecasts against three naive 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, and ratios greater than 1 indicate 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 naive benchmarks by chance, our main analyses focused on performance across all three benchmarks (see the Supplementary Information for the rationale favouring this method over averaging across the three benchmarks), and we adjusted the CIs of the ratio scores for simultaneous inference of 12 domains in each tournament by simulating a multivariate *t* distribution²⁰. Figures 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 1 out of 12 domains, which concerned explicit gender-career bias ($1.53 < \text{ratio} \leq 1.60$, $1.16 < 95\% \text{ CI} \leq 2.910$). In the remaining 11 domains, scientific predictions were either no different from 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 naive benchmarks in 5 out of 12 domains. These domains reflected explicit racial bias (African American bias, $2.20 < \text{ratio} \leq 2.86$, $1.55 < 95\% \text{ CI} \leq 4.05$; Asian American bias, $1.39 < \text{ratio} \leq 3.14$, $1.01 < 95\% \text{ CI} \leq 4.40$) and implicit racial and gender-career biases (African American bias, $1.35 < \text{ratio} \leq 2.00$, $1.35 < 95\% \text{ CI} \leq 2.78$; Asian American bias, $1.36 < \text{ratio} \leq 2.73$, $1.001 < 95\% \text{ CI} \leq 3.71$; gender-career bias, $1.59 < \text{ratio} \leq 3.22$, $1.15 < 95\% \text{ CI} \leq 4.46$). In the remaining seven domains, the forecasts were not significantly different from the naive benchmarks. Moreover, as Fig. 3 shows, scientific forecasts for political polarization in Tournament 2 were significantly less accurate than estimates from a naive linear regression (ratio = 0.51; $95\% \text{ CI}$, (0.38, 0.68)). Figure 3 also shows that in most domains at least one of the naive forecasting methods produced errors that were

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

Which domains were harder to predict?

Figure 4 shows that some societal trends were significantly harder to forecast than others (Tournament 1: $F(11,295.69) = 41.88$, $P < 0.001$, $R^2 = 0.450$; Tournament 2: $F(11,469.49) = 26.87$, $P < 0.001$, $R^2 = 0.291$). Forecast accuracy was the lowest in politics (underestimating Democratic support, 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 the quality of ground truth markers: on the basis of the sampling frequency and representativeness of the data, most reliable ground truth markers concerned

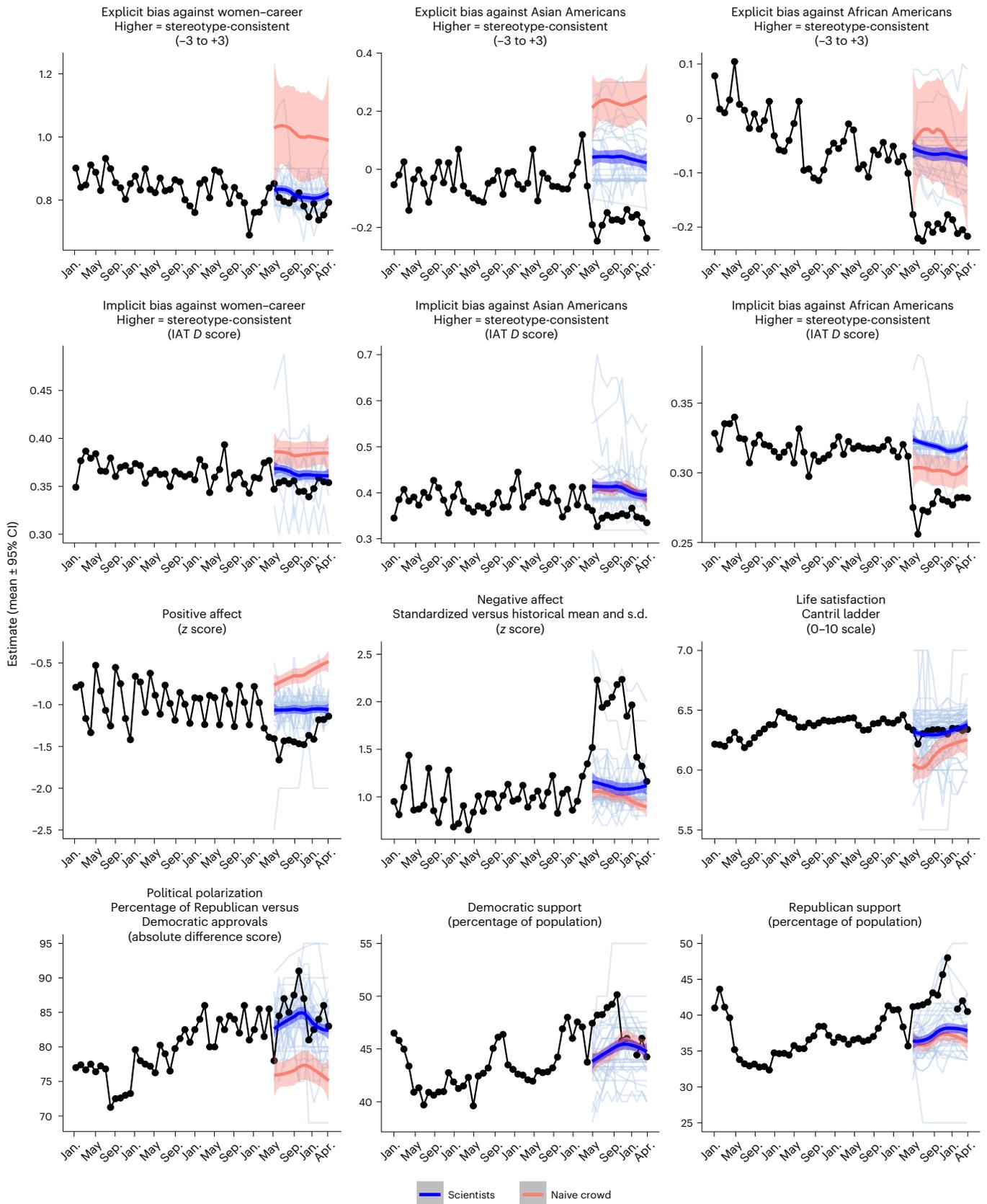


Fig. 2 | Forecasts and ground truth—are forecasts anchoring on the last few historical data points? Historical time series (40 months before Tournament 1) and ground truth series (12 months over Tournament 1), along with forecasts of individual teams (light blue), lowess curves and 95% CIs across social scientists' forecasts (blue), and lowess curves and 95% CIs across the naive crowd's forecasts

(salmon). For most domains, Tournament 1 forecasts of both scientists and the naive crowd start near the last few historical data points they received prior to the tournament (January–March 2020). Note that the April 2020 forecast was not provided to the participants. IAT, implicit association test.

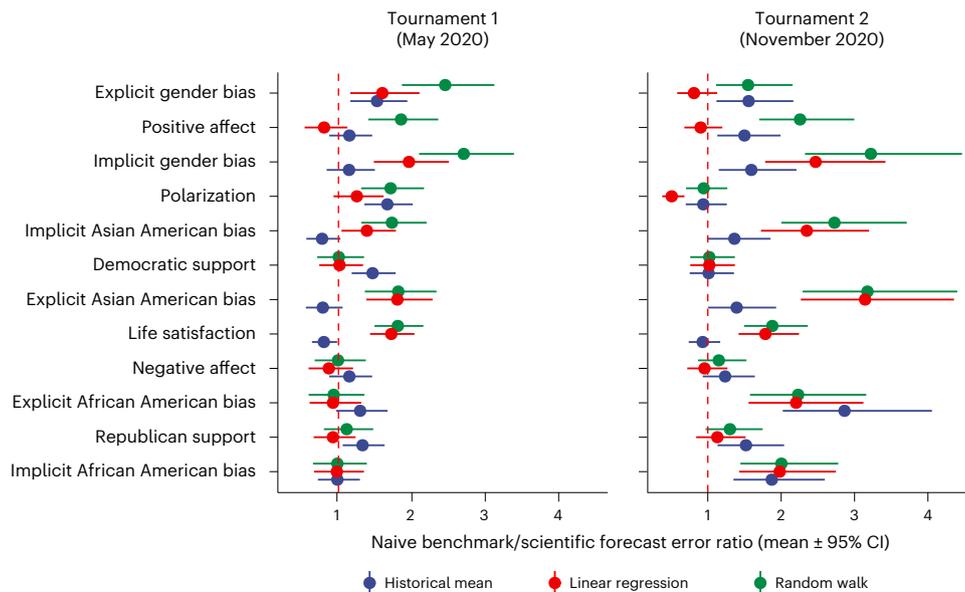


Fig. 3 | Ratios of forecasting errors among benchmarks compared to scientific forecasts. Scores greater than 1 indicate greater accuracy of scientific forecasts. Scores less than 1 indicate greater accuracy of naive benchmarks. The domains are ranked from least to most error among scientific teams in Tournament 1. The estimated means indicate the fixed-effect coefficients of linear mixed models 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% CIs. The CIs were adjusted for simultaneous inference of estimates for 12 domains in each tournament by simulating a multivariate t distribution²⁰.

societal change in political ideology, obtained via an aggregate of multiple nationally representative surveys by reputable pollsters, yet this domain was among the 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 (Fig. 4).

Domain differences in forecasting accuracy corresponded to differences in the complexity of historical data: domains that were more variable in terms of standard deviation 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: $\rho(\text{s.d.}) = 0.19$, $\rho(\text{MAD}) = 0.20$; Tournament 2: $\rho(\text{s.d.}) = 0.48$, $\rho(\text{MAD}) = 0.36$), and domain changes in the variability of historical data across tournaments corresponded to changes in accuracy ($\rho(\text{s.d.}) = 0.27$, $\rho(\text{MAD}) = 0.28$).

Comparison of accuracy across tournaments

Forecasting error was higher in the first tournament than in the second tournament (Fig. 4) ($F(1, 889.48) = 64.59$, $P < 0.001$, $R^2 = 0.063$). We explored several possible differences between the tournaments that may account for this effect. One possibility is that the characteristics of teams differed between tournaments (such as team size, gender, number of forecasted domains, field specialization and team diversity, number of PhDs on a team, and prior experience with forecasting). However, the difference between the tournaments remained equally pronounced when we ran parallel analyses with team characteristics as covariates ($F(1, 847.79) = 90.45$, $P < 0.001$, $R^2 = 0.062$).

Another hypothesis is that forecasts for 12 months (Tournament 1) include further-removed data points than forecasts for 6 months (Tournament 2), and the greater temporal distance between the tournament and the moment to forecast resulted in greater inaccuracy in 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 constants. Contrary to this seemingly straightforward hypothesis, error for the forecasts for the first six months was in fact significantly greater (MASE = 3.16; s.e. = 0.21; 95% CI, (2.77, 3.60)) than for the last six months (MASE = 2.59; s.e. = 0.17; 95% CI, (2.27, 2.95)) ($F(1, 621.41) = 29.36$, $P < 0.001$, $R^2 = 0.012$). 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 than in the last six months. This suggests that for several domains, social scientists anchored 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 than for the November 2020–April 2021 forecasts.

Finally, we tested whether providing the teams an additional six months of historical data capturing the onset of the pandemic in Tournament 2 may have contributed to lower error than in Tournament 1. To this end, we compared the inaccuracy of forecasts for the six-month period of November 2020–April 2021 done in May 2020 (Tournament 1) and those done 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 mean,

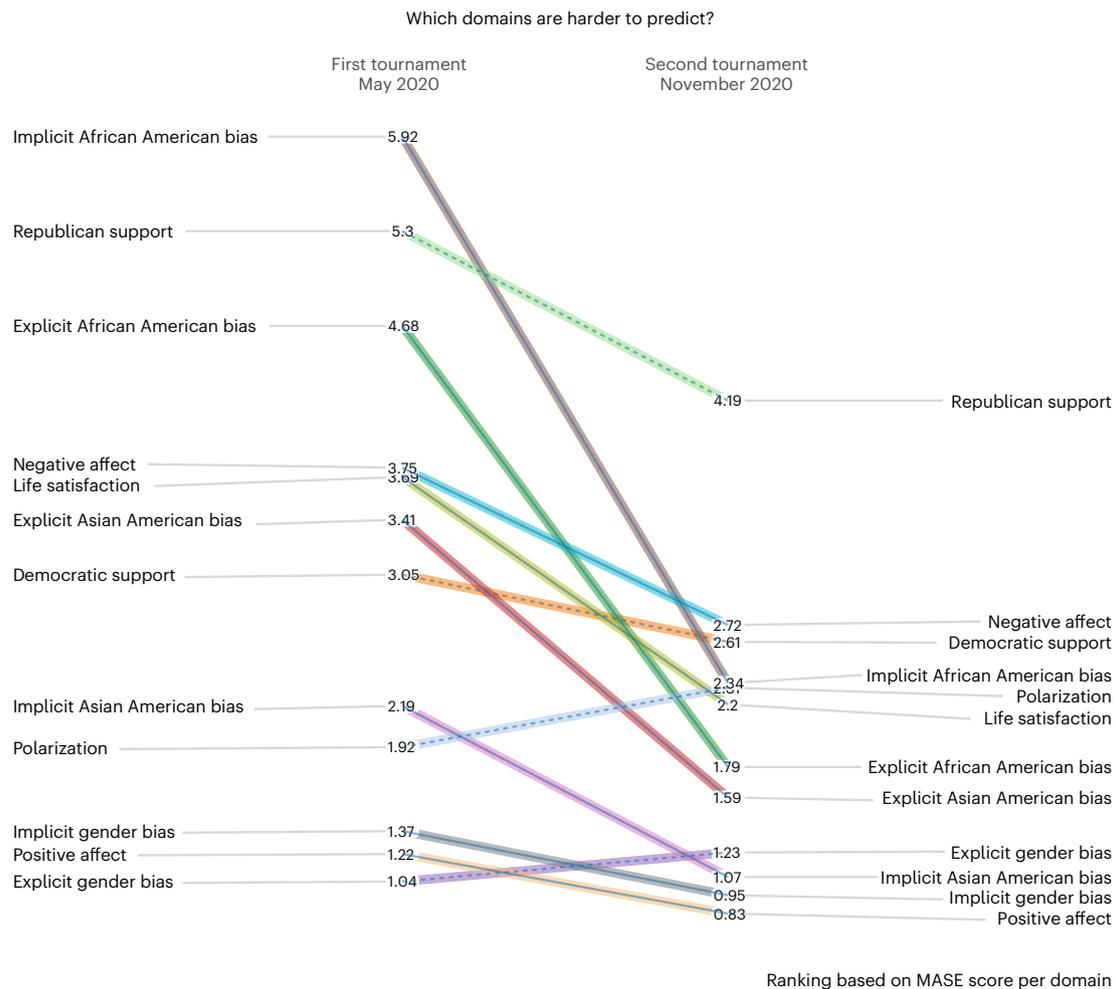


Fig. 4 | Cross-tournament consistency in the ranking of domains in terms of forecasting inaccuracy. Cross-tournament consistency in the ranking of domains in terms of forecasting inaccuracy. Left part of the graph shows ranking of domains in terms of the estimated mean forecasting error, assessed via MASE, across all teams in the first tournament (May 2020) from most to least inaccurate. Right part of the graph shows corresponding ranking of domains for the second tournament (November 2020). A solid line of the slope graph indicates that the

change in accuracy between tournaments is statistically significant ($P < 0.05$); a dashed line indicates a non-significant change. Significance was determined via pairwise comparisons of $\log(\text{MASE})$ 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 $\log(\text{MASE})$ scores, with responses nested in scientific teams ($N_{\text{teams}} = 120$, $N_{\text{observations}} = 905$).

2.54; s.e. = 0.17; 95% CI, (2.23, 2.90)) than Tournament 2 forecasts (MASE mean, 1.99; s.e. = 0.13; 95% CI, (1.74, 2.27)) ($F(1, 607.79) = 31.57$, $P < 0.001$, $R^2 = 0.017$), suggesting that it was the availability of new (pandemic-specific) information rather than temporal distance that contributed to more accurate forecasts in the second than in 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 $r_{\text{across domains}} = 0.88$; 95% CI, (0.85, 0.90); $t(357) = 34.80$; $P < 0.001$; domain-specific $0.55 < r_s \leq 0.99$; second tournament: multilevel $r_{\text{across domains}} = 0.72$; 95% CI, (0.67, 0.75); $t(544) = 23.95$; $P < 0.001$; domain-specific $0.24 < r_s \leq 0.96$). Furthermore, the results of a linear mixed model with MASE scores in Tournament 1, domain, and their interaction predicting MASE in Tournament 2 showed that for 11 out of 12 domains, accuracy in Tournament 1 was associated with greater accuracy in Tournament 2 (median of standardized $\beta_s = 0.26$).

Moreover, the ranking of models based on performance in the initial 12-month tournament corresponds to the ranking of the updated models in the follow-up 6-month tournament (Fig. 4). Harder-to-predict domains in the initial tournament remained the most inaccurate in the second tournament. Figure 3 shows one notable exception. Bias against African Americans was easier to predict than other domains in the second tournament. This exception appears consistent with the idea that George Floyd's death catalysed movements in racial awareness just after the first tournament, although this explanation is speculative (see 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 five performers in each domain (Supplementary Figs. 4–11). We compared forecasting approaches relying on (1) no data modelling (but possible consideration of theories), (2) pure

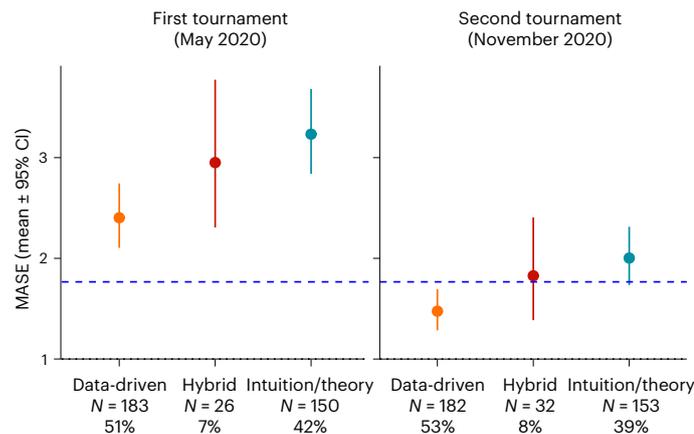


Fig. 5 | Forecasting errors by prediction approach. The estimated means and 95% CIs are based on a restricted information maximum likelihood linear mixed-effects model with model type (data-driven, hybrid or intuition/theory-based) as a fixed-effects predictor of the log(MASE) scores, domain as a fixed-effects 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 horizontal line show better performance than a naive in-sample random walk. Scores below the dashed horizontal line show better performance than the median performance in M4 tournaments⁷.

data modelling (but no consideration of subject matter theories) and (3) hybrid approaches. Roughly half of the teams relied on data-based modelling 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: $F(2, 149.10) = 5.47$, $P = 0.005$, $R^2 = 0.096$; second tournament: $F(2, 177.93) = 5.00$, $P = 0.008$, $R^2 = 0.091$) (Fig. 5). Forecasts that considered historical data as part of the forecast modelling were more accurate than models that did not (first tournament: $F(1, 56.29) = 20.38$, $P < 0.001$, $R^2 = 0.096$; second tournament: $F(1, 159.11) = 8.12$, $P = 0.005$, $R^2 = 0.084$). Model comparison effects were qualified by a significant model type \times domain interaction (first tournament: $F(11, 278.67) = 4.57$, $P < 0.001$, $R^2 = 0.045$; second tournament: $F(11, 462.08) = 3.38$, $P = 0.0002$, $R^2 = 0.028$). Post-hoc comparisons in Supplementary Table 4 revealed that data-inclusive (data-driven and hybrid) models were significantly more accurate than data-free models in three domains (explicit and implicit racial bias against Asian Americans and implicit gender-career bias) in Tournament 1 and two domains (life satisfaction and 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 ($t(577) = 0.87$, $P = 0.385$), whereas data-inclusive models tended to perform significantly better than lay estimates ($t(470) = 3.11$, $P = 0.006$, Cohen's $d = 0.391$).

To examine the incremental contributions 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 ($R^2 = 0.314$), though much of it was accounted for by differences in accuracy between domains (non-domain R^2 (partial), 0.043). Consistent with prior research²¹, model sophistication—that is, considering a larger number of exogenous predictors, COVID-19 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, as evidenced by the negative effect of statistical model complexity for accuracy ($B = 0.14$, s.e. = 0.06, $t(220.82) = 2.33$, $P = 0.021$, R^2 (partial) = 0.010).

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 had published academic research on the forecasted domain ($B = -0.26$, s.e. = 0.09, $t(711.64) = 3.01$, $P = 0.003$, R^2 (partial) = 0.007) and who had taken part in prior forecasting competitions ($B = -0.35$, s.e. = 0.17, $t(56.26) = 2.02$, $P = 0.049$, R^2 (partial) = 0.010) (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 0.009). Additional testing of whether the inclusion of US-based scientists influenced forecasting accuracy did not yield significant effects ($F(1, 106.61) < 1$).

In the second tournament, we provided the teams with the opportunity to compare their original forecasts (Tournament 1, May 2020) with new data at a later time point and to update their predictions (Tournament 2, November 2020). We therefore tested whether updating improved people's predictive accuracy. 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). The updated forecasts in the second tournament (November) were significantly more accurate than the original forecasts in the first tournament (May) ($t(94.5) = 6.04$, $P < 0.001$, Cohen's $d = 0.804$), but so were the forecasts from the 34 new teams recruited in November ($t(75.9) = 6.30$, $P < 0.001$, Cohen's $d = 0.816$). Furthermore, the updated forecasts were not significantly different from the forecasts provided by new teams recruited in November ($t(77.8) < 0.10$, $P = 0.928$). 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 change²²? The 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. Furthermore, apart from a few domains concerning racial and gender-career bias, scientists' original forecasts were typically not much better than naive statistical benchmarks derived from historical averages, linear regressions or random walks. Even when we confined the analysis to the top five 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 towards African Americans were the 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^{23,24}. Moreover, racial attitudes showed more change than attitudes

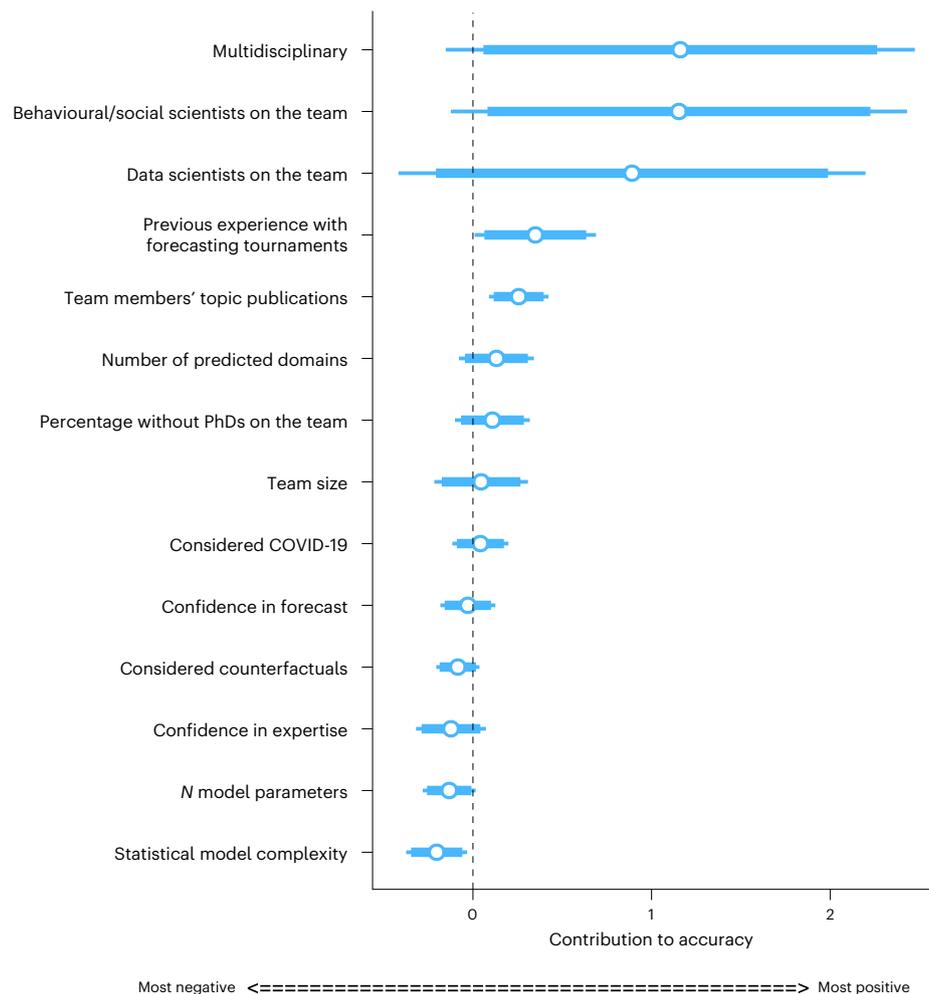


Fig. 6 | Contributions of specific forecasting strategies and team characteristics to forecasting accuracy. Contributions of specific forecasting strategies (n parameters, statistical model complexity, consideration of exogenous events and counterfactuals) and team characteristics to forecasting accuracy (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 line contribute negatively. The analyses control for domain type. All continuous predictors are mean-centred and scaled by two standard deviations, to afford comparability⁶⁴. The reported standard errors are heteroskedasticity robust. The thicker bands show the 90% CIs, and the thinner lines show the 95% CIs. The effects are statistically significant if the 95% CI does not include zero (dashed vertical line).

regarding gender during this period (perhaps due to movements such as 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 judgements²¹. 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 the tournaments (R^2 (partial) = 0.010). This is in line with some prior research on the value of subject matter expertise for geopolitical forecasts⁶ and for the prediction of success of behavioural science interventions²⁵. 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^{26,27,28} (versus intellectual humility). There is some evidence that overconfident forecasters are less accurate^{29,30}.

These findings, along with the lack of a domain-general effect of social science expertise on performance compared with the general public, invite consideration of whether what usually counts as expertise in the social sciences translates into a greater ability to predict future real-world trends.

The nature of our forecasting tournaments allowed social scientists to self-select any of the 12 forecasting domains, inspect three years of historical trends for each domain and update their predictions on the basis of feedback on their initial performance in the first tournament. These features emulated typical forecasting platforms (for example, 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 that foster 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 the determination of domains that one may want participants to forecast. In designing the present tournaments, we provided the 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 that the main sources of information about the forecasting domains remain identical across all participants. However, this approach restricts the types of social issues that participants can forecast. A simpler design without the inclusion of historical data would have had the advantage of greater flexibility in selecting forecasting domains.

Why were forecasts of societal change largely inaccurate, even though the participants had data-based resources and ample time to deliberate? One possibility concerns self-selection. Perhaps the participants in the Forecasting Collaborative were unusually bad at forecasting compared with 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 subdisciplines (Supplementary Information). Indeed, many of our forecasters are well-established scholars. We thus do not expect members of the Forecasting Collaborative to be worse at forecasting than other members of the social science community. Nevertheless, only a random sample of social scientists (albeit impractical) 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. We provided reputational incentives by announcing the winners and rankings of participating teams, but like other big-team science projects^{8,31}, we did not provide performance-based monetary incentives³², because they may not be key motivating factors for intrinsically motivated social scientists³³. Indeed, the drop-out rate between Tournaments 1 and 2 was marginal, suggesting that the 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^{8,31,34}. 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 on the basis of their lab studies, and in fact, such effects may be spurious given the lack of external validity^{35,36}. Social scientists may thus overestimate and misestimate the impacts of the effects they study in the lab on real-world phenomena^{37,38}.

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 the scale of entire societies³⁹. 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 modelling is not a requirement in many social sciences programmes¹⁰. Social scientists often prioritize explanations over formal predictions⁵. For instance, statistical training in the social sciences typically emphasizes unbiased estimation of model parameters in the sample over predictive out-of-sample accuracy⁴⁰. Moreover, typical graduate curricula in many areas of social science, such as social or clinical psychology, do not require computational training in predictive modelling. The formal empirical study of societal change is relatively uncommon in these disciplines. Most social scientists approach individual- or group-level phenomena in an atemporal fashion³⁹. Scientists may favour 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 the social sciences, such as field theory⁴¹, 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 (a 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. The 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 whether a given theory is suitable for making societal predictions in the first place or whether 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 whether 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 modelling the 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 predictors for a stationary time series⁴³. 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 behaviour, with a promise of greater generalizability in the real world.

Given the broad societal impact of phenomena such as prejudice, political polarization and well-being, the ability to accurately predict trends in these variables is crucially important for policymakers and the experts guiding them. But despite common beliefs that social science experts are better equipped to accurately predict these trends than non-experts¹, the current findings suggest that social and behavioural scientists have a lot of room for growth⁴⁴. The good news is that forecasting skills can be improved. Consider the growing accuracy of forecasting models in meteorology in the second part of the twentieth century⁴⁵. 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 no. 42142.

Pre-registration and deviations

The forecasts of all participating teams along with their rationales were pre-registered on the 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), the operationalization of the covariates and benchmarks (that is, the use of naive forecasting methods), and the key analytic procedures (linear mixed models and contrasts being different forecasting approaches; <https://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 were collected, though we did pre-register this benchmark in early September 2020, prior to data pre-processing or analyses. Deviating from the pre-registration, we performed a single analysis with all covariates in the same model rather than performing separate analyses for each set of covariates, to protect against inflating *P* values. Furthermore, 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 the corresponding analyses on the GitHub site for the project, <https://github.com/grossmania/Forecasting-Tournament>, which replicate the key effects presented in the main manuscript).

Participants and recruitment

We initially aimed for a minimum sample of 40 forecasting teams in our tournament after prescreening to ensure that the participants possessed at minimum a bachelor's degree in the behavioural, social or computer sciences. To ensure a sufficient sample for comparing groups of scientists employing different forecasting strategies (for example, data-free versus data-inclusive methods), we subsequently tripled the target size of the final sample ($N = 120$) and accomplished this target by the November phase of the tournament (see Supplementary Table 1 for the demographics).

The Forecasting Collaborative website that we used for recruitment (<https://predictions.uwaterloo.ca/faq>) outlined the guidelines for eligibility and the approach for prospective participants. We incentivized the participating teams in two ways. First, the prospective participants had an opportunity for co-authorship in a large-scale citizen science publication. Second, we incentivized accuracy by emphasizing throughout the recruitment that we would be announcing the winners and would share the rankings 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 (for example, model-based) or expertise-based (for example, general intuition or theory-based) forecasts from any field. As part of the survey, the 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 were no restrictions, though all teams were encouraged to report their education as well as areas of knowledge or expertise. The participants were recruited via large-scale advertising on social media; mailing lists in the behavioural and social sciences, the decision sciences, and data science; advertisement on academic social networks including ResearchGate; and word of mouth. To ensure broad representation across the academic spectrum of relevant disciplines, we targeted groups of scientists working on computational modelling, social psychology, judgement 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 US Institute for Health Metrics and Evaluation projected the initial peak of the COVID-19 pandemic in the United States. The recruitment phase continued until mid-June 2020, to ensure that at least 40 teams joined the initial tournament. We were able to recruit 86 teams for the initial 12-month tournament (mean age, 38.18; s.d. = 8.37; 73% of the forecasts were made by scientists with a doctorate), each of which provided forecasts for at least one domain (mean = 4.17; s.d. = 3.78). At the six-month mark after the 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 (mean age, 36.82; s.d. = 8.30; 67% of the forecasts were made by scientists with a doctorate; mean number of forecasted domains, 4.55; s.d. = 3.88). Supplementary

analyses showed that the updating likelihood did not significantly differ between data-free and data-inclusive models ($z = 0.50$, $P = 0.618$).

Procedure

Information for this project was available on the designated website (<https://predictions.uwaterloo.ca>), which included objectives, instructions and prior monthly data for each of the 12 domains that the participants could use for modelling. Researchers who decided to partake in the tournament signed up via a Qualtrics survey, which asked them to upload their estimates for the 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 (Supplementary Appendix 1) and to answer a set of questions about their rationale and forecasting team composition. Once all data were received, the de-identified responses were used to pre-register the forecasted values and models on the Open Science Framework (<https://osf.io/6wgbj/>).

At the halfway point (that is, at six months), the participants were provided with a comparison summary of their initial point estimate forecasts versus 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 modelling. On the basis of prior recommendations⁴⁶, in the first tournament we provided each team with 39 monthly estimates—from January 2017 to March 2020—for each of the domains that the participating teams chose to forecast. This approach enabled the teams to perform data-driven forecasting (should the teams choose to do so) and to establish a baseline estimate prior to the US peak of the pandemic. In the second tournament, conducted six months later, we provided the forecasting teams with 45 monthly time points—from January 2017 to September 2020.

Because of the requirement for rich standardized data for computational approaches to forecasting⁹, we limited the forecasting domains to issues of broad societal importance. Our domain selection was guided by the discussion of broad social consequences associated with these issues at the beginning of the pandemic^{47,48}, along with general theorizing about psychological and social effects of threats of infectious disease^{49,50}. An additional pragmatic consideration concerned 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, and stereotypes regarding gender and career versus family. To establish the common task framework—a necessary step for the evaluation of predictions in data science^{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 would evaluate accuracy at the end of the tournament.

Furthermore, each team had to (1) download and inspect the historical trends (visualized on an Excel plot; an example is in the Supplementary Information); (2) add their forecasts in the same document, which automatically visualized their forecasts against the historical trends; (3) confirm their forecasts; and (4) answer prompts concerning their forecasting rationale, theoretical assumptions, models, conditionals and consideration of additional parameters in the model. This procedure ensured that 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 relied 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-being⁵¹. For each month, we used previously validated predictive models of well-being, as measured by affective well-being and life satisfaction scales⁵². Affective well-being was calculated by applying a custom lexicon⁵³ 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 polls⁵⁴. 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 (<https://projects.fivethirtyeight.com/polls/generic-ballot/>), which ask representative samples of Americans to indicate which party they would support in an election. We weighed the polls on the basis of 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 cell phones and landlines. On the basis of these data, we then estimated monthly averages for support of the Democratic and Republican parties across pollsters (for example, Marist College, NBC/Wall Street Journal, CNN and 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 as the percentage of Republican versus Democratic approval ratings and estimated monthly averages for the period of interest. The absolute value of the difference score ensures that possible changes following the 2020 presidential election do 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 Americans (versus European Americans). To further probe racial bias, we sought to examine forecasted racial bias in attitudes towards African American (versus European American) people. Finally, we sought to examine gender bias in associations of the female (versus male) gender with family versus career. For each domain, we sought to obtain both estimates of explicit attitudes⁵⁵ and estimates of implicit attitudes⁵⁶. 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^{57,58} and when studying cross-temporal societal shifts in US attitudes⁵⁹. Despite the non-representative nature of the Project Implicit data, recent analyses suggest that the bias scores captured by Project Implicit are highly correlated with nationally representative estimates of explicit bias ($r = 0.75$), indicating that group aggregates of the bias data from Project

Implicit can reliably approximate group-level estimates⁵⁸. 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 IAT⁶⁰. The IAT is a computerized task comparing reaction times to categorize paired concepts (in this case, social groups—for example, Asian American versus European American) and attributes (in this case, valence categories—for example, good versus bad). Average response latencies in correct categorizations were compared across two paired blocks in which the 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’ and ‘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 the 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, the participants provided ratings on feeling thermometers towards Asian Americans and European Americans (to assess Asian American bias) and towards white and Black Americans (to assess racial bias), on a seven-point scale ranging from –3 to +3. Explicit gender–career bias was measured using seven-point Likert-type scales assessing the degree to which an attribute was female or male, from strongly female (–3) to strongly male (+3). Two questions assessed explicit stereotypes for each attribute (for example, 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 (for example, (male–career versus female–career) – (male–family versus 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 or African American) group. Similarly, for gender–career bias, –6 reflects a strong counter-stereotype association (for example, male–arts/female–science), and +6 reflects a strong stereotypical association (for example, female–arts/male–science). In both cases, the midpoint of 0 represents 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 the population estimates by weighting the monthly scores on the basis of their representativeness of the demographic frequencies in the US population (age, race, gender and education, estimated biannually by the US Census Bureau; <https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html>). Furthermore, we adjusted the weights on the basis of 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 weighted values for each participant, we computed weighted monthly means for each attitude test. These procedures ensured that the weighted monthly averages approximated the demographics of the US population. We cross-validated this procedure by comparing the weighted annual scores to nationally representative estimates

for feeling thermometers 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 American 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 we received feedback from co-authors, we adopted a more optimal approach wherein weights were computed on a monthly as opposed to yearly basis. This was necessary because demographic characteristics varied from month to month each year. This meant that using yearly weights had the potential to amplify bias instead of reducing it. Consequently, our new procedure ensured that sample representativeness was maximized. This insight affected forecasts from seven teams who had provided them before the change. The teams were informed, and four teams chose to provide updated estimates using the newly weighted historical data.

For each of these domains, the forecasters were provided with 39 monthly estimates in the initial tournament (45 estimates in the follow-up tournament), as well as detailed explanations of the origin and calculation of the respective indices. We thereby aimed to standardize the data source for the purpose of the forecasting competition⁹. See Supplementary Appendix 1 for example worksheets provided to the participants for submissions of their forecasts.

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

Confidence in forecasts. The participants rated their confidence in their forecasted points for each forecast model they submitted. These ratings were on a seven-point scale from 1 (not at all) to 7 (extremely).

Confidence in expertise. The participants provided ratings of their teams' expertise for a particular domain by indicating their extent of agreement with the statement "My team has strong expertise on the research topic of [field]." These ratings were on a seven-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, the participants reported whether they had used the past or predicted trajectory of the COVID-19 pandemic (as measured by the number of deaths or the prevalence of cases or new infections) as a conditional in their model, and if so, they 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. The participants described the key counterfactual events between December 2019 and April 2020 that they theorized would have led to different forecasts (for example, US-wide implementation of social distancing practices in February). Two independent coders evaluated the distinctiveness of the counterfactuals (interrater $\kappa = 0.80$). When discrepancies arose, the coders discussed individual cases with other members of the Forecasting Collaborative to make the final evaluation. In the primary analyses, we focus on the presence of counterfactuals (yes/no).

Team expertise. Because expertise can mean many things^{2,61}, we used a telescopic approach and operationalized expertise in four ways of varying granularity. First, we examined broad, domain-general expertise in the social sciences by comparing social scientists' forecasts with 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 the social sciences. To this end, we asked each participating team to report how many team members had a doctorate in the social sciences and calculated the percentage of doctorates on each team. Moving to domain-specific expertise, we instructed the participating teams to report whether 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 their team's expertise in a given domain (Supplementary Information).

General public benchmark. In parallel to the tournament with 86 teams, on 2–3 June 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 (well-being (life satisfaction and positive and negative sentiment on social media), politics (political polarization and ideological support for Democrats and Republicans), Asian American bias (explicit and implicit trends), African American bias (explicit and implicit trends) and gender-career bias (explicit and implicit trends)). During recruitment, the participants were informed that in exchange for 3.65 GBP, they had 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 the predictions did not exceed the possible range for a given domain (for example, polarization above 100%). Moreover, three coders (intercoder $\kappa = 0.70$ unweighted, $\kappa = 0.77$ weighted) reviewed all submitted rationales from lay people and excluded any submissions where the participants either misunderstood the task or wrote bogus bot-like responses. Coder disagreements were resolved via a discussion. Finally, we excluded responses if the participants spent under 50 seconds making their forecasts, which included reading instructions, downloading the files, providing forecasts and re-uploading their forecasts (final $N = 802$, 1,467 forecasts; mean age, 30.39; s.d. = 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; median 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 that the independent coders flagged as either misunderstood ($n = 6$) or bot-like bogus responses ($n = 26$). See Supplementary Appendix 2 for verbatim examples of each screening category and the exact coding instructions. Finally, we removed responses where the 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 on the basis of open-ended information where the participant

indicated they had made forecasts for a different country than the United States.

Naive statistical benchmarks. There is evidence from data science forecasting competitions that the dominant statistical benchmarks are the Theta method, ARIMA and ETS⁷. Given the socio-cultural context of our study and to avoid loss of generality, we decided to employ more traditional benchmarks such as naive/random walk, historical average and the basic linear regression model—that is, the method that is used more than anything else in practice and science. In short, we selected three benchmarks on the basis of their common application in the forecasting literature (historical mean and random walk are the most basic forecasting benchmarks) or the behavioural/social science literature (linear regression is the most basic statistical approach to test inferences in the 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 and 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 of scientists' performance, we examined whether social scientists' performance was better than each of the three benchmarks. To obtain metrics of uncertainty around the naive statistical estimates, we chose to simulate these three naive approaches for making forecasts: (1) random resampling of historical data, (2) a naive out-of-sample random walk based on random resampling of historical change and (3) extrapolation from a naive regression based on a randomly selected interval of historical data. We describe each approach in Supplementary Information.

Analytic plan

Categorization of forecasts. We categorized the forecasts on the basis of modelling approaches. Two independent research associates categorized the forecasts for each domain on the basis of the following justifications: (1) theoretical models only, (2) data-driven models only or (3) a combination of theoretical and data-driven models—that is, computational models that rely on specific theoretical assumptions. See Supplementary Appendix 3 for the exact coding instructions and a description of the classification (interrater $\kappa = 0.81$ unweighted, $\kappa = 0.90$ weighted). We further examined the modelling complexity of approaches that relied on the extrapolation of time series from the data we provided (for example, ARIMA or moving average with lags; yes/no; see Supplementary Appendix 4 for the exact coding instructions). Disagreements between coders here (interrater $\kappa = 0.80$ unweighted, $\kappa = 0.87$ weighted) and on 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 impacted forecasting accuracy. To this end, we ensured that additional variables were distinct from one another. Two independent coders evaluated the distinctiveness of each reported parameter (interrater $\kappa = 0.56$ unweighted, $\kappa = 0.83$ weighted).

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

Forecasting update justifications. Given that the participants received both new data and a summary of diverse theoretical positions that they could use as a basis for their updates, two independent

research associates scored the participants' justifications for forecasting updates on three dummy categories: (1) the new six months of data that we provided, (2) new theoretical insights and (3) consideration of other external events (interrater $\kappa = 0.63$ unweighted/weighted). See Supplementary Appendix 5 for the 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 the analyses, we inspected the MASE scores to determine violations of linearity, which we corrected via log-transformation before performing the 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% CIs by domain, we simulated a multivariate *t* distribution²⁰ to adjust the scores for simultaneous inference of estimates for 12 domains in each tournament.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in the main text and supplementary analysis are accessible on GitHub (<https://github.com/grossmania/Forecasting-Tournament>). All prior data presented to the forecasters are available at <https://predictions.uwaterloo.ca/>. Historical and ground truth markers were obtained from Project FiveThirtyEight (<https://projects.fivethirtyeight.com/polls/generic-ballot>), Gallup (<https://news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx>), Project Implicit (see the Open Science Framework website at <https://osf.io/t4bnj>) and the US 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 the Reporting Summary for the R packages and their versions.

References

- Hutcherson, C. et al. On the accuracy, media representation, and public perception of psychological scientists' judgments of societal change. Preprint at <https://doi.org/10.31234/osf.io/g8f9s> (2023).
- Collins, H. & Evans, R. *Rethinking Expertise* (Univ. of Chicago Press, 2009).
- Fama, E. F. Efficient capital markets: a review of theory and empirical work. *J. Finance* **25**, 383–417 (1970).
- Tetlock, P. E. *Expert Political Judgement: How Good Is It?* (Princeton University Press, 2017).
- Hofman, J. M. et al. Integrating explanation and prediction in computational social science. *Nature* **595**, 181–188 (2021).
- Mandel, D. R. & Barnes, A. Accuracy of forecasts in strategic intelligence. *Proc. Natl Acad. Sci. USA* **111**, 10984–10989 (2014).
- Makridakis, S., Spiliotis, E. & Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. *Int. J. Forecast.* **36**, 54–74 (2020).
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
- Hofman, J. M., Sharma, A. & Watts, D. J. Prediction and explanation in social systems. *Science* **355**, 486–488 (2017).
- Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).
- 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**, 0003 (2016).
13. Schaller, M. & Murray, D. R. Pathogens, personality, and culture: disease prevalence predicts worldwide variability in socio-sexuality, 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. Hawkey, 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. USA* **117**, 8398–8403 (2020).
17. Liberman, M. *Reproducible Research and the Common Task Method* (2015); <https://www.simonsfoundation.org/event/reproducible-research-and-the-common-task-method/>
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* <https://doi.org/10.3758/s13428-021-01694-3> (2021).
20. Genz, A. & Bretz, F. *Computation of Multivariate Normal and t Probabilities* (Springer, 2009).
21. Green, K. C. & Armstrong, J. S. Simple versus complex forecasting: the evidence. *J. Bus. Res.* **68**, 1678–1685 (2015).
22. Grossmann, I., Twardus, O., Varnum, M. E. W., Jayawickreme, E. & McLevey, J. Expert predictions of societal change: insights from the World After COVID Project. *Am. Psychol.* **77**, 276–290 (2022).
23. Grossmann, I., Huynh, A. C. & Ellsworth, P. C. Emotional complexity: clarifying definitions and cultural correlates. *J. Pers. Soc. Psychol.* **111**, 895–916 (2016).
24. Alves, H., Koch, A. & Unkelbach, C. Why good is more alike than bad: processing implications. *Trends Cogn. Sci.* **21**, 69–79 (2017).
25. 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).
26. Dunning, D., Heath, C. & Suls, J. M. Flawed self-assessment. *Psychol. Sci. Public Interest* **5**, 69–106 (2004).
27. Grossmann, I. et al. The science of wisdom in a polarized world: knowns and unknowns. *Psychol. Inq.* **31**, 103–133 (2020).
28. Porter, T. et al. Predictors and consequences of intellectual humility. *Nat. Rev. Psychol.* **1**, 524–536 (2022).
29. Mellers, B., Tetlock, P. E. & Arkes, H. R. Forecasting tournaments, epistemic humility and attitude depolarization. *Cognition* **188**, 19–26 (2019).
30. Grossmann, I. et al. Training for wisdom: the distanced-self-reflection diary method. *Psychol. Sci.* **32**, 381–394 (2021).
31. 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).
32. Voslinsky, A. & Azar, O. H. Incentives in experimental economics. *J. Behav. Exp. Econ.* **93**, 101706 (2021).
33. 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).
34. Richard, F. D., Bond, C. F. Jr. & Stokes-Zoota, J. J. One hundred years of social psychology quantitatively described. *Rev. Gen. Psychol.* **7**, 331–363 (2003).
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. Cesario, J. What can experimental studies of bias tell us about real-world group disparities? *Behav. Brain Sci.* <https://doi.org/10.1017/S0140525X21000017> (2021).
38. IJzerman, H. et al. Use caution when applying behavioural science to policy. *Nat. Hum. Behav.* **4**, 1092–1094 (2020).
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**, 199–231 (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. & Gavrillets, S. War, space, and the evolution of Old World complex societies. *Proc. Natl Acad. Sci. USA* **110**, 16384–16389 (2013).
43. Brockwell, P. J. & Davis, R. A. *Introduction to Time Series and Forecasting* (Springer, 2016); <https://doi.org/10.1007/978-3-319-29854-2>
44. Makridakis, S. & Taleb, N. Living in a world of low levels of predictability. *Int. J. Forecast.* **25**, 840–844 (2009).
45. Hitchens, N. M., Brooks, H. E. & Kay, M. P. Objective limits on forecasting skill of rare events. *Weather Forecast.* **28**, 525–534 (2013).
46. Jebb, A. T., Tay, L., Wang, W. & Huang, Q. Time series analysis for psychological research: examining and forecasting change. *Front. Psychol.* **6**, 727 (2015).
47. Van Bavel, J. et al. Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **4**, 460–471 (2020).
48. Seitz, B. M. et al. The pandemic exposes human nature: 10 evolutionary insights. *Proc. Natl Acad. Sci. USA* **117**, 27767–27776 (2020).
49. Schaller, M. & Park, J. H. The behavioral immune system (and why it matters). *Curr. Dir. Psychol. Sci.* **20**, 99–103 (2011).
50. Wang, I. M., Michalak, N. M. & Ackerman, J. M. in *The SAGE Handbook of Personality and Individual Differences: Origins of Personality and Individual Differences* Vol. 2 (eds Zeigler-Hill, V. & Shackelford, T. K.) 321–345 (2018); <https://doi.org/10.4135/9781526451200.n18>
51. Luhmann, M. Using Big Data to study subjective well-being. *Curr. Opin. Behav. Sci.* **18**, 28–33 (2017).
52. Schwartz, H. A. et al. Predicting individual well-being through the language of social media. *Biocomputing 2016* https://doi.org/10.1142/9789814749411_0047 (2016).
53. Kiritchenko, S., Zhu, X. & Mohammad, S. M. Sentiment analysis of short informal texts. *J. Artif. Intell. Res.* **50**, 723–762 (2014).
54. Witters, D. & Harter, J. In U.S., *Life Ratings Plummet to 12-Year Low* (2020); <https://news.gallup.com/poll/391331/life-ratings-drop-month-low.aspx>
55. Axt, J. R. The best way to measure explicit racial attitudes is to ask about them. *Soc. Psychol. Pers. Sci.* **9**, 896–906 (2018).
56. Nosek, B. A. et al. Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur. Rev. Soc. Psychol.* **18**, 36–88 (2007).
57. Hehman, E., Flake, J. K. & Calanchini, J. Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Soc. Psychol. Pers. Sci.* **9**, 393–401 (2018).
58. 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. USA* **116**, 8846–8851 (2019).
59. 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).
60. 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).

61. Gobet, F. The future of expertise: the need for a multidisciplinary approach. *J. Expertise* **1**, 107–113 (2018).
62. Lenth, R., Singmann, H., Love, J. & Maxime, H. emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.0 (2020).
63. R Core Team. R: A Language and Environment for Statistical Computing (2022).
64. Gelman, A. Scaling regression inputs by dividing by two standard deviations. *Stat. Med.* **27**, 2865–2873 (2008).

Acknowledgements

This programme of research was supported by the Basic Research Program at the National Research University Higher School of Economics (M. Fabrykant), John Templeton Foundation grant no. 62260 (I.G. and P.E.T.), Kega 079UK-4/2021 (P.K.), Ministerio de Ciencia e Innovación España grants no. PID2019-111512RB-I00-HMDM and no. HDL-HS-280218 (A.A.), the National Center for Complementary & Integrative Health of the National Institutes of Health under award no. K23AT010879 (S.B.G.), National Science Foundation RAPID grant no. 2026854 (M.E.W.V.), PID2019-111512RB-I00 (M.S.), NPO Systemic Risk Institute grant no. LX22NPO5101 (I.R.), the Slovak Research and Development Agency under contract no. APVV-20-0319 (M.A.), Social Sciences and Humanities Research Council of Canada Insight grant no. 435-2014-0685 (I.G.), Social Sciences and Humanities Research Council of Canada Connection grant no. 611-2020-0190 (I.G.), and Swiss National Science Foundation grant no. PPOOP1_170463 (O. Strijbis). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank J. 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

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., E.S., 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 and 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., E.S., 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

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41562-022-01517-1>.

Peer review information *Nature Human Behaviour* thanks Richard Klein and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© His Majesty the King in Right of Canada as represented by Department of National Defence 2023

The Forecasting Collaborative

Igor Grossmann¹, Amanda Rotella^{1,2}, Cendri A. Hutcherson³, Konstantyn Sharpinskiy¹, Michael E. W. Varnum⁴, Sebastian Achter⁵, Mandeep K. Dhami⁶, Xinqi Evie Guo⁷, Mane Kara-Yakoubian⁸, David R. Mandel^{9,10}, Louis Raes¹¹, Louis Tay¹², Aymeric Vie^{13,14}, Lisa Wagner¹⁵, Matus Adamkovic^{16,17}, Arash Arami^{18,19}, Patrícia Arriaga²⁰, Kasun Bandara²¹, Gabriel Banik¹⁶, František Bartoš²², Ernest Baskin²³, Christoph Bergmeir²⁴, Michał Białek²⁵, Caroline K. Børsting²⁶, Dillon T. Browne¹, Eugene M. Caruso²⁷, Rong Chen²⁸, Bin-Tzong Chie²⁹, William J. Chopik³⁰, Robert N. Collins⁹, Chin Wen Cong³¹, Lucian G. Conway³², Matthew Davis³³, Martin V. Day³⁴, Nathan A. Dhaliwal³⁵, Justin D. Durham³⁶, Martyna Dziekan³⁷, Christian T. Elbaek²⁶, Eric Shuman³⁸, Marharyta Fabrykant^{39,40}, Mustafa Firat⁴¹, Geoffrey T. Fong^{1,42}, Jeremy A. Frimer⁴³, Jonathan M. Gallegos⁴⁴, Simon B. Goldberg⁴⁵, Anton Gollwitzer^{46,47}, Julia Goyal⁴⁸, Lorenz Graf-Vlachy^{49,50}, Scott D. Gronlund³⁶, Sebastian Hafenbrädl⁵¹, Andree Hartanto⁵², Matthew J. Hirshberg⁵³, Matthew J. Hornsey⁵⁴, Piers D. L. Howe⁵⁵, Anoosha Izadi⁵⁶, Bastian Jaeger⁵⁷, Pavol Kačmár⁵⁸, Yeun Joon Kim⁵⁹, Ruslan Krenzler^{60,61}, Daniel G. Lannin⁶², Hung-Wen Lin⁶³, Nigel Mantou Lou^{64,65}, Verity Y. Q. Lua⁵², Aaron W. Lukaszewski^{66,67}, Albert L. Ly⁶⁸, Christopher R. Madan⁶⁹, Maximilian Maier⁷⁰, Nadyanna M. Majeed⁷¹, David S. March⁷², Abigail A. Marsh⁷³, Michal Misiak^{25,74}, Kristian Ove R. Myrseth⁷⁵, Jaime M. Napan⁶⁸, Jonathan Nicholas⁷⁶, Konstantinos Nikolopoulos⁷⁷, Jiaqing O⁷⁸, Tobias Otterbring^{79,80}, Mariola Paruzel-Czachura^{81,82}, Shiva Pauer²², John Protzko⁸³, Quentin Raffaelli⁸⁴, Ivan Ropovik^{85,86}, Robert M. Ross⁸⁷, Yefim Roth⁸⁸, Espen Røysamb⁸⁹, Landon Schnabel⁹⁰, Astrid Schütz⁹¹, Matthias Seifert⁹², A. T. Sevincer⁹³, Garrick T. Sherman⁹⁴, Otto Simonsson^{95,96}, Ming-Chien Sung⁹⁷, Chung-Ching Tai⁹⁷, Thomas Talhelm⁹⁸, Bethany A. Teachman⁹⁹, Philip E. Tetlock^{100,101}, Dimitrios Thomakos¹⁰²,

Dwight C. K. Tse¹⁰³, Oliver J. Twardus¹⁰⁴, Joshua M. Tybur⁵⁷, Lyle Ungar⁹⁴, Daan Vandermeulen¹⁰⁵, Leighton Vaughan Williams¹⁰⁶, Hrag A. Vosgerichian¹⁰⁷, Qi Wang¹⁰⁸, Ke Wang¹⁰⁹, Mark E. Whiting^{110,111}, Conny E. Wollbrant¹¹², Tao Yang¹¹³, Kumar Yogeeswaran¹¹⁴, Sangsuk Yoon¹¹⁵, Ventura R. Alves¹¹⁶, Jessica R. Andrews-Hanna^{84,117}, Paul A. Bloom⁷⁶, Anthony Boyles¹¹⁸, Loo Charis¹¹⁹, Mingyeong Choi¹²⁰, Sean Darling-Hammond¹²¹, Z. E. Ferguson¹²², Cheryl R. Kaiser⁴⁴, Simon T. Karg¹²³, Alberto López Ortega⁵⁷, Lori Mahoney¹²⁴, Melvin S. Marsh¹²⁵, Marcellin F. R. C. Martinie⁵⁵, Eli K. Michaels¹²⁶, Philip Millroth¹²⁷, Jeanean B. Naqvi¹²⁸, Weiting Ng¹²⁹, Robb B. Rutledge¹³⁰, Peter Slattery¹³¹, Adam H. Smiley⁴⁴, Oliver Strijbis¹³², Daniel Szyner¹³³, Eli Tsukayama¹³⁴, Austin van Loon¹³⁵, Jan G. Voelkel¹³⁵, Margaux N. A. Wienk⁷⁶ & Tom Wilkening¹³⁶

¹Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada. ²Department of Psychology, Northumbria University, Northumbria, UK. ³Department of Psychology, University of Toronto Scarborough, Toronto, Ontario, Canada. ⁴Department of Psychology, Arizona State University, Tempe, AZ, USA. ⁵Institute of Management Accounting and Simulation, Hamburg University of Technology, Hamburg, Germany. ⁶Department of Psychology, Middlesex University London, London, UK. ⁷Department of Experimental Psychology, University of California, San Diego, San Diego, CA, USA. ⁸Department of Psychology, Toronto Metropolitan University, Toronto, Ontario, Canada. ⁹Defence Research and Development Canada, Toronto, Ontario, Canada. ¹⁰Department of Psychology, York University, Toronto, Ontario, Canada. ¹¹Department of Economics, Tilburg University, Tilburg, the Netherlands. ¹²Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA. ¹³Mathematical Institute, University of Oxford, Oxford, UK. ¹⁴Institute of New Economic Thinking, University of Oxford, Oxford, UK. ¹⁵Jacobs Center for Productive Youth Development, University of Zurich, Zurich, Switzerland. ¹⁶Institute of Psychology, University of Prešov, Prešov, Slovakia. ¹⁷Institute of Social Sciences, CSPA, Slovak Academy of Sciences, Bratislava, Slovakia. ¹⁸Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, Ontario, Canada. ¹⁹Toronto Rehabilitation Institute (KITE), University Health Network, Toronto, Canada. ²⁰Iscte-University Institute of Lisbon, CIS, Lisbon, Portugal. ²¹Melbourne Centre for Data Science, University of Melbourne, Melbourne, Victoria, Australia. ²²Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, the Netherlands. ²³Department of Food Marketing, Haub School of Business, Saint Joseph's University, Philadelphia, PA, USA. ²⁴Department of Data Science and Artificial Intelligence, Monash University, Melbourne, Victoria, Australia. ²⁵Institute of Psychology, University of Wrocław, Wrocław, Poland. ²⁶Department of Management, Aarhus University, Aarhus, Denmark. ²⁷Anderson School of Management, University of California, Los Angeles, Los Angeles, CA, USA. ²⁸Department of Psychology, Dominican University of California, San Rafael, CA, USA. ²⁹Department of Industrial Economics, Tamkang University, New Taipei City, Taiwan. ³⁰Department of Psychology, Michigan State University, East Lansing, MI, USA. ³¹Independent Researcher, Penang, Malaysia. ³²Psychology Department, Grove City College, Grove City, PA, USA. ³³Department of Economics, Siena College, Loudonville, NY, USA. ³⁴Department of Psychology, Memorial University of Newfoundland, St. John's, Newfoundland, Canada. ³⁵UBC Sauder School of Business, University of British Columbia, Vancouver, British Columbia, Canada. ³⁶Department of Psychology, University of Oklahoma, Norman, OK, USA. ³⁷Faculty of Psychology and Cognitive Science, Adam Mickiewicz University, Poznań, Poland. ³⁸Department of Psychology, University of Groningen, Groningen, the Netherlands. ³⁹Laboratory for Comparative Studies in Mass Consciousness, Expert Institute, HSE University, Moscow, Russia. ⁴⁰Faculty of Philosophy and Social Sciences, Belarusian State University, Minsk, Belarus. ⁴¹Department of Sociology, Radboud University, Nijmegen, the Netherlands. ⁴²Ontario Institute for Cancer Research, Toronto, Ontario, Canada. ⁴³Department of Psychology, University of Winnipeg, Winnipeg, Manitoba, Canada. ⁴⁴Department of Psychology, University of Washington, Seattle, WA, USA. ⁴⁵Department of Counseling Psychology, University of Wisconsin–Madison, Madison, WI, USA. ⁴⁶Department of Leadership and Organizational Behaviour, BI Norwegian Business School, Oslo, Norway. ⁴⁷Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. ⁴⁸School of Public Health Sciences, University of Waterloo, Waterloo, Ontario, Canada. ⁴⁹TU Dortmund University, Dortmund, Germany. ⁵⁰ESCP Business School, Paris, France. ⁵¹IESE Business School, Barcelona, Spain. ⁵²School of Social Sciences, Singapore Management University, Singapore, Singapore. ⁵³Center for Healthy Minds, University of Wisconsin–Madison, Madison, WI, USA. ⁵⁴University of Queensland Business School, Brisbane, Queensland, Australia. ⁵⁵Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Victoria, Australia. ⁵⁶Department of Marketing, University of Massachusetts Dartmouth, Dartmouth, MA, USA. ⁵⁷Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands. ⁵⁸Department of Psychology, Faculty of Arts, Pavol Jozef Šafárik University in Košice, Košice, Slovakia. ⁵⁹Cambridge Judge Business School, University of Cambridge, Cambridge, UK. ⁶⁰Hermes Germany GmbH, Hamburg, Germany. ⁶¹University of Hamburg, Hamburg, Germany. ⁶²Department of Psychology, Illinois State University, Normal, IL, USA. ⁶³Department of Business Administration, National Pingtung University, Pingtung City, Taiwan. ⁶⁴Department of Psychology, University of Victoria, Victoria, British Columbia, Canada. ⁶⁵Centre for Youth and Society, University of Victoria, Victoria, British Columbia, Canada. ⁶⁶Department of Psychology, California State University, Fullerton, Fullerton, CA, USA. ⁶⁷Center for the Study of Human Nature, California State University, Fullerton, Fullerton, CA, USA. ⁶⁸Department of Psychology, Loma Linda University, Loma Linda, CA, USA. ⁶⁹University of Nottingham, Nottingham, UK. ⁷⁰Department of Experimental Psychology, University College London, London, UK. ⁷¹Singapore Management University, Singapore, Singapore. ⁷²Department of Psychology, Florida State University, Tallahassee, FL, USA. ⁷³Department of Psychology, Georgetown University, Washington, DC, USA. ⁷⁴School of Anthropology & Museum Ethnography, University of Oxford, Oxford, UK. ⁷⁵School for Business and Society, University of York, York, UK. ⁷⁶Department of Psychology, Columbia University, New York, NY, USA. ⁷⁷IHRR Forecasting Laboratory, Durham University, Durham, UK. ⁷⁸Department of Psychology, Aberystwyth University, Aberystwyth, UK. ⁷⁹School of Business and Law, Department of Management, University of Agder, Kristiansand, Norway. ⁸⁰Institute of Retail Economics, Stockholm, Sweden. ⁸¹Institute of Psychology, University of Silesia in Katowice, Katowice, Poland. ⁸²Department of Neurology, Penn Center for Neuroaesthetics, University of Pennsylvania, Philadelphia, PA, USA. ⁸³Central Connecticut State University, New Britain, CT, USA. ⁸⁴Department of Psychology, University of Arizona, Tucson, AZ, USA. ⁸⁵Faculty of Education, Institute for Research and Development of Education, Charles University, Prague, Czech Republic. ⁸⁶Faculty of Education, University of Prešov, Prešov, Slovakia. ⁸⁷School of Psychology, Macquarie University, Sydney, New South Wales, Australia. ⁸⁸Department of Human Service, University of Haifa, Haifa, Israel. ⁸⁹Promenta Center, Department of Psychology, University of Oslo, Oslo, Norway. ⁹⁰Department of Sociology, Cornell University, Ithaca, NY, USA. ⁹¹Institute of Psychology, University of Bamberg, Bamberg, Germany. ⁹²IE Business School, IE University, Madrid, Spain. ⁹³Faculty of Psychology and Human Movement Science, University of Hamburg, Hamburg, Germany. ⁹⁴Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. ⁹⁵Department of Clinical Neuroscience, Karolinska Institutet, Solna, Sweden. ⁹⁶Department of Sociology, University of Oxford, Oxford, UK. ⁹⁷Department of Decision Analytics and Risk, University of Southampton, Southampton, UK. ⁹⁸University of Chicago Booth School of Business, Chicago, IL, USA. ⁹⁹Department of Psychology, University of Virginia, Charlottesville, VA, USA. ¹⁰⁰Psychology Department, University of Pennsylvania, Philadelphia, PA, USA. ¹⁰¹Wharton School of Business, University of Pennsylvania, Philadelphia, PA, USA. ¹⁰²Department of Economics, National and Kapodistrian University of Athens, Athens, Greece. ¹⁰³School of Psychological Sciences and Health, University of Strathclyde, Glasgow, UK. ¹⁰⁴Department of Psychology, University of Guelph, Guelph, Ontario, Canada. ¹⁰⁵Psychology Department, Hebrew University of Jerusalem, Jerusalem, Israel. ¹⁰⁶Department of Economics, Nottingham Trent University, Nottingham, UK.

¹⁰⁷Department of Management and Organizations, Northwestern University, Evanston, IL, USA. ¹⁰⁸College of Human Ecology, Cornell University, Ithaca, NY, USA. ¹⁰⁹Harvard Kennedy School, Harvard University, Cambridge, MA, USA. ¹¹⁰Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. ¹¹¹Operations, Information, and Decisions Department, the Wharton School, University of Pennsylvania, Philadelphia, PA, USA. ¹¹²School of Economics and Finance, University of St. Andrews, St. Andrews, UK. ¹¹³Department of Management, Cameron School of Business, University of North Carolina Wilmington, Wilmington, NC, USA. ¹¹⁴School of Psychology, Speech and Hearing, University of Canterbury, Christchurch, New Zealand. ¹¹⁵Department of Marketing, University of Dayton, Dayton, OH, USA. ¹¹⁶ISG Universidade Lusofona, Lisbon, Portugal. ¹¹⁷Cognitive Science, University of Arizona, Tucson, AZ, USA. ¹¹⁸Ephemer AI, Atlanta, GA, USA. ¹¹⁹Questrom School of Business, Boston University, Boston, MA, USA. ¹²⁰Institute of Social Science Research, Pusan National University, Busan, South Korea. ¹²¹Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, USA. ¹²²Psychology Department, University of Washington, Seattle, WA, USA. ¹²³Department of Political Science, Aarhus University, Aarhus, Denmark. ¹²⁴College of Science and Mathematics, Wright State University, Fairborn, OH, USA. ¹²⁵Department of Psychology, Georgia Southern University, Statesboro, GA, USA. ¹²⁶Division of Epidemiology, School of Public Health, University of California, Berkeley, Berkeley, CA, USA. ¹²⁷Department of Psychology, Uppsala University, Uppsala, Sweden. ¹²⁸Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA. ¹²⁹School of Humanities & Behavioral Sciences, Singapore University of Social Sciences, Singapore, Singapore. ¹³⁰Department of Psychology, Yale University, New Haven, CT, USA. ¹³¹BehaviourWorks Australia, Monash University, Melbourne, Victoria, Australia. ¹³²Institute of Political Science, University of Zurich, Zurich, Switzerland. ¹³³Department of Psychology, Oklahoma State University, Stillwater, OK, USA. ¹³⁴Department of Business Administration, University of Hawaii–West Oahu, Kapolei, HI, USA. ¹³⁵Department of Sociology, Stanford University, Stanford, CA, USA. ¹³⁶Department of Economics, University of Melbourne, Melbourne, Victoria, Australia.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-----|-----------|
| n/a | Confirmed |
|-----|-----------|
- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
 - A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
 - The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
 - A description of all covariates tested
 - A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
 - A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
 - For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
 - For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
 - For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
 - Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Qualtrics Excel
Data analysis	R version 4.2.2 (2022-10-31 ucrt) with the following packages: tidyquant_1.0.4 quantmod_0.4.20 TTR_0.24.3 PerformanceAnalytics_2.0.4 xts_0.12.1 zoo_1.8-10 ggdist_3.2.0 bayestestR_0.12.1 rstanarm_2.21.3 Rcpp_1.0.9 ggpubr_0.4.0 moments_0.14.1 partR2_0.9.1 CGPfunctions_0.6.3 tsibble_1.1.2 statcomp_0.1.0 lubridate_1.8.0 Hmisc_4.7-0 Formula_1.2-4 survival_3.4-0 lattice_0.20-45 ggsci_2.9 jtools_2.2.0 car_3.1-0 carData_3.0-5 emmeans_1.8.0 lme4_1.1-30 Matrix_1.5-1 irr_0.84.1 lpSolve_5.6.15 forcats_0.5.1 stringr_1.4.0 dplyr_1.0.9 purrr_0.3.4 readr_2.1.2 tidyr_1.2.0 tibble_3.1.8 ggplot2_3.3.6 tidyverse_1.3.2 psych_2.2.5 forecast_8.17.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

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/>.

Ground truth markers:

- projects.fivethirtyeight.com/congress-generic-ballot-polls
- Gallup Presidential Approval Ratings <https://news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx>
- Project Implicit Open Science Framework website <https://osf.io/t4bnj>
- U.S. Census Bureau <https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html>

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

For the scientist teams, we only collected data on self-identified gender of teams and quantified % of team members who indicated their gender was either female or other.

For lay sample, we included relevant gender info in the methods section, when describing the sample. Both indices were included for descriptive purposes only, without hypotheses about the role of gender.

Population characteristics

We were able to recruit 86 scientist 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 forecasted domains = 4.55; SD = 3.88).

General public benchmark: final N = 802, 1,467 forecasts; Mage = 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).

Recruitment

Scientists. 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 Table S1 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 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, $z = 0.50$, $P = .618$.

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)

Ethics oversight

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

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

In our quantitative study, we conducted two forecasting tournaments through the Forecasting Collaborative—a crowdsourced initiative among scientists interested in ex-ante testing of their theoretical or data-driven models. 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 S1). 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 examined accuracy of teams by comparing their predictions against ground truth markers we gathered a year later. 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.

Research sample

We were able to recruit 86 scientist 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 forecasted domains = 4.55; SD = 3.88). The same of scientists was not representative, as were trying to recruit scientists from a range of fields, but had to do it during the first peak of a COVID-19 pandemic.

General public benchmark: final $N = 802, 1,467$ forecasts; Mage = 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). We recruited a regionally-stratified, age-and gender-balanced sample of US Americans via Prolific. Thus, it can be considered largely representative for the online US population of crowdworkers.

Sampling strategy

Convenience sample for forecasting teams of scientists. Stratified sample for lay people.

Scientists: 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 Table S1 for demographics).

Lay sample: 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%).

Data collection

Data from scientist teams and lay people was collected via the online Qualtrics survey platform. Participants had to upload a filled out Excel worksheet onto the Qualtrics platform, which connected to their unique survey link. Forecasting teams could fill out any of the 12 domains and were therefore aware of other domains. Forecasting teams did not know of other teams taking part in the initial tournament. Lay people were randomly assigned to 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).

Domains were presented in Qualtrics online in a randomized order. 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/>).

Timing

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 forecasted domains = 4.55; SD = 3.88).

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).

Data exclusions

Scientists: We included all submissions, as long as participants provided information about their rationales for their forecasts.

General Public Sample. We considered lay 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; Mage = 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).

Table S7 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.

Non-participation

no participant declined to participate.

Randomization

To maximize number of scientist submissions, forecasting teams could fill out any of the 12 domains and were therefore aware of other domains. Lay people were randomly assigned to 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).
Both for scientists and lay people, (selected) domains were presented in a randomized order on the Qualtrics submission form

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

- | n/a | Involvement in the study |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |

Methods

- | n/a | Involvement in the study |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |