nature human behaviour

Article https://doi.org/10.1038/s41562-022-01517-1

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

Table of Contents

Supplementary Methods

Mean Absolute Scaled Errors (MASE) as a Marker of Forecasting Accuracy. Conventional measures of forecasting accuracy for time series typically relies on Mean Absolute Error (MAE) across timepoint estimate. Hereby, a forecast "error" is the difference between an observed value and its forecast. Because forecast errors are on the same scale as the data, MAE is therefore scale-dependent, as well, and cannot be used to estimate accuracy *across* time series involving different units [1]. To circumvent the scale-dependency, Hyndman and Koehler [2] introduced scaled errors to afford comparability of forecast accuracy across series with different units. They proposed scaling the errors based on the training MAE from a simple forecast method representing a one-step naïve (random walk) forecast, resulting in Mean Absolute Scaled Error (MASE). By using this scaling factor as denominator, the resulting scores will be independent of the data scale:

$$MASE = \frac{MAE}{MAE_{in\ sample,naive}}$$

Here, a scaled error < 1 arises from a better forecast than the average one-step naïve forecast computed on the historical data. Conversely, a scaled error > 1 arises if the forecast is worse than the average one-step naïve forecast computed on the training data. Because MASE reflects a scaled version of the means in errors when comparing predictions against real trends, it is a standardized measure of mean-level error. Since its introduction, MASE has been adopted by most forecasting competitions [1,2]. Recent empirical forecasting competition also emphasize assessing prediction intervals along with point forecast estimates[1], employing metrics like MSIS and Cover rate to see how often the actual outcome falls within the estimated a priori prediction interval, as many decisions in real life setting are based in the intervals rather than the points forecasts. For pragmatic reasons, such prediction interval metrics were out of the scope of this specific research, but the interested reader can follow up the recent M4, M5, & M6 (ongoing) forecasting studies for further information.

MASE below 1.38 can be considered a threshold for a good forecast in the context of homogeneous trends from the tourism industry[2]. Results of a more heterogeneous M4 forecasting competition of 100,000 real-life time series suggest a threshold 1.89 for the out-of-sample naïve (random walk) statistical benchmark, a threshold of 1.70 for the Theta statistical benchmark, and a threshold of 1.77 when considering the median of all M4 forecasting teams[1]._We used estimates of median performance in M4 as one possible benchmark in our tournament, because M4 tournament focused on a range of heterogeneous domains, including forecasts for industries, services, tourism, imports & exports, demographics, education, labor & wage, government, households, bonds, stocks, insurances, loans, real estate, transportation, and natural resources & environment. Due to their heterogeneity and content, we assumed at least some generalizability to social issues in our forecasting tournament.

Comparison of MASE against an alternative scale-free metric of forecasting precision. In addition to MASE as a measure of forecasting *accuracy*, one can examine forecasting precision across domains on different scales by testing the correlation between predicted time series and observed time series. Like MASE, a correlation-based marker of *precision* is also standardized. Yet, it is conceptually different from MASE in several ways. First, whereas MASE concerns mean-level error, a correlation-based marker can account for raters being sensitive to the variability of the outcome across time points, as reflected in the variability of their predictions across those same time points. Second, whereas MASE takes historical time series into account when computing the scaling factor in denominator, a correlation measure does not. Further, by definition, correlation can only be applied when variance in time series exist (*SD* > 0); if a forecaster predicts no change in time series over time, it would be impossible to calculate the correlation. In exploratory analyses, we examined convergence between a correlation-based marker of forecasting precision and MASE marker of forecasting accuracy, as well as to estimate average correlation between predicted and observed time series along with effects of domain-general expertise.

In the First Tournament, multi-level correlation with participant and domain as random factors revealed a significant negative association between scaled mean-level errors (MASE) and correlation-based precision marker among social scientists, *r* = -.14, 95% *CI* [-.25, -.03], *t*(298) = 2.44, *P* = .015, and a non-significant trend in the same direction among the general public, *r* = -.05, 95% *CI* [-.10, .01], *t*(1458) = 1.75, *P* = .081. Similar correlations in the Second Tournament revealed a non-significant negative trend among social scientists, *r* = -.07, 95% *CI* [-.16, .02], *t*(460) = 145, *P* = .148. Evaluating central tendencies of associations between two markers across domains revealed a small-moderate negative effect among social scientists in the First Tournament, *M*(*r*) = -.16; *Md*(*r*) = - .20, a negligible negative effect among the general public in the First Tournament, *M*(*r*) = -.08; *Md*(*r*) = -.09, and a very small negative effect among the social scientists in the Second Tournament, *M*(*r*) = -.10; *Md*(*r*) = -.05. Overall, participants with lower MASE tended to be somewhat more sensitive to accuracy in variability across time series. However, the effect size of this association is modest, suggesting that the markers are largely distinct.

In the First Tournament, average correlation between predicted and observed time series was small for the lay crowd, *M* = .06, 95%*CI* [.04, .08], and statistically indistinguishable from zero among social scientists, *M* = .05, 95%*CI* [-.01, .10]. Similarly, in the Second Tournament, average correlation between predicted and observed time series was very small among social scientists, *M* = .07, 95%*CI* [.02, .13].

Moreover, social scientists did not significantly differ from the lay crowd, *F*(1,1683.6) = 0.84, *P* = .358, nor was this effect qualified by a significant domain X expertise interaction, *F*(11, 1397.9) = 1.65, *P* = .080. Post-hoc analyses for each domain revealed no significant differences between social scientists and the lay crowd, *t*s < 1.83, *P*s > .741, with one exception. For the domain of negative affect, lay people showed a positive correlation of predicted and observed time series, *M* = .15, 95%*CI* [.08, .21], whereas social scientists showed a *negative* correlation, *M* = -.12, 95%*CI* [-.27, .03], *t*(1719.4) = 3.61, *P* = .004.

Overall, using a theoretically and empirically distinct metric of forecasting accuracy (MASE) and precision (correlation-based marker), we confirm the general picture of negligible degree of forecasting accuracy among social scientists, as well as parallel results for lack of advantage for domain-general expertise for forecasting accuracy.

Distribution of Social Scientists' MASE scores in each Tournament. Figures S12-S13 show that in most domains social scientists' forecasts were right skewed. To account for the non-normal distributions, we transformed the MASE scores for linear mixed model analyses. Upon inspecting residuals from different models, we zeroed in on *log*(MASE) scores for analyses reported in Figure 1 and Table 1, back-transforming estimates (and 95% confidence intervals) for visualizations in Figure 1. Figures S12-S13 show estimated means from such linear mixed models in red crossed squares, overlayed on top of box-and-dot-plots: For most domain, these estimates were either very close to the median forecast or presented an overestimation of social scientists' performance (lower MASE score than suggested by the medians – 3 domains in Tournament 1 / 2 domains in Tournament 2). Only in two cases (implicit African American bias in Tournament 1, Republican support in Tournament 2) median MASE scores were slightly lower than estimated model parameters, but these differences were negligible. Consequently, the present analyses take non-normal distribution into account such that possible outliers have limited weight for the overall estimation of social scientists' performance.

Estimating naïve benchmarks. We estimated three naïve statistical benchmarks, as described below.

Benchmark 1: Random resampling of historical data. One potential approach to prediction is to simply assume that future observations will resemble previous observations, with no assumptions about additional temporal structure. This approach essentially assumes that the COVID-19 pandemic should result in little change in a measure compared to historical values. To capture this kind of naïve approach, we predicted future values by randomly resampling 12 values for Tournament 1 and 6 values for Tournament 2, sampled with replacement from the observed historical data points in each domain. To determine the expected distribution of predictions based on this method, we simulated 10,000 predictions (12/6 observations each) per domain. We then calculated MASE scores for each simulation. Estimates from this approach are equivalent to the historical mean for each domain, but also give a sense of the expected range around that mean (see Supplement for additional details how resampled estimates compare to an estimate obtained via averaging of historical trends).

Benchmark 2: Naïve Auto-regressive Random-Walk. How dependent are the conclusions above on the specific choice of naïve prediction? We computed predictions from several other approaches to prediction. For Benchmark 2, we asked about predictions made by a forecaster who assumes that monthly changes will resemble previously observed monthly changes. In contrast to the naïve random resampling approach (which assumes no temporal autocorrelation over time points), the random walk approach attempts to capture the intuition that temporally close data points are autocorrelated, and thus change less markedly than points that are temporally distant. Thus, a naïve forecaster might try to take advantage of this autocorrelation structure from previous data. Rather than assuming any sort of specific distribution for these changes, we simulated naïve predictions in which 12/6 changes were drawn randomly with replacement from previously observed values of change in that measure. These changes were then added one by one to the previous time point, starting with April/October of 2020, until a random path prediction for 12/6 months had been obtained. We then calculated MASE scores for each of 10,000 simulations using this method. Finally, we compared these simulated MASE scores to observed MASE scores of our expert sample. Note that the results were nearly identical when using a similar approach, closer to the mathematical random walk, where the monthly changes were drawn from a Gaussian distribution fitting the mean and standard deviation of changes in previous data. We thus focus here on the version that makes fewer assumptions and denote for simplicity this first naïve method "random walk."

Benchmark 3: Naïve Regression Based on Random Intervals. In the previous two approaches, we have assumed that a naïve prediction approach would use all the previous data at hand to predict future outcomes. However, an alternative strategy is to assume that change over a subset of time in the previous data might best resemble future change. For example, imagine a forecaster who believes that, since the upcoming months include a national election, then the period around the previous national election might provide the best window into how observations should change. Another forecaster might adopt a similar approach but assume instead that the relevant period for comparison dates back to a previous pandemic (e.g., the H1N1 outbreak). Thus, both forecasters might run a regression to determine how a particular measure changed over time from before to during the period of interest and use the slope of that change to extrapolate change in the current context. Of course, which forecaster is "correct" about the relevant comparison period is unclear, and it could be that selection of a time interval that resembles the current context could occur simply due to chance. To capture the naïve version of this approach, we simulated 10,000 predictions by first selecting a subset of the observed historical data, defined as a random continuous interval of $2 < n < 40$ time points. For each simulation, we performed a regression on the observed data, using time as the predictor, and calculated the slope of change. Finally, we used this slope to predict linear change over the upcoming 12/6 months, starting with April/October of 2020 as the intercept. We then calculated MASE scores for each of 10,000 simulated forecasts using this method and compared these simulated MASE scores to observed MASE scores of our expert sample.

Were forecasting teams wrong for the right reasons? We estimated the forecasting accuracy of the COVID-19 trajectory by evaluating MASE scores for COVID-19 cases and death against the actual number of cases and deaths. We use these conditional forecasting accuracy scores to evaluate accuracy of each of the targeted domains – a significant association between inaccuracy of predicted COVID-19 trajectory and societal predictions would speak to the possibility of participants being wrong for right reasons. Because both prediction inaccuracy (MASE) for COVID-19 and societal change were skewed, in each case we applied a log-transformation. We also included the domain as a set of dummy-covariates. The results showed no significant association between accuracy of COVID predictions and accuracy of predictions of societal change, B = -0.59, SE = 0.37, $t(14.75)$ = 1.61, P = .130.

Multiverse analyses of domain-general accuracy. Our analyses comparing forecasting accuracy of social scientists and the general public yield little overall evidence of domain-general expertise affording greater accuracy. To interpret these findings, we used multiple methods as robustness checks to confirm these results. Here, we outline our analytic approach.

Linear mixed-effect models (LME). Scientists were invited to make predictions about any number of the 12 forecasting domains, which resulted in a nested (and somewhat imbalanced) LME design, with some teams making predictions for a few domains, others for many. Due to the imbalance, we examined the full model by considering both main effects and the interaction term between domain and expertise. Additionally, we performed linear mixed model analyses with *lme4* package [3] with main effects of domain and expertise only, thereby treating domain as a dummy-coded set of covariates. These analyses resulted in slightly different results, with the question of possible differences between the scientist and lay groups depending on the inclusion of this interaction term in the full model. Specifically, in contrast to the full model results reported in the main text, results of analyses without accounting for the domain × expertise (lay crowd/scientist) interaction reveal a main effect of social scientists performing better than their lay counterparts, $F(1, 628) = 17.67$, $P < .001$, $R^2 = 0.013$. As reported in the main text, results of a full model revealed no main effect of expertise, but a significant interaction, $F(11, 1304) = 2.00$, $P = .026$. Simple effects show that social scientists were significantly more accurate than lay people when forecasting change in life satisfaction, political polarization, and both implicit and explicit gender-career bias. However, the scientific teams were no better in the remaining eight out of twelve domains (Table 1 and Fig. S2).

Bayes Factors. We computed Bayes Factor approximations using *rstanarm* [4] package in *R* [5]. Following guidelines [8], we relied on weakly informative priors[6] for our linear mixed model and used *emmeans* [7] and *bayestestR* [8] packages to obtain Bayes Factors for individual marginal means of social scientists and lay people. Only a single domain – life satisfaction – showed strong evidence in support of greater accuracy of scientists' predictions compared to lay predictions, $BF = 22.72$ (see Table 1 in main text) and eight domains in support of the null hypothesis, where there was moderate to strong evidence that there were no differences between the predictive accuracy of scientist and lay people, $BF \leq 0.12$. For the remaining three domains, there was either weak evidence in support of the difference (explicit and implicit gender career bias), or too little evidence of support for either hypothesis (political polarization).

Interpretation. These analyses reveal consistent results across different analytic approaches. There is support for greater accuracy of scientists' predictions compared to lay people in some domains, but we caution the reader that there is little evidence in support of a difference between social scientists and lay people in most other domains.

Rationale for testing performance against three naïve benchmarks. We sought to test if social scientists perform better than the three naïve benchmarks. To this end, we examined performance against each benchmark (Figure 3 in the main text). To reduce the likelihood that social scientists' forecasts beat naïve benchmarks by chance, our main analyses examine if scientists performed better than each of the three benchmarks. Further, we adjusted CI estimates for 12 domains in each tournament for simultaneous inference by simulating a multivariate *t* distribution[9]. We focused on performance across all three benchmarks in the spirit of quality control tests across a heterogeneous set of standards. As an analogy, consider three distinct benchmarks testing performance of a nuclear reactor. When such benchmarks probe against different aspects of overall performance, it would not be informative to know whether the reactor passes such tests on average—even if one of the tests is not passed, it may still lead to a nuclear meltdown. Additionally, if the number of tests is small and the likelihood of heterogeneity is high (which is the case when tests aim to examine different aspects of performance), measures of central tendency may be simply unreliable.

We selected the three benchmarks based on their prior use in forecasting and because they target distinct features of performance (historical mean addresses base rate sensitivity, linear regression speaks to the sensitivity to the overall trend, **and random walk captures** random fluctuations along consecutive time points). Each of these benchmarks may perform better in some but not other circumstances. Consequently, to test the limits in scientists' performance, we examine if performance is better than each of the three benchmarks.

We recommend interested readers to start by examining whether scientists passed all three quality checks – if the performance was superior to all three benchmarks (Figures 1 and 3) and subsequently inspect performance against individual benchmarks (see Figures 3 and S2). Though we do not believe inspection of averages of benchmark MASE scores adds on informational value, we leave it up to the educated reader to draw their inferences. As Figure S15 shows, in the First Tournament, for 9 out of 12 domains scientist forecasts 95% CI included the naïve statistic average. In the Second Tournament, this was the case for 5 domains. These inferences are not that different from what we find if we examine if scientists performed better than all three benchmarks (T1: 10 out of 12; T2: 7 out of 12), which we report in the main text.

In our view, the main take home message from relevant analyses paints a similar picture to the one obtained when testing performance against the best of the three benchmarks: Social scientists' forecasts are hardly better than naïve methods such as historical mean, random walk, or a linear regression.

Raw forecasts and ground truth markers. Figure S1 shows forecasts from both tournaments along with aggregated estimates, ground truth markers (i.e., measured societal change), and the last three historical data points preceding the start of the tournaments (also see Figure 2 in the main text with the whole historical time series forecasting teams received and the lowess estimate of the naïve crowd). As an example, consider the featured domain of negative affect on social media. As Fig. S1 shows, average forecasts differed widely from the observed estimates in negative affect over time: whereas the ground truth markers show a substantial swing of over 1 *SD* over the period of a year, average forecasts were less volatile over the same timeframe. Notably, both at the beginning of the first and the second tournament, the initial forecasts are in line with the ground truth and start to diverge over time.

Overall, Fig. S1 and Fig. 2 in the main text show large variability in forecasts among scientists' teams in each domain. Though in some domains the average of the forecasts appeared close to the ground truth markers (e.g., gender stereotypes, life satisfaction), in many other domains forecasts were less accurate (e.g., negative affective sentiment on social media or ideological support for political parties). Additionally, for half of the domains the average forecasts were highly similar or nearly indistinguishable from the last historical data points (January-March) provided to forecasting teams.

Comparison of two naïve benchmarks: Resampled historical mean versus averaging of historical data. We used resampling of historical data as one of our three naïve statistical benchmarks. We randomly drew forecasting estimates for each of the 12 (Tournament 1) / 6 (Tournament 2) months from the past historical data participants received for respective domains and calculated MASE score for each, repeating this procedure 10,000 times. Our estimates were the averages of MASE scores obtained from resampling. Because of resampling, this approach can also provide an expected range around the mean. An alternative approach would be to average historical data for a given domain, and then estimate the MASE scores from this averaged historical prediction. This type of benchmark cannot capture the variability that one might expect due to chance from analyzing many teams making predictions in different ways, but it has the advantage of being extremely simple to calculate. As Table S8 demonstrates for Tournament 1 data, for each domain resampling produces almost identical predictions to the averaging of historical data. Further, for most domains, the MASE scores difference from the resampling mean and simple historical mean approaches was close to identical, too. We do note two domains where this pattern of MASE scores slightly diverged: for explicit gender bias and life satisfaction, the resampling approach produces noticeably less accurate forecasts (i.e., higher MASE scores) compared to the mean of the historical data, despite producing nearly identical mean predictions. Thus, similar performance of scientists to the resampled historical mean for explicit gender bias and poorer performance compared to resampled mean benchmark for life satisfaction (see Figure 3 and Figure S2) can be considered underestimations: For these domains, scientists on average did worse than one would have done by taking a historical average of the last 12 data points. Overall, resampling and averaging historical data yield close to identical benchmarks, and very similar degrees of forecasting errors. In domains where differences in forecasting errors were larger, resampling produced an underestimation of the benchmark performance and overestimation of scientists' relative performance compared to this benchmark.

Societal change over the pandemic. We examined if societal change trends in each domain were unusual during the time immediately following the first pandemic lockdowns in the US. To this end, we computed the mean absolute difference between observations for the May-Oct 2020 6-month period (the first six months of the Tournament 1) and the average of the last 3 time points of the historical data (Jan-March 2020). We then repeated this procedure for the next 6 months period (Nov 2020-Apr 2021), again computing the difference from the average of the last 3 time points of the historical data (see Fig. 3 in the main text for historical trends). The results showed that the absolute difference was significantly higher across domains for the initial 6-month period than for the subsequent 6-month period, paired *t*-test across domains for the two time points ($df = 11$) = 2.474, $P = .03$.

Top scorers and their strategies. Who won the tournaments? As Tables S2-S3 show, Tournament 1 winners were not necessarily Tournament 2 winners, except for two overlaps (**lmielin**, and **AbCdEfG)**. "Goodness" of forecasts is not an abstract entity and depends on the number of time points and the domain. Still, we can use 1.38 as a rough benchmark suggested by Athanasopoulos and Hyndman [2] or the M4-based benchmark[10] of 1.77 (the median teams in the tournament) [3], keeping in mind that there is no such thing as a universal threshold of what constitutes a good forecast.

In Tournament 1 ($N = 86$), we see that 8 domains have better (lower) scores than the Hyndman benchmark, and 9 domains had better (lower) scores than the M4 benchmark. Indeed, 4 of them – explicit and implicit gender-career bias, positive affect, and political polarization are < 1, i.e., these out of sample predictions perform better than in-sample naïve forecast (random walk) used as a scaling factor denominator. Four domains were relatively hard to predict: ideological support for Republicans, negative affect, explicit and implicit African American bias. In the Tournament 2 ($N = 120$), top predictions for all domains but one (ideological support for Republicans) did better than the in-sample naïve random walk forecast and better than benchmarks derived from prior forecasting competitions.

Overlap in top scorers across tournaments and domains. Except for one team, the top forecasting teams from the first tournament did not appear among the winners of the same domains in the second tournament. Expanding the scope to top five forecasting teams, only in five out of 12 domains one top team from the first tournament appeared among the top five teams of a given domain in the second tournament: **Compassionate Values** for Explicit African American bias; **fearfulastra** for Explicit Gender-Career bias; **FMTeam** for Implicit Asian American bias; **AbCdEfG** for Ideological Support of Democrats; **A Woman Scientist** for Negative Sentiment; **NYHC** for political polarization. The remaining top five teams were unique across tournaments.

Some top five performers in one domain also performed well in other domains, First Tournament: $n = 14$; Second Tournament: $n = 17$. However, most of these repeats occurred only in one other domain, First Tournament: $M = 1.62$; Second Tournament: $M = 1.67$. In the First Tournament, 5 teams appeared in three top five domains, and 2 teams appeared in four top five domains. However, most of these teams also made predictions for a large number of domains (*M*

= 7.75, *SD* = 3.94), thus their relative success across domains may be due to chance. Indeed, only one team among those who were among the top five in more than 2 domains had a reasonably small number of domains they made predictions for, such that they were in the top five in 4 out of 6 domains (67%). For the other top five teams with success in more than two domains, number of hits (top five placements) were below half of the domains they were making predictions for.

In the Second Tournament, two teams appeared in three top five domains, and one team appeared in the four out out of five domains. Again, most of these teams also made predictions for many domains (*M* = 8.50, *SD* = 3.54), thus their relative success across domains may be due to chance. Indeed, only one team among those who were among the top five in more than 2 domains had a reasonably small number of domains they made predictions for, such that they were in the top five in 5 out of 9 domains (55.56%). For other top five teams with success in more than two domains, number of hits (top five placements) were at or below half of the domains they were making predictions for.

Modeling strategies and rationales in Tournament 1. The domains below are ranked from the most to least accurate domains, precising the characteristics of the best forecasting methods used in each domain:

1. *explicit gender bias* (Time-series regression with monthly adjustments)
2. *implicit gender bias* (assumption of steady/slow change in societal phenomena, linear interpolation from prior years, focus on -long-term trend, adjustment based on the domain and possibility of a domain-specific plateau)
3. *positive affect on social media* (intuition-based forecast, considering covid deaths, unemployment, taking care of children, home-office, political changes, social changes in communication & groups prejudice, health problems connected with medical treatment, generalized fear & anxiety)
4. *political polarization* (an ARIMA model, assumption that polarization of opinions increases before decision-making and decreases immediately after - but before the outcomes of the consequences of the decision have the time to unfold)
5. *life satisfaction* (best was intuition/guess without any information; second/third best were data driven using insights about general trends in the past or mean reversion)
6. *explicit Asian American bias* (assumption of steady/slow change in societal phenomena, linear interpolation from prior years, focus on -long-term trend, adjustment based on the domain and possibility of a domain-specific plateau)
7. *Implicit Asian American bias* (data-driven; didn't believe in strong effects on implicit Asian American bias since we doubt the reliability and validity of the measure. Tested seasonal monthly effects using cyclical p-spline in a GAM model. After rejecting the model, defaulted to a model using a constant value across the whole range)
8. *Democratic support -congressional ballot* (data-driven; Vector autoregression with a constant and one lag term; Used the lower and upper end of the predictions to differentiate more between Republicans and Democrats assuming that the current affairs were going to turn out to be relevant)
9. *Republican support - congressional ballot* (theory-based; Previous research has shown a tendency to gravitate towards conservatism when faced with system threatening events, be it terrorist attacks or a pandemic (Jost et al., 2017; Economou & Kollias, 2015; Berrebi & Klor,

2008; Canetti-Nisim et al., 2009; Schaller, 2015; Beall et al., 2016; Schaller et al., 2017). With this in mind, we believe that the COVID-19 outbreak, as a system-threatening event, will precipitate a shift towards supporting the Republican party. In addition, relative deprivation, the perception that one's self or group does not receive valued resources, goals, or standards of living (Kunst & Obaidi, 2020) might play a role in ideological preferences. Assuming that Americans value their economic resources, and that they perceive themselves to be victims of their country's challenging economic circumstances, this relative deprivation during COVID-19 may trigger supporting the Republican party to counter Americans lost economic resources. The Republican party, and not the Democratic party, could be more supported because it is perceived as the party that owns the economy.)

10. *Negative affect on social media* (intuition-based forecast, considering covid deaths, unemployment, taking care of children, home-office, political changes, social changes in communication & groups prejudice, health problems connected with medical treatment, generalized fear & anxiety; second/third were also theory-based)
11. *Explicit African American bias* (data-driven; Time-series Regression with Monthly Adjustments)
12. *Implicit African American bias* (data-driven; Time series regression with monthly adjustments)

In total, **among the top 5 teams per domain**, 62% were data-driven, 30% were based on intuition/theory, and 8% were hybrid. As Fig. S4 shows, by domain, intuition/theory dominated among top performers for positive affect, polarization, and negative affect, whereas data-driven models dominated for everything else (except for implicit African American bias where it was evenly split). Only for explicit gender bias, *out of sample* forecasts by top teams for implicit gender bias, polarization, and positive affect were as good/better than *in-sample* naïve forecasting models (red dotted line).

Benchmarking top teams against native statistical methods in Tournament 1. Fig. S5 shows comparison of top forecasts to *out of sample* naïve forecasts (linear regression and random walk). Results in this figure demonstrate whether the top 5 forecasts in each domain had higher error (blue / triangle) or lower error (orange / circle) than naïve (linear regression / random walk) estimators. By sampling the top five teams, we can evaluate if these teams are *consistently* more accurate (i.e., below the benchmark) than naïve estimators. Thus, we count performance as on-par or inferior to a naïve estimator if at least one of the top five teams produces more error than the benchmark. Following this analytic rationale, linear regression was better than at least one of the top 5 teams for half of the domains, whereas out-of-sample random walk was worse than all teams.

Disciplinary orientation of top teams in Tournament 1. Fig. S6 shows that data scientists and social scientists were most prevalent among top 5 teams (total: 38% social science, and 30% data/CS), whereas 15% were multidisciplinary. Social scientists dominated for political forecasts and forecasts of positive affect, whereas data scientists dominated among top forecasts of ethnic and gender bias. Behavioral (incl. decision-) scientists dominated among top scorers for negative affect and life satisfaction.

Forecasting experience among top teams in Tournament 1. Most of the top teams in Tournament 1 (78%) did not have prior forecasting experience. But note that for base rate (all submissions in Tournament 1), 83% did not have prior experience. In other words, the top 5 teams had a 4% greater chance of having a prior forecasting experience compared to the base rate. Exceptions where majority of top performers have prior tournament experience: explicit gender bias, implicit Asian-American bias. For life satisfaction, 2 out of 5 teams had prior experience with tournaments.

Modeling strategies and rationales in Tournament 2. The domains below are ranked from the most to least accurate domains, precising the characteristics of the best forecasting methods used in each domain:

1. *Implicit African American bias* - (data-driven; a statistical model to leverage autocorrelation in the data. Specifically, each estimated value is the weighted average of the previous five values of the same variable, where the weights are obtained through an ordinary least squares regression in which the current value is predicted from the five previous values. We further sought to leverage the correlations among the variables. To do this, we had each variable's final [model-based] prediction result from a regularized regression (specifically a ridge regression) where each variable's predicted value from the OLS described above is used to predict the current value of each variable.)
2. *Positive affect on social media* - (intuition-based forecast, considering Biden winning election; US Presidential inauguration; Quieting of Trump rhetoric; COVID infections; COVID deaths; Family separation over the winter holidays; Vaccine development; Vaccine distribution; Lifting lockdowns; Return to normal; life Demographics of twitter users)
3. *Implicit Asian American bias* - (hybrid; Critically, the measure used by PI is not a measure of generalized bias, but "Americanness" bias (or associating European American faces with domestic / not foreign concepts more easily than Asian American faces). We have previously conducted analysis on nearly a decade of Implicit Americanness Bias data from PI and have found that trends are generally conceivable as linear over sufficient time horizons. We thus assume that trends will eventually be linear. We layer in our theory about how trends will shift in the short term as conservative media entities lose incentive to depict COVID-19 as the "China Virus" and use like terminology that suggests China is responsible for the virus)
4. *Explicit African-American bias* - (hybrid; I compared my estimates for the past 6 months with the actual data during that period and updated my estimates. I also used the past data to make the new estimates. Following discussions on Twitter, it seems to me that expressing negative attitudes toward African Americans is even less socially acceptable than before the death of George Floyd)
5. *Life satisfaction* - (data-driven forecast; My predictions were simple and based on the idea that the past is a good prediction of the future. I took the current satisfaction during the pandemic and predicted that it would slowly shift back to the previous mean satisfaction prior to the pandemic. Essentially, I took the current satisfaction and predicted that it would slowly shift back to the previous mean satisfaction)
6. *Implicit gender bias* – (data-driven; time series analysis - Holt Winters Exponential Smoothing)
7. *Explicit Asian American bias* - (intuition-based forecast, considering covid deaths)
8. *Explicit gender bias* - (data-driven; a statistical model to leverage autocorrelation in the data. Specifically, each estimated value is the weighted average of the previous five values of the same variable, where the weights are obtained through an OLS regression in which the current value is predicted from the five previous values. Leveraged the correlations among the variables: each variable's final [model-based] prediction result from a regularized regression (specifically a ridge regression) where each variable's predicted value from the OLS described above is used to predict the current value of each variable)

9. *Democratic support -congressional ballot* - (data; I revised my earlier estimates by looking at how much I missed in the last six months (overlap period). I calculated the average error and corrected with this factor the previous estimates. After six months, I just keep my estimate constant.)
10. *Political polarization* - (data-driven forecast, employed a standard LSTM model to predict the values for the next 12-months. The model uses data from the past 4 months to predict values for the next 4 months, and we iterate this process to obtain the forecasts for the next 12 months (i.e., forecasts of the first 4 months are used as inputs for the next 4 months, and so on)
11. *Negative affect on social media* - (data-driven forecast, considering regression to the mean and introduction of COVID-19 vaccines)
12. *Republican support - congressional ballot* - (data-driven; "We applied an auto-regressive integrated moving average (ARIMA) model. Before modeling the time series, we plotted the data, checked for the need to stabilize the variance, and automatically applied Box-Cox transformation (with and adjusted back-transformation to produce mean forecasts) if so. Then, we employed a variation of the Hyndman-Khandakar algorithm, implemented in the forecast::auto.arima() function. After estimating and selecting the best-fitting model (based on AICs), we checked for the autocorrelation of the residuals by plotting the ACF plot and the portmanteau test whether the residuals are consistent with white noise (randomness). In case the residuals showed a significant pattern of autocorrelation, we went on to determine the ARIMA model parameters and selecting a better model manually. If the autocorrelation of residuals test was non-significant at alpha = .05, we calculated numerical point-estimate forecasts for the next 12 months. We did not further adjust these estimates in any way. Given theoretical considerations and the examination of past data, we assumed stationarity (lack of prolonged time trends) and restricted the model search to non-seasonal models.[…] stationarity (including lack of seasonality) - we consider this variable a relatively stable population characteristic (at least in the short run of 12 months), that did not seem to be markedly affected by the rather heated 2020 U.S. presidential election campaigns.)

In total, among **the top 5 teams per domain**, 65% were data-driven, 28% were based on intuition/theory, and 7% were hybrid. By domain, intuition/theory dominated among top performers for positive affect and negative affect, whereas data-driven models dominated for all other domains (Fig. S8). For most domains, *out of sample* top forecasts were as good/better than *in-sample* naïve random walk (below red dotted line); exception – Republican support.

Benchmarking top teams against native statistical methods in Tournament 2. Fig. S9 above shows comparison of top forecasts to *out of sample* naïve forecasts (linear regression and random walk). By sampling the top five teams, we can evaluate if these teams are *consistently* more accurate (i.e., below the benchmark) than naïve estimators. Thus, we count performance as on-par or inferior to a naïve estimator if at least one of the top five teams produces more error than the benchmark. Following this analytic rationale, linear regression was better than at least one of the top 5 teams for 3 out of 12 domains, whereas in all domains random walk was worse than the top 5 teams.

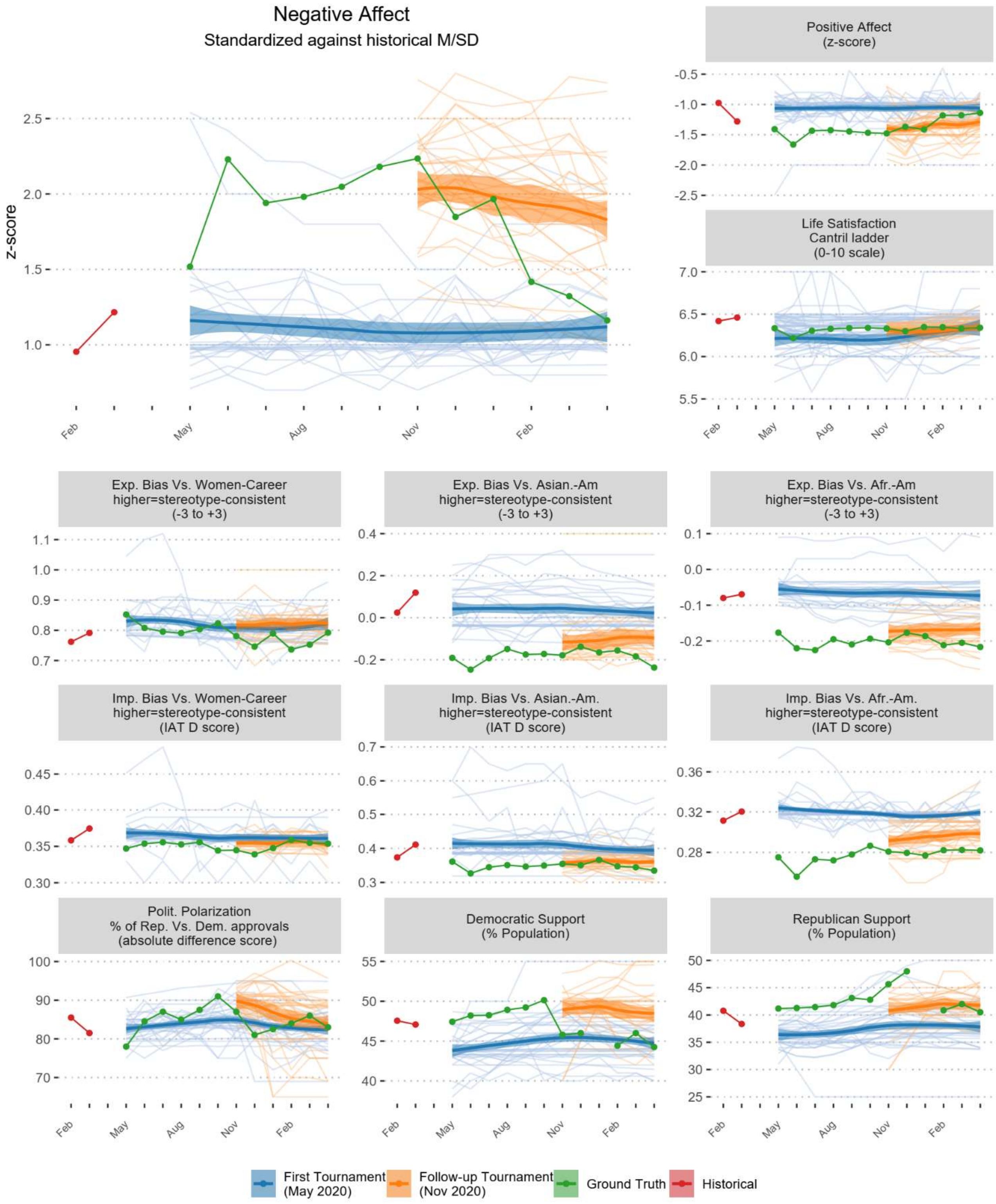
Disciplinary orientation of top teams in Tournament 2. Fig. S10 shows that behavioral/decision scientists and social scientists were most prevalent among the top 5 teams (total: 52% social science & 18% behavioral science), whereas 15% were multidisciplinary, and 12% were data scientists. Social scientists dominated most domains except for negative affect (dominated by behavioral scientists), & polarization (dominated by multidisciplinary teams).

Forecasting experience among top teams in Tournament 2. Most (78%) top teams did not have prior forecasting experience. But note that the base rate (all submissions in Tournament 2), 85% did not have prior experience. In other words, among the top 5 teams, we observe 6.65% more chance of having prior forecasting experience compared to the base rate. Also, at least one of the top performing teams had prior experience in most domains.

Code review process. Members of the Forecasting Collaborative performed an extensive code review during the preparation of the manuscript to assess the reproducibility of the code and cross-validation. Two members volunteered for the code review. The two code reviewers downloaded and ran the data and code in the "R Markdown file of the main analyses" shared through the GitHub repository of the project. The code reviewers carried out the code check on their own computers (i.e., using different operating system environments from where the code was initially developed), and let the main authors know that the code was possible to run, and the code did carry out the intended analyses. The shared code was divided into sections that correspond to the statistical output and visualization in the manuscript. If the code reviewers spotted an error, they were instructed to take a note of the line and then write a short description that explains what the potential issues are. Any typos and coding errors were reported to the main author prior to manuscript submission. The short description of the errors was emailed to the main authors and the code updates were pushed through GitHub. When needed, the code reviewers also made efforts to improve the readability of the code by breaking up long lines of code and adding comments.

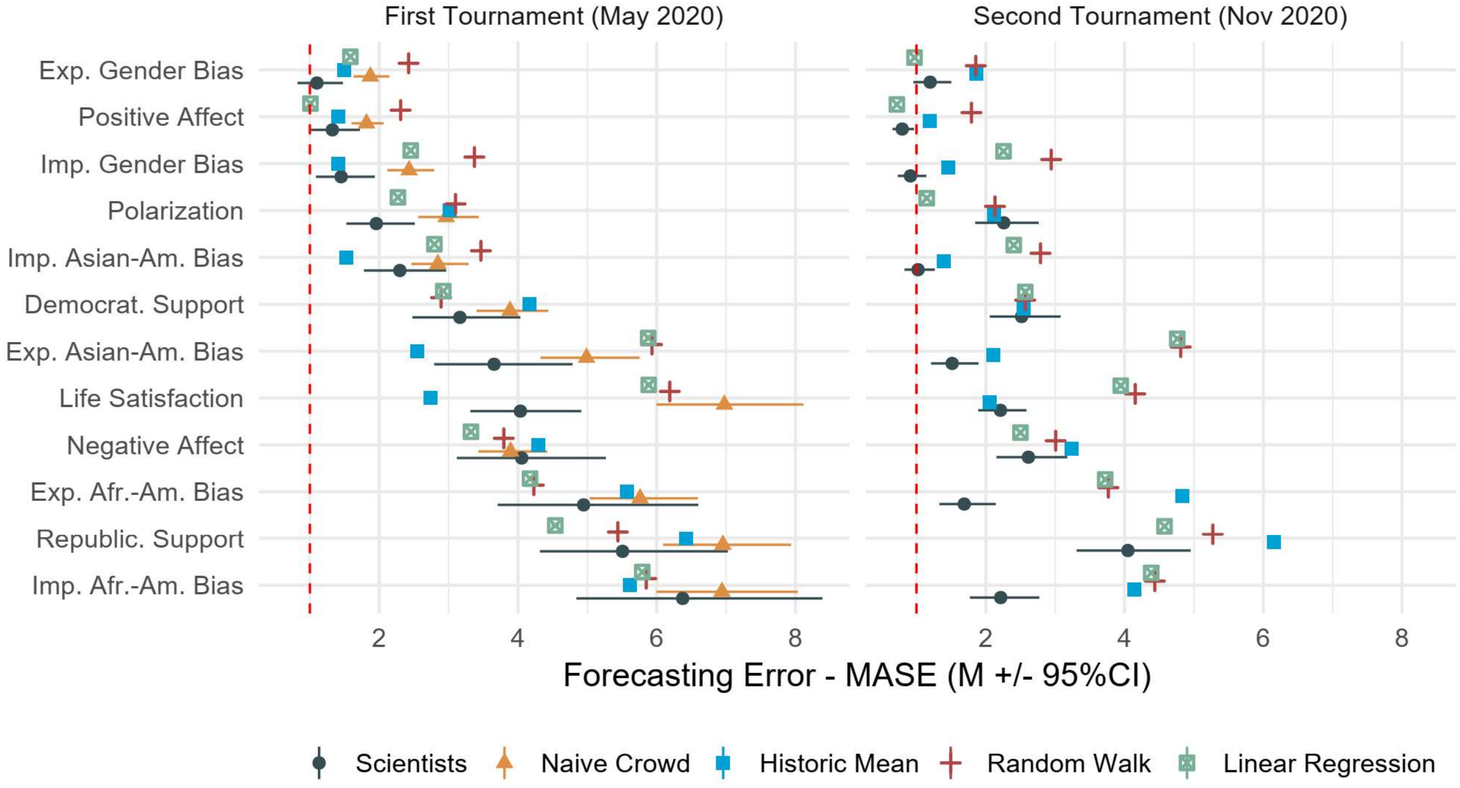
Reproducibility. Reproducible code is available in the GitHub repository of the project. See Table S9 for the table of content (summary of outputs) for the main R Markdown analyses.

Supplementary Figure 1. Forecasted models in the tournaments, and ground truth markers



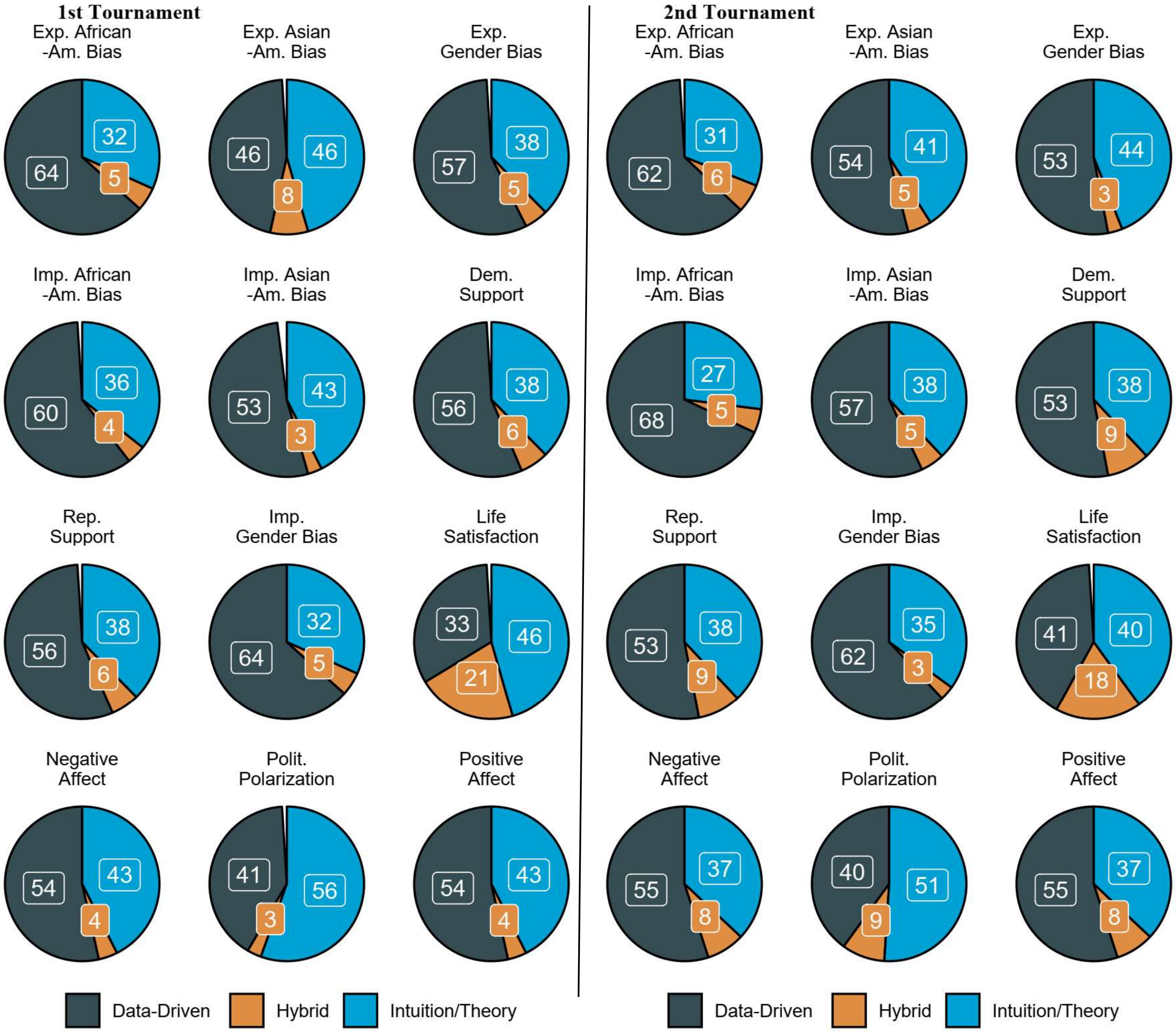
Forecasted models in the 12th and 6th-months tournaments, and ground truth markers, along with the last two historical data points teams received. Blue [orange] lines = forecasts in the May [November] 2020 Tournament. Aggregate trends (+/- 95% confidence band) obtained via a loess estimator. Negative affect is highlighted for visualization purposes.

Supplementary Figure 2. Average forecasting error, compared against benchmarks



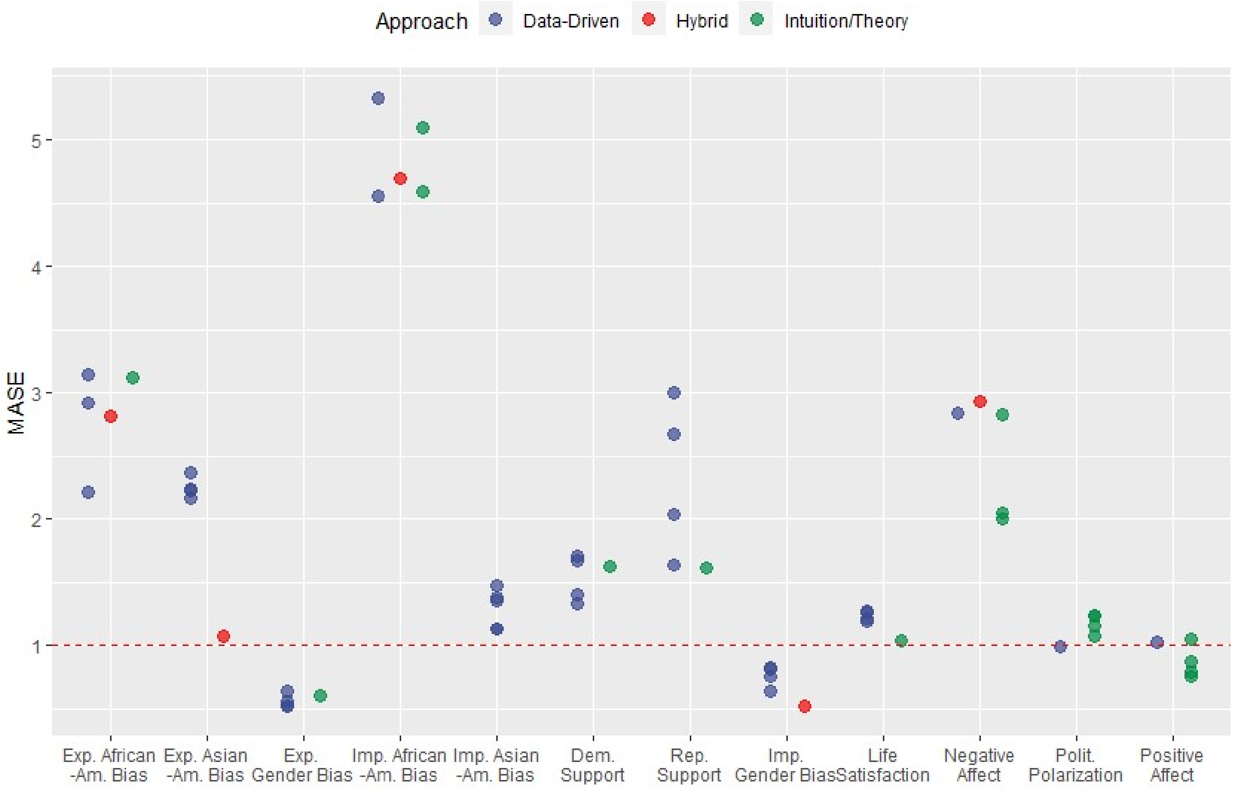
Average forecasting error, compared against benchmarks. We rank domains from least to most error in Tournament 1, assessing forecasting errors via mean absolute scaled error (MASE). Estimated means for Scientists and Naïve Crowd indicate the fixed effect coefficients of a linear mixed model with domain (k = 12) and group (in Tournament 1: $n_{\text{scientists}}$ = 86, $n_{\text{naïve crowd}}$ = 802; only scientists in Tournament 2: n = 120) as a predictor of forecasting error (MASE) scores nested in teams (Tournament 1 observations: $n_{\text{scientists}}$ = 359, $n_{\text{naïve crowd}}$ = 1467; Tournament 2 observations: n = 546), using restricted maximum likelihood estimation. To correct for right skew, we used log-transformed MASE scores, which are subsequently back-transformed when calculating estimated means and 95% confidence intervals. In each tournament, confidence intervals are adjusted for simultaneous inference of estimates for 12 domains in each tournament by simulating a multivariate t distribution. Benchmarks represent the historic mean, average random walk with an autoregressive lag of one, linear regression, and naïve crowd. Statistical benchmarks obtained via simulations (k = 10,000) with resampling. Dashed vertical line represents MASE = 1, with lower scores reflecting better performance than naïve in-sample random walk.

Supplementary Figure 3. Distribution of forecasting methods by domain



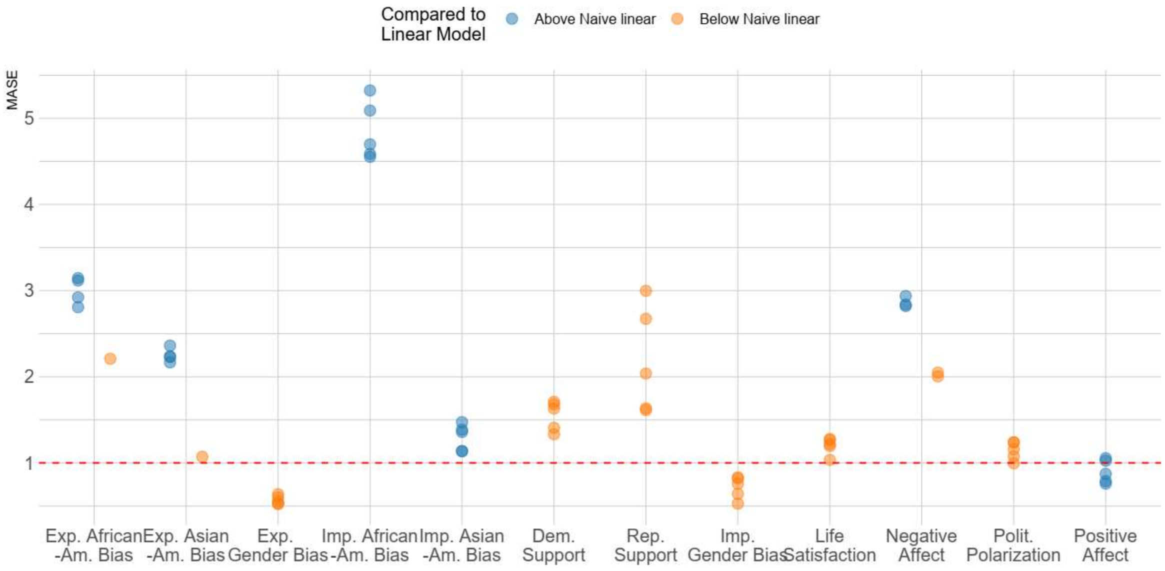
Distribution of forecasting methods -- data-driven / pure theory / hybrid (in %) – by domain

Supplementary Figure 4. Error among top 5 teams in each domain in Tournament 1, by approach



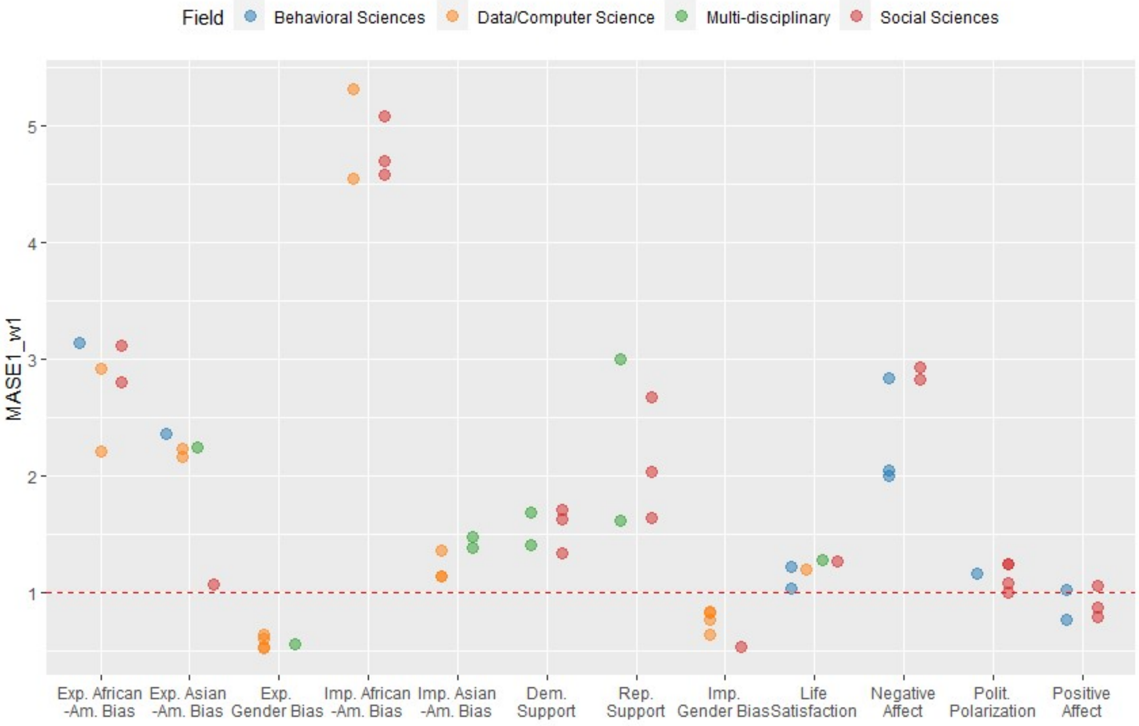
Error among top 5 teams in each domain in Tournament 1, by approach.

Supplementary Figure 5. Error among top 5 teams in Tournament 1 versus statistical benchmarks



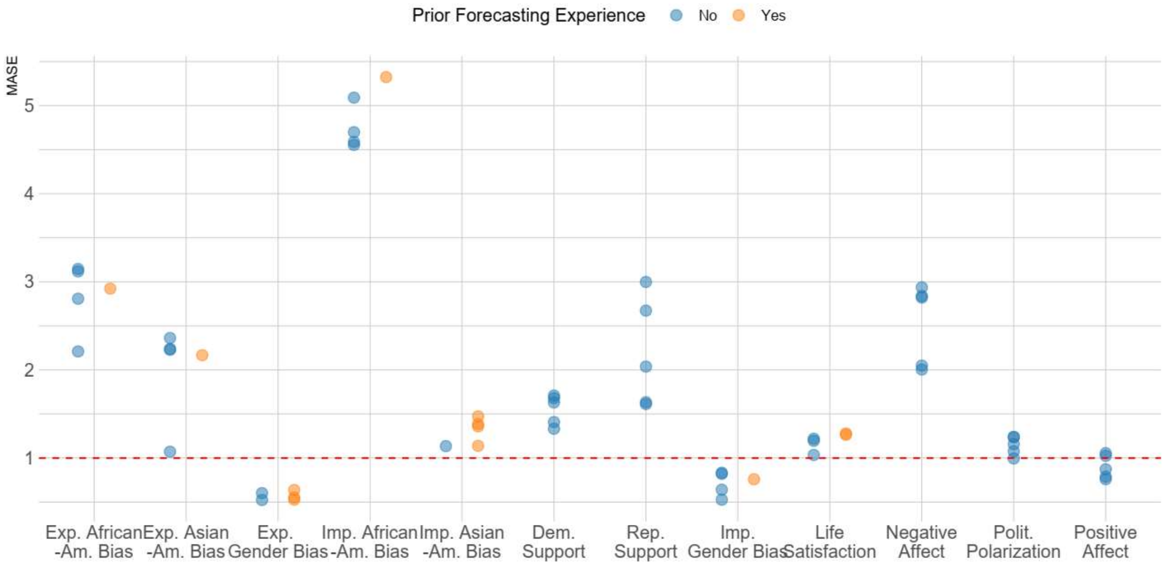
Error among top 5 teams in each domain in Tournament 1, contrasted with one naïve out of sample benchmarks - linear regression. All top five teams performed better than out of sample random walk.

Supplementary Figure 6. Disciplinary affiliation of top 5 teams in each domain in Tournament 1



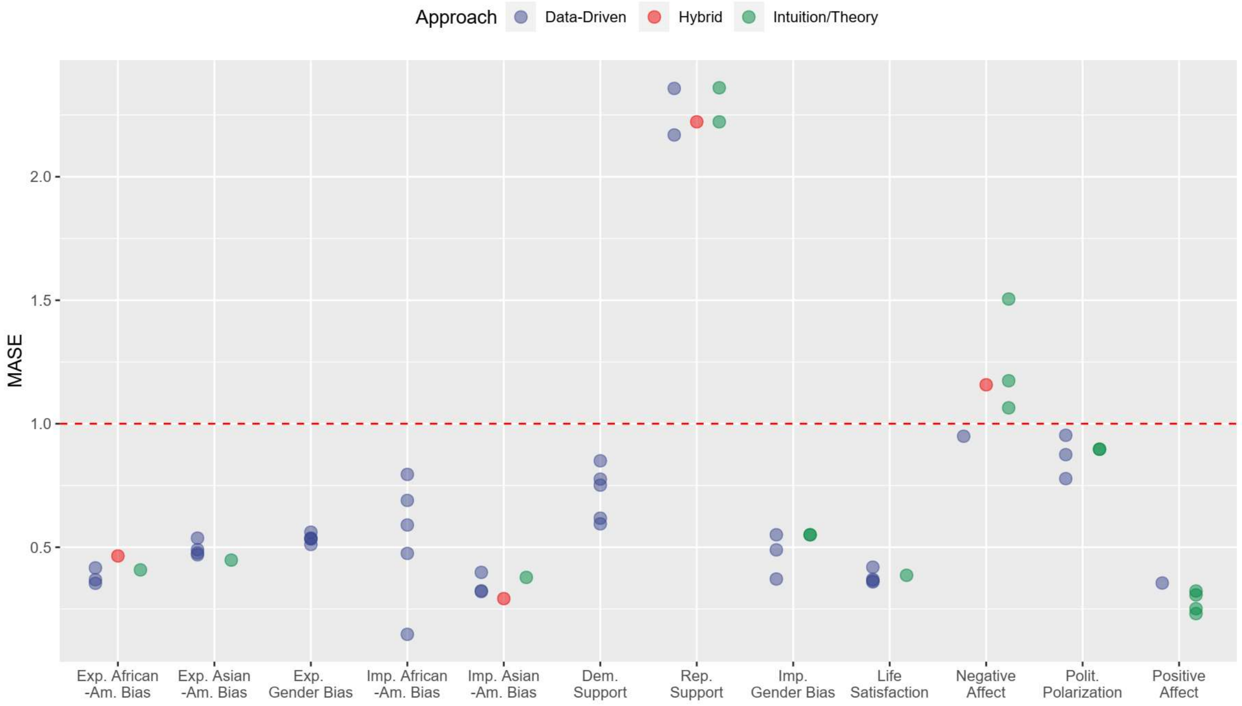
Disciplinary affiliation of top 5 teams in each domain in Tournament 1.

Supplementary Figure 7. Prior forecasting experience among the top 5 teams in each domain in Tournament 1



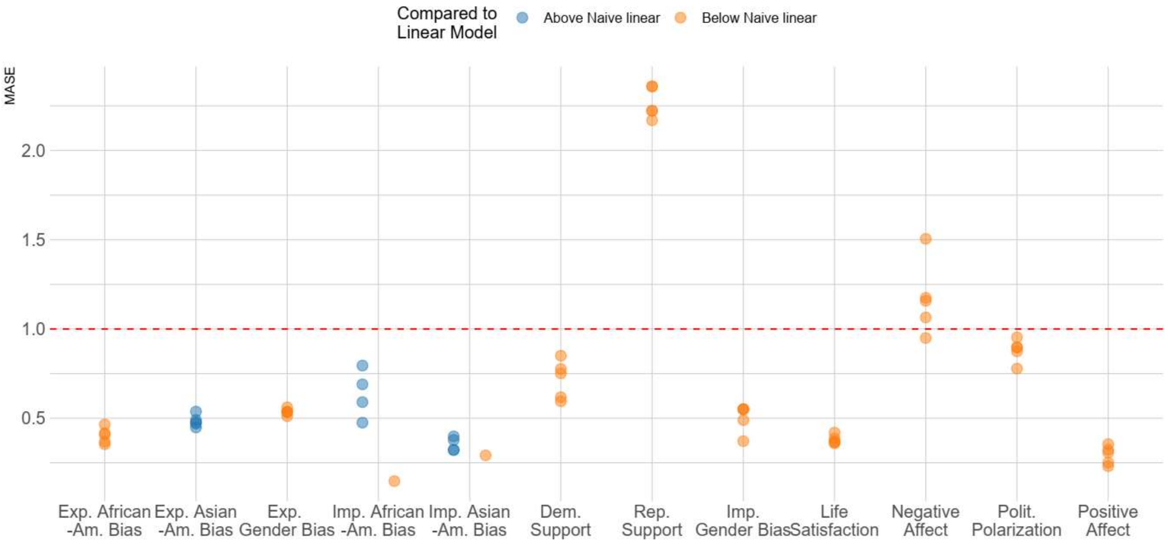
Prior forecasting experience of top 5 teams in each domain in Tournament 1.

Supplementary Figure 8. Error among top 5 teams in each domain in Tournament 2, by approach



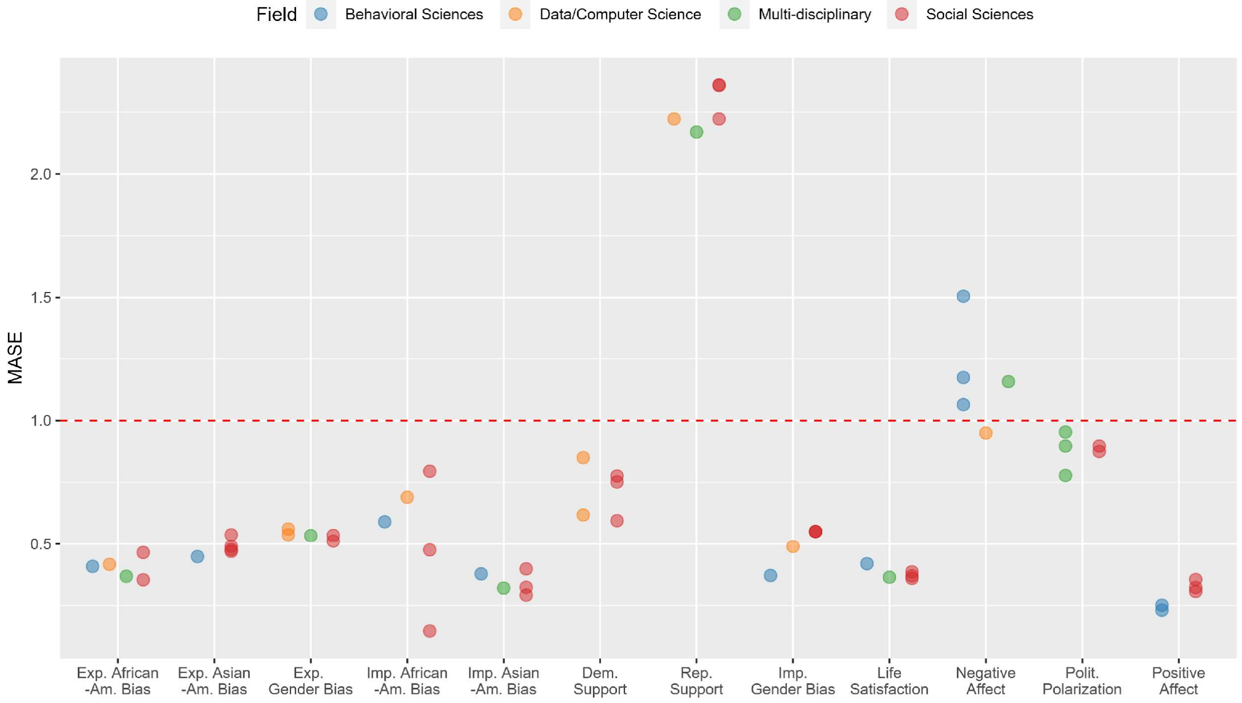
Error among top 5 teams in each domain in Tournament 2, by approach.

Supplementary Figure 9. Error among top 5 teams in Tournament 2 versus statistical benchmarks



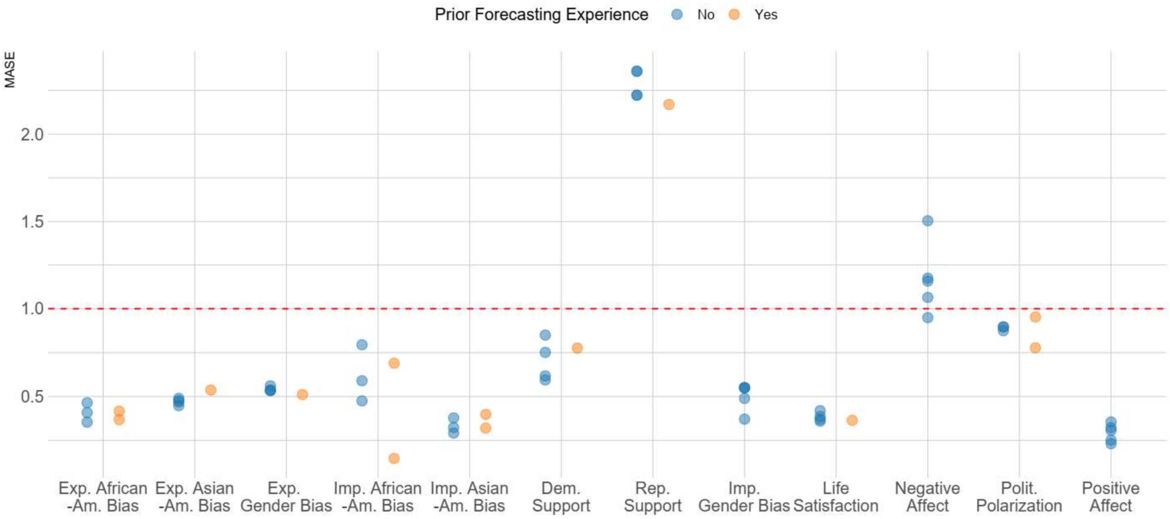
Error among top 5 teams in each domain in Tournament 2, contrasted with the naïve benchmark - linear regression. All five teams performed better than the out of sample random walk.

Supplementary Figure 10. Disciplinary affiliation of top 5 teams in each domain in Tournament 2



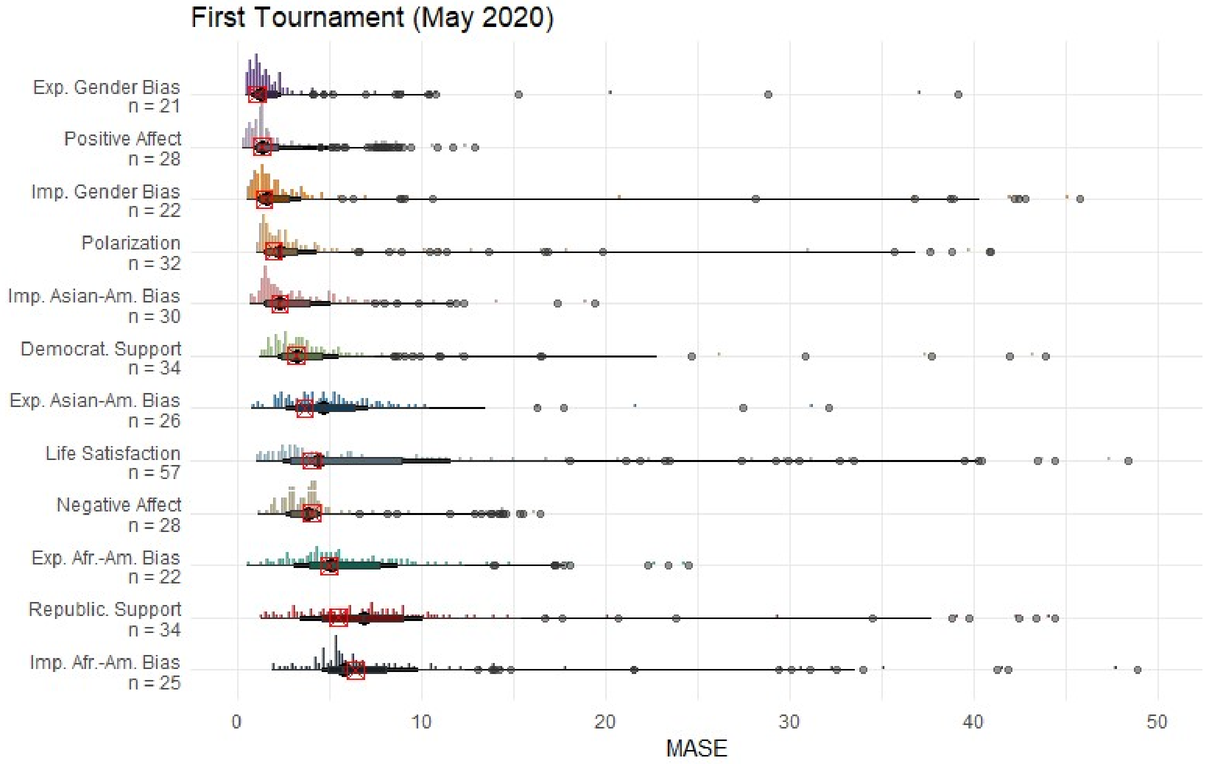
Disciplinary affiliation of top 5 teams in each domain in Tournament 2.

Supplementary Figure 11. Prior forecasting experience among the top 5 teams in each domain in Tournament 2



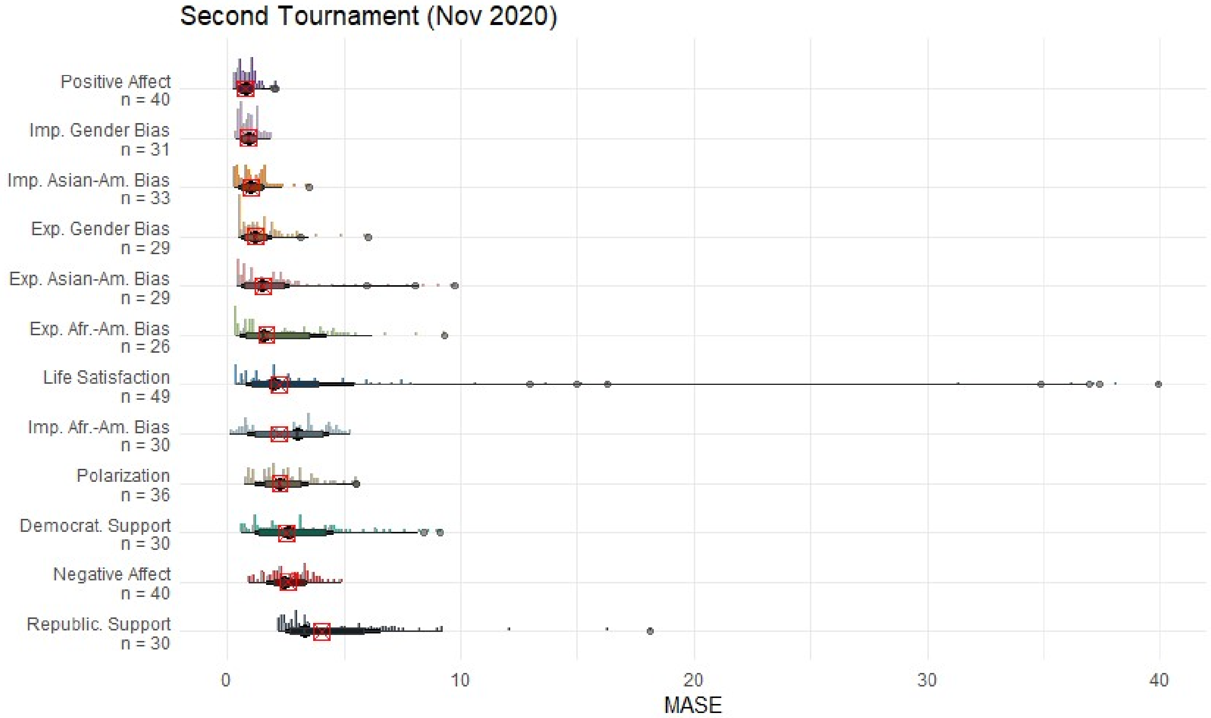
Prior forecasting experience of top 5 teams in each domain in Tournament 2.

Supplementary Figure 12. Social scientists' forecasts compared to estimated means in Tournament 1



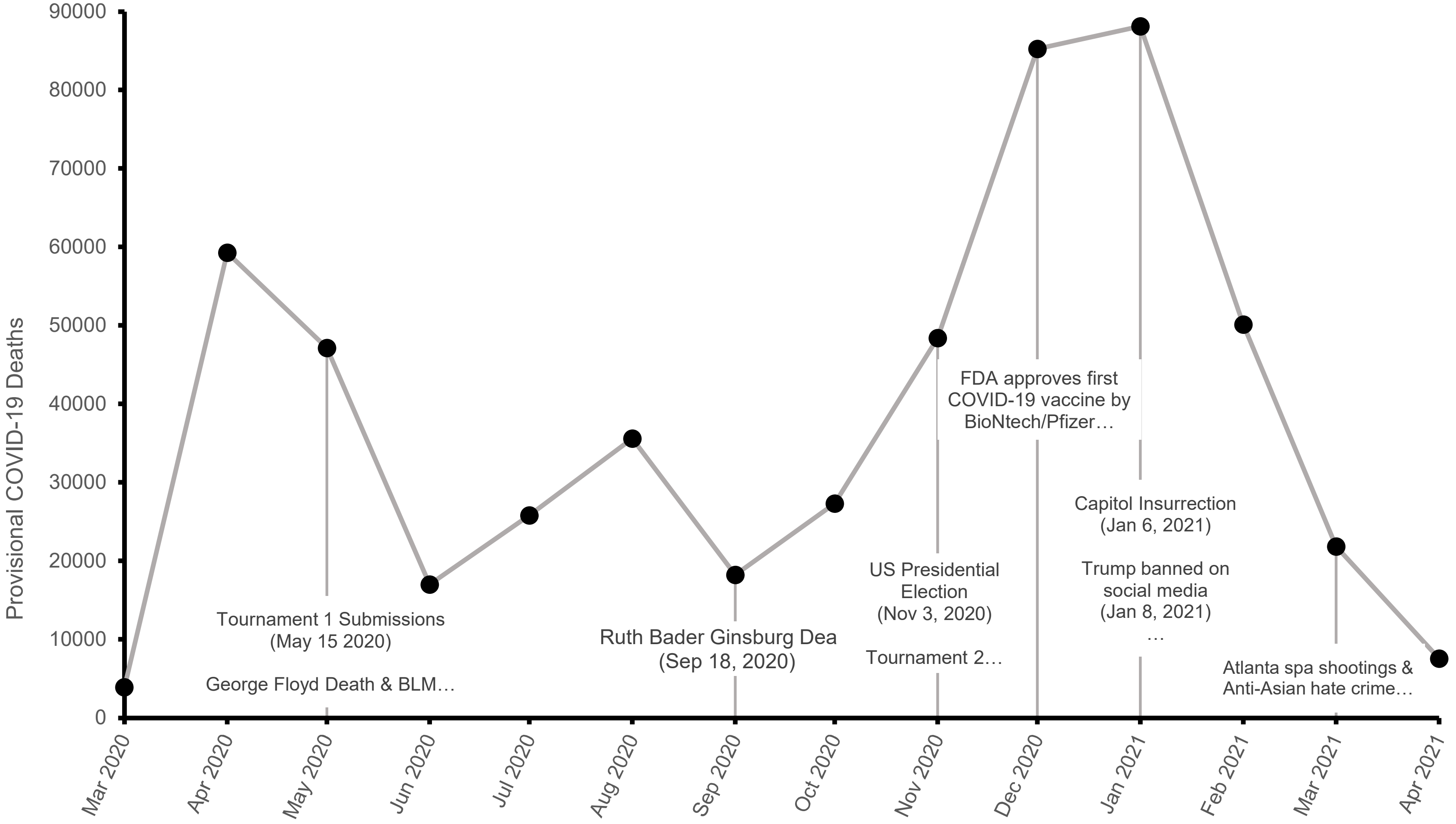
Boxplots and dot-plots with medians and outliers (round dots), as well as domain estimates from linear mixed model (red crossed square) in Tournament 1. The box of the box plot shows the 25th percentile, the median and the 75th percentile. The length of the whiskers is 1.5 × interquartile range. *n* = number of teams with submissions for a given domain.

Supplementary Figure 13. Social scientists' forecasts compared to estimated means in Tournament 2.

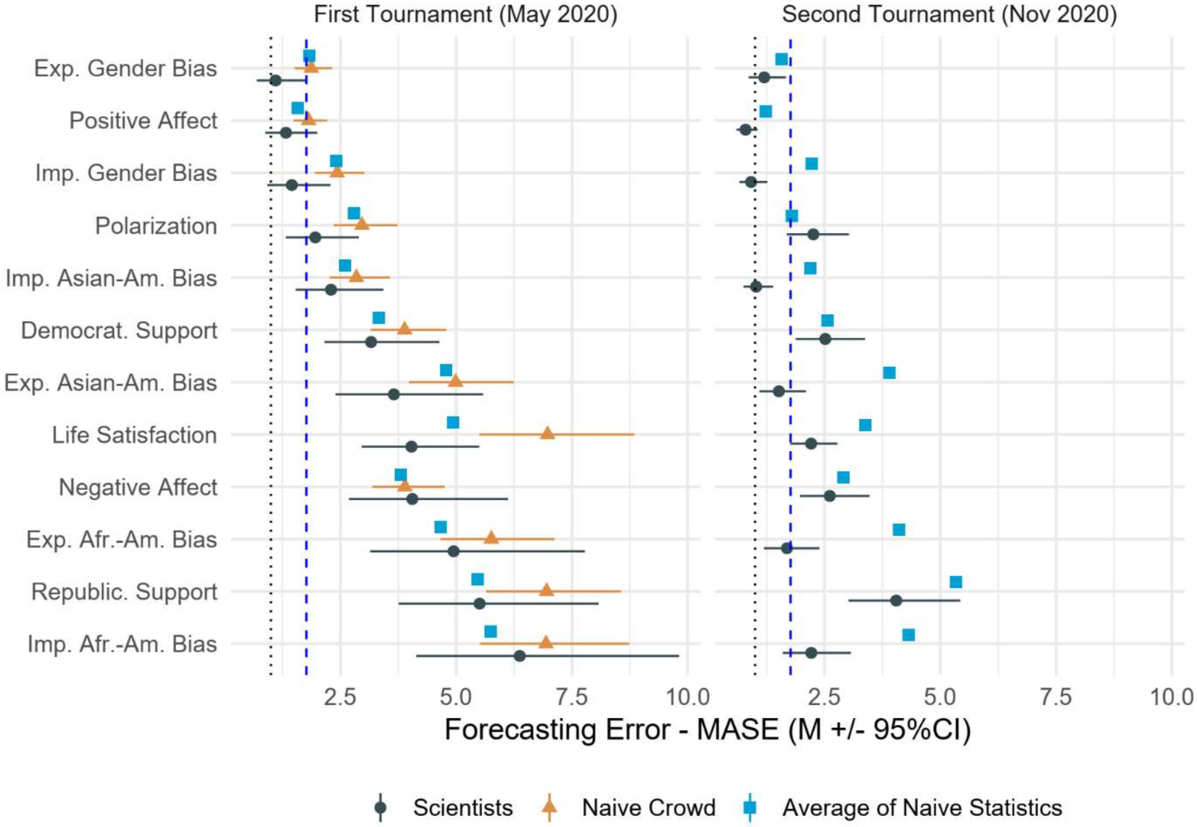


Boxplots and dot-plots with medians and outliers (round dots), as well as domain estimates from linear mixed model (red crossed square) in Tournament 2. The box of the box plot shows the 25th percentile, the median and the 75th percentile. The length of the whiskers is 1.5 × interquartile range. *n* = number of teams with submissions for a given domain.

Supplementary Figure 14. Timeline of major historical events in the US during the tournament



Supplementary Figure 15. Forecasts of scientists, naïve crowd, and the average of statistical benchmarks



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

Supplementary Table 1. Demographic characteristics of forecasting teams.

	First Tournament (May 2020)	**Second Tournament (Nov 2020)**
N Teams	86	120
N Participants	135	190
Time size *M* (*SD*)	1.57 (1.12)	1.58 (1.10)
N Forecasted domains *M* (*SD*)	4.17 (3.78)	4.55 (3.88)
% Teams who predicted (published on):		
Explicit gender-career bias	24 (14)	30 (14)
Implicit gender-career bias	26 (9)	31 (11)
Positive affect on social media	33 (25)	43 (33)
Political polarization	37 (41)	39 (43)
Life satisfaction	66 (33)	68 (37)
Explicit Asian American bias	30 (12)	31 (16)
Implicit Asian American bias	35 (13)	35 (19)
Democratic support -congressional ballot	40 (41)	39 (45)
Republican support - congressional ballot	40 (41)	39 (45)
Negative affect on social media	33 (25)	43 (33)
Explicit African American bias	26 (18)	27 (31)
Implicit African American bias	29 (16)	31 (27)
Age *M* (*SD*)	38.18 (8.37)	36.82 (8.30)
% Without Doctoral Degree on a team *M* (*SD*)	21.79 (39.58)	25.59 (41.88)
% non-US on a team *M* (*SD*)	59.69 (47.61)	59.72 (47.82)
% non-male on a team *M* (*SD*)	19.76 (36.36)	21.78 (37.48)
% Teams updated predictions in Nov 2020	44	

Note. The percentage of forecasting teams with self-identified expertise—i.e., team members with publications on a particular topic—are relative to the total number of teams that chose to forecast a given topic.

Demographic characteristics of forecasting teams.

Supplementary Table 2. Top performers in each domain in Tournament 1.

Rank	Domain	Scale of predicted value	12m *SD* of observed data	12m *M* absolute error	MASE	Team name
1	explicit gender-career bias	Minus 6 (career is strongly female, whereas family is strongly male) to 6 (career is strongly male, whereas family is strongly female)	0.033	0.02	0.52	**fearfulastra**
2	implicit gender-career bias	IAT *D* score (diff score of latencies / SD)	0.006	0.01	0.53	**The Well-Adjusted R Squares**
3	positive affect on social media	z-scores, standardized on all prior time series data (M = 0.145 / SD = 0.0215)	0.150	0.22	0.76	**Imielin**
4	political polarization	Abs difference of % support by Dem vs. Reps	3.394	2.04	0.99	**Junesix**
5	life satisfaction	0-10 scale	0.036	0.03	1.04	**polarization-2020CO**
6	explicit Asian American bias	difference in majority - minority groups feeling *t* (0-100)	0.032	0.06	1.07	**The Well-Adjusted R Squares**
7	implicit Asian American bias	IAT *D* score (diff score of latencies / SD)	0.008	0.03	1.13	**NotGreatWithNamesTeam**
8	Democratic support - congressional ballot	%	1.977	1.42	1.33	**AbCdEfG**
9	Republican support - congressional ballot	%	2.286	1.57	1.61	**Erebuni**
10	Negative affect on social media	z-scores, standardized on all prior time series data (M = 0.145 / SD = 0.0215)	0.372	0.39	2.00	**Imielin**
11	Explicit African American bias	difference in majority - minority groups feeling *t* (0-100)	0.016	0.07	2.21	**fearfulastra**
12	Implicit African American bias	IAT *D* score (diff score of latencies / SD)	0.011	0.03	4.55	**fearfulastra**

Top performers in each domain in Tournament 1, ranked by accuracy across domains.

Supplementary Table 3. Top performers in each domain in Tournament 2.

Rank	Domain	Scale of predicted value	6m *SD* of observed data	6m *M* absolute error	MASE	Team name
1	Implicit African American bias	IAT *D* score (diff score of latencies / *SD*)	0.010	0.001	0.15	**Polarization Disciples**
2	positive affect on social media	z-scores, standardized on all prior time series data (*M* = 0.145 / *SD* = 0.0215)	0.094	0.062	0.23	**The Forecasting Four**
3	implicit Asian American bias	IAT *D* score (diff score of latencies / *SD*)	0.011	0.007	0.29	**TAPE-Measurement**
4	Explicit African American bias	difference in majority - minority groups feeling *t* (0-100)	0.018	0.011	0.35	**BlackSwan**
5	life satisfaction	0-10 scale	0.047	0.012	0.36	**Sociology**
6	Implicit gender-career bias	IAT *D* score (diff score of latencies / SD)	0.005	0.004	0.37	**datamodelers**
7	Explicit Asian American bias	difference in majority - minority groups feeling *t* (0-100)	0.033	0.026	0.45	**Imielin**
8	Explicit gender-career bias	−6 (career is strongly female, whereas family is strongly male) to 6 (career is strongly male, whereas family is strongly female)	0.023	0.020	0.51	**Polarization Disciples**
9	Democratic support - congressional ballot	%	0.944	0.632	0.59	**AbCdEfG**
10	political polarization	Abs difference of % support by Dem vs. Reps	4.336	1.807	0.78	**BRTWN**
11	Negative affect on social media	z-scores, standardized on all prior time series data (M = 0.145 / SD = 0.0215)	0.254	0.191	0.95	**Team Supreme Ignorance**
12	Republican support - congressional ballot	%	0.822	2.204	2.17	**teamGreenLake**

Top performers in each domain in Tournament 2, ranked by accuracy across domains

Supplementary Table 4. Inaccuracy of Data-inclusive vs. Data-free Models.

Tournament	Domain	*no data / data-inclusive Ratio*	*SE*	*df*	*t -ratio*	*P-value*
First (May 2020)	Exp. African American Bias	1.201	0.261	326.962	0.843	. 515
	Exp. Asian American Bias	**2.008**	**0.389**	**334.969**	**3.593**	**.004**
	Explicit Gender-career bias	1.729	0.372	329.761	2.542	.103
	Imp. African American Bias	1.142	0.233	334.319	0.651	.515
	Imp. Asian American Bias	**2.237**	**0.414**	**333.967**	**4.356**	**<.001**
	Ideology Democrats	1.440	0.255	333.820	2.053	.286
	Ideology Republicans	1.256	0.223	333.820	1.285	.515
	Implicit Gender-career Bias	**2.106**	**0.460**	**328.407**	**3.409**	**.007**
	Life Satisfaction	1.402	0.197	298.454	2.398	.137
	Negative Affect	0.768	0.143	333.845	-1.414	.515
	Political Polarization	0.871	0.157	331.211	-0.767	.515
	Positive Affect	0.772	0.144	333.845	-1.384	.515
Second (Nov 2020)	Exp. African American Bias	1.107	0.276	510.618	0.406	.685
	Exp. Asian American Bias	1.125	0.252	521.995	0.527	.685
	Explicit Gender-career Bias	**2.089**	**0.463**	**519.138**	**3.321**	**.011**
	Imp. African American Bias	1.235	0.303	520.239	0.862	.685
	Imp. Asian American Bias	1.276	0.273	520.237	1.140	.685
	Ideology Democrats	1.611	0.325	521.668	2.368	.182
	Ideology Republicans	1.167	0.235	521.668	0.768	.685
	Implicit Gender-career Bias	1.145	0.259	515.449	0.596	.685
	Life Satisfaction	**2.067**	**0.322**	**478.215**	**4.663**	**<.001**
	Negative Affect	0.790	0.155	519.240	-1.204	.685
	Political Polarization	1.086	0.214	517.811	0.418	.685
	Positive Affect	0.708	0.139	519.240	-1.762	.685

Note. Ratio = data-free MASE / data-inclusive MASE. Scores >1: greater accuracy of data-inclusive forecasts. Scores < 1: greater accuracy of forecasts based on intuition / theory alone. Pairwise contrasts obtained via *emmeans* package in *R*, drawing on the restricted information maximum likelihood linear mixed effects model with model type (data-inclusive or not data-inclusive), domain, and their interaction as predictors of the *log*(MASE) scores, with responses nested in participants. We ran separate models for each tournament. To avoid skew, tests performed on log-transformed MASE scores. Degrees of freedom obtained via Kenward-Roger approximation.

Contrasts of Mean-level Inaccuracy (MASE) among Scientists Using Data-inclusive vs. Data-free Models.

Supplementary Table 5. Effect of strategies & individual characteristics on forecasting accuracy.

	Unstandardized Coefficients	Standardized Coefficients
(Intercept)	-1.96 **	-2.15 **
	(0.66)	(0.66)
N model parameters	-0.02	-0.13
	(0.01)	(0.08)
Statistical model complexity	-0.14 *	-0.20 *
	(0.06)	(0.09)
Considered COVID-19 (yes/no)	0.04	0.04
	(0.08)	(0.08)
Considered counterfactuals (yes/no)	-0.08	-0.08
	(0.06)	(0.06)
Number of predicted domains	0.02	0.13
	(0.01)	(0.11)
Data Scientists on the team (yes/no)	0.89	0.89
	(0.67)	(0.67)
Behav.-Soc Scientists on the team (yes/no)	1.15	1.15
	(0.65)	(0.65)
Multidisciplinary	1.16	1.16
	(0.67)	(0.67)
Team size	0.02	0.05
	(0.06)	(0.13)
% Team members without a PhD	0.001	0.11
	(0.001)	(0.11)
Confidence in forecast	-0.01	-0.03
	(0.03)	(0.08)
Self-reported expertise	-0.03	-0.12
	(0.03)	(0.10)
Team members published in the forecasted domain (yes/no)	0.26 **	0.26 **
	(0.09)	(0.09)
Prior engagement with forecasting tournaments	0.35 *	0.35 *
	(0.17)	(0.17)
N	905	
N (teams)	120	
AIC	1940.35	
BIC	2074.97	
*R*2 (fixed)	0.31	
*R*2 (total)	0.58	

Note: Results from a linear mixed effect model (restricted information maximum likelihood estimator) across both tournaments, with candidate variables in the table as predictors of *log* MASE scores, with domain (11 dummy-variables) as a covariate, and responses nested in participants. All continuous predictors are mean-centered. Standard errors, in parentheses, are heteroskedasticity robust. In the standardized model, predictors are scaled by 2 standard deviations. *** $p < .001$; ** $p < .01$; * $p < .05$.

Regression model coefficients demonstrating effects of forecasting strategies and individual characteristics predicting forecasting accuracy across two tournaments.

Supplementary Table 6. Effects of different updating rationales for forecasting inaccuracy.

Variable	*Est.*	*S.E.*	*t*	*df*	*P*
Intercept	0.359	0.117	3.068	48.213	.004
Update based on Data Received	-0.135	0.169	-0.801	61.505	.426
Update based on theory	0.782	0.662	1.181	78.712	.241
Update based on consideration of external events	0.081	0.167	0.484	98.156	.630
N			162		
N (teams)			38		
AIC			412.53		
BIC			431.06		
*R*2 (fixed)			0.02		
*R*2 (total)			0.17		

Note: Results from a linear mixed effect model (restricted information maximum likelihood estimator) for a subset of scientist teams in the Second Tournament who chose to update their predictions. The model includes the updating type as a fixed effect predictor of *log* MASE scores with responses nested in participants. All continuous predictors are mean-centered. Standard errors, in parentheses, are heteroskedasticity robust. In the standardized model, predictors are scaled by 2 standard deviations. *** $p < .001$; ** $p < .01$; * $p < .05$.

Regression model coefficients demonstrating effects of different updating rationales for forecasting inaccuracy in the Second Tournament.

Supplementary Table 7. Exclusions by category when processing the general public sample.

Data Filtering Step	**Unique Participants**	**Participants Removed**	**N Predictions**	**Predictions Removed**
Original *N*	1891			
Remove incomplete 12-month predictions	1173	718	2638	
Remove Duplicate submissions	1155	18	2611	27
Remove Outliers	1094	59	2448	163
Remove misunderstood responses	1088	6	2426	22
Remove bogus responses	1062	26	2330	96
Remove Responses below 50 sec	803	259	1469	861
Removed flagged Covid responses*	802	1	1467	2

Note. *Covid-related open-ended responses were not originally reviewed in the coding and were coded at a later stage. Two participants were flagged – one for indicating their predictions were based on Russia, not the U.S., and the other for responding in Spanish instead of English. One of these participants was already flagged in a previous step, which is why only one participant/responses were removed at this stage.

Breakdown of exclusions of the general public sample by category.

Supplementary Table 8. Comparison of resampled historical mean and averaged historical mean in Tournament 1.

Domain	Historical Mean	Resample Mean	|Diff|	*ln*(MASE) historical mean	*ln*(MASE) resample mean	|Diff|
Explicit Prejudice – Af Am	-0.035	-0.035	0.000	1.718	1.713	0.005
Explicit Prejudice – Asian	-0.037	-0.037	0.000	0.932	0.925	0.006
Explicit Prejudice – Gender	0.837	0.837	0.000	0.184	0.381	0.197
Implicit Prejudice – Af Am	0.319	0.319	0.000	1.726	1.725	0.001
Implicit Prejudice – Asian	0.386	0.386	0.000	0.404	0.410	0.006
Implicit Prejudice – Gender	0.365	0.365	0.000	0.241	0.326	0.085
Support for Democrats	43.291	43.299	0.008	1.348	1.414	0.066
Support for Republicans	36.721	36.708	0.013	1.845	1.855	0.010
Life Satisfaction	6.362	6.362	0.000	0.304	0.998	0.694
Negative Affect	0.973	0.974	0.001	1.453	1.455	0.002
Polarization	79.385	79.391	0.005	1.021	1.091	0.070
Positive Affect	-0.973	-0.973	0.000	0.310	0.333	0.023

Note. Historical Mean was calculated by averaging all historical data per domain and using this value (repeated for each of the 12 months) as the forecast for computing MASE scores. Resample Mean was calculated by creating 10,000 simulated forecasts and computing the average MASE/ln(MASE) across these 10000 forecasts. Each forecast was created by sampling with replacement 12 values from the historical observations, then computing the MASE score for these 12 values compared to the observed domain values.

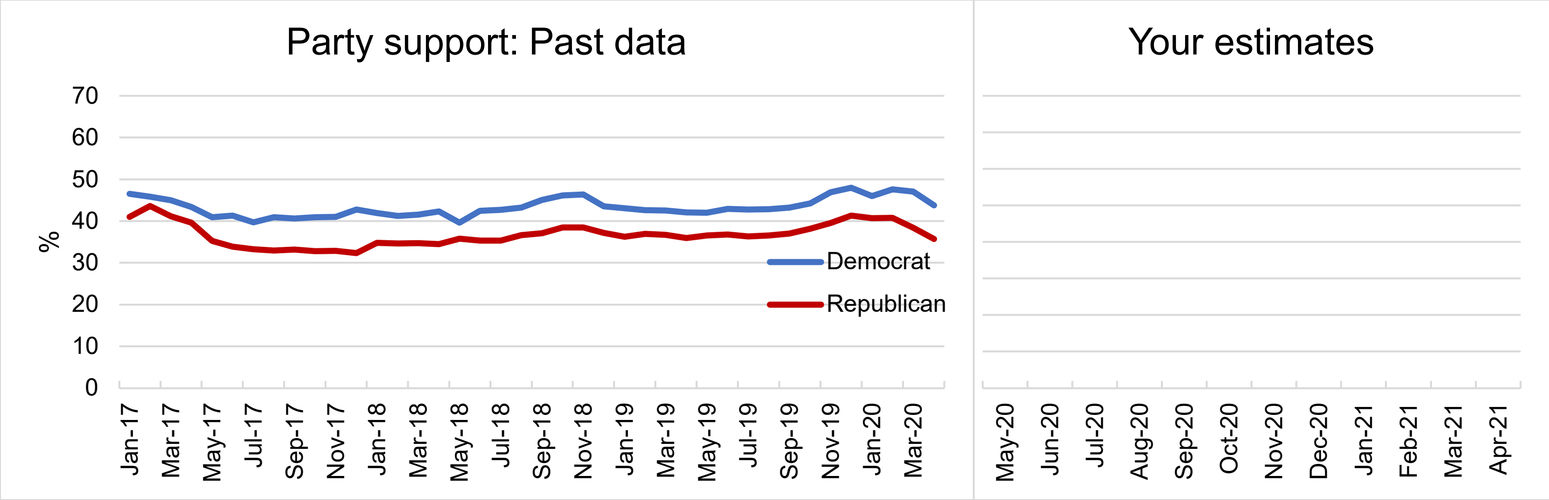
Supplementary Table 9. Table of content for the R Markdown analyses

Name of .Rmd chunks	Output / Specific Goal
Chunk 1 “setup”	Load R packages needed
Chunk 2 “setup working directory”	Setup working directory
Chunk 3 “Import data”	Import data
Chunk 4 “get simulated benchmark data & add RW to data”	MASE differences between the participants’ MASE and the MASE of random walk of Tournament 1 and 2. Model comparison categories based on MASE cutoff scores.
Chunk 5 “set subsets of data for analyses”	Data subsets that fulfill certain conditions for later analysis. For instance, only Tournament 1 data, only Tournament 2 data, objective experts, etc.
Chunk 6 “create subsets for separate visualizations”	Dataset with labels that are close to natural language for data visualization. Dataset with important statistics for visualization. For instance, mean, median, top performance for different domains and level of expertise
Chunk 7 “create a file to share with teams to announce how they did”	Prepare both tournament datasets to share with the academic teams about how they did (MASE) compared with the Ground truth marker and their rank.
Chunk 8 “visualize top performers”	Fig. S4. - Fig. S11. Error among top 5 teams in each domain in Tournament 1, by approach.
Chunk 9 “Inspect historical trends and calculate complexity”	Markers of complexity for the tournament, including sd, mad, and an supplementary metric of permutation entropy
Chunk 10 “PHASE 1 prep and simple visualizations of trends”	Fig. S1 and Figure 2 in the manuscript
Chunk 11 “Phases 1 and 2 along with sims”	graph individual predictions and ground truth markers - Fig. S2 THE SUPPLEMENT in the PAPER, as well as one Figure 1 in the MAIN TEXT) analyses of scientists versus lay people in tournament 1
Chunk 12	statistical tests of difference from benchmark. Figure 3.
Chunk 13	compare scores from tournament 1 and tournament 2 test complexity associations
Chunk 14	graph change in ranking. Figure 4.
Chunk 15 “ranking in phase 1 and complexity”	ranking of domain by forecasting error and correlation to historical variability in trends
Chunk 16	compare scores from tournament 1 - first six months vs. last six months
Chunk 17	Consistency in forecasting
Chunk 18	visualize by method (phase 1 and 2) Figure 5
Chunk 19	examine effects of covariates across both tournaments, Figure 6
Chunk 20	Role of Updating - Phase 2
Chunk 21	demographics Phase 1
Chunk 22	demographics lay people

Supplementary References

1. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* **36**, 54–74 (2020).
2. Athanasopoulos, G., Hyndman, R. J., Song, H. & Wu, D. C. The tourism forecasting competition. *International Journal of Forecasting* **27**, 822–844 (2011).
3. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).
4. Goodrich, B., Gabry, J., Ali, I. & Brilleman, S. rstanarm: Bayesian applied regression modeling via Stan. (2022).
5. R Core Team. R: A Language and Environment for Statistical Computing. (2022).
6. Rouder, J. N. & Morey, R. D. Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research* **47**, 877–903 (2012).
7. Lenth, R., Singmann, H., Love, J. & Maxime, H. emmeans: Estimated Marginal Means, aka Least-Squares Means. (2020).
8. Makowski, D., Ben-Shachar, M. & Lüdecke, D. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *JOSS* **4**, 1541 (2019).
9. Genz, A. & Bretz, F. *Computation of multivariate normal and t probabilities*. (Springer, 2009).
10. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting* **34**, 802–808 (2018).

Supplementary Appendix 1. Example submission forms.

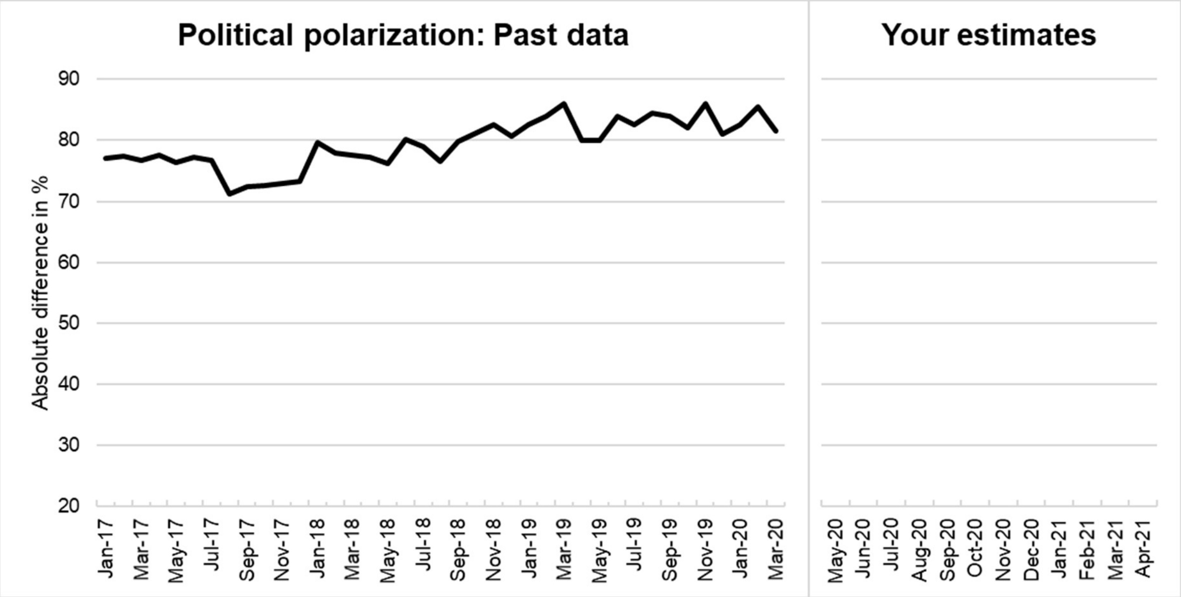


Time	**Democrat**	**Republican**		
May-20			←	**enter your monthly estimates in the box to the left**
Jun-20			←	**enter your monthly estimates in the box to the left**
Jul-20			←	**enter your monthly estimates in the box to the left**
Aug-20			←	**enter your monthly estimates in the box to the left**
Sep-20			←	**enter your monthly estimates in the box to the left**
Oct-20			←	**enter your monthly estimates in the box to the left**
Nov-20			←	**enter your monthly estimates in the box to the left**
Dec-20			←	**enter your monthly estimates in the box to the left**
Jan-21			←	**enter your monthly estimates in the box to the left**
Feb-21			←	**enter your monthly estimates in the box to the left**
Mar-21			←	**enter your monthly estimates in the box to the left**
Apr-21			←	**enter your monthly estimates in the box to the left**

Data Source: Aggregated weighted data from the Congressional Generic Ballot polls conducted between January 2017 and March 2020.

Congressional generic Ballot asks representative samples of Americans to indicate which party they would support in an election.

Obtained from projects.fivethirtyeight.com/congress-generic-ballot-polls



Political polarization: Past data

Your estimates

Time	**Polarization Score**		
May-20		←	**enter your monthly estimates in the box to the left**
Jun-20		←	**enter your monthly estimates in the box to the left**
Jul-20		←	**enter your monthly estimates in the box to the left**
Aug-20		←	**enter your monthly estimates in the box to the left**
Sep-20		←	**enter your monthly estimates in the box to the left**
Oct-20		←	**enter your monthly estimates in the box to the left**
Nov-20		←	**enter your monthly estimates in the box to the left**
Dec-20		←	**enter your monthly estimates in the box to the left**
Jan-21		←	**enter your monthly estimates in the box to the left**
Feb-21		←	**enter your monthly estimates in the box to the left**
Mar-21		←	**enter your monthly estimates in the box to the left**
Apr-21		←	**enter your monthly estimates in the box to the left**

Data Source: Aggregated data from the Gallup polls conducted between January 2017 and March 2020.
Data represents absolute value of the difference score in Presidential Job Approval by Party Identification (Democrat vs. Republican)

Obtained from news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx

Supplementary Appendix 2. Screening Low Comprehension Responses in the Public

Verbatim information provided to coders, including standardized information about the study.

Participants were asked to submit predictions of how they expected different domains to change over the course of a year. The 12 domains were: Life satisfaction; Affect (both positive and negative); Political polarization; Ideological preferences (Democrat & Republican); African American bias (implicit and explicit); Asian American bias (implicit and explicit); Gender-career bias (implicit and explicit). In addition to submitting their forecasts, we asked them to answer questions regarding how they went about creating those forecasts. Participants provided written responses to 3 different questions which are denoted by the following columns:

Column Name	Question Asked
Theory	Please describe your thought process for generating your predictions **for life satisfaction**. If your forecast is based on theoretical assumptions, please describe your assumptions.
Parameters	**Did you consider additional variables in your predictions?** By additional variables we mean variables other than prior data on **life satisfaction** itself. These variables can include single-shot events (e.g., political leadership change; implementation of a particular policy) or variables that change over time (e.g., COVID-19 deaths; unemployment rate). Describe your decision process when making your predictions.
Coviddesc	What is the role of COVID-19 pandemic trajectory in your Life Satisfaction forecast? Please, describe **how you think this variable impacts your Life Satisfaction predictions.**

Note. Not all participants were required to answer the *Coviddesc* question as it was only presented if they selected a particular response to a previous question.

Your task is to look at their written responses for these 3 columns and assign a value of 0 (no issue), 1 (did not understand), or 2 (bogus response) to each row.

<u>2 Bogus responses</u> are responses where the participant did not take the study seriously. For instance, participant wrote the following responses for each column:

Theory	Parameters	Coviddesc
"Ok"	"Ok"	"good"

In this case, the participant avoids answering all 3 questions and we cannot assume they took the survey seriously as a result.

<u>1 Did not understand responses</u> refer to participants who did not understand the question posed or the predictions they were supposed to make. **Keep in mind that you are not coding them on the quality of their response.** If their response includes a rationale that informed their prediction and doesn't explicitly indicate they misunderstood the question, **do not code as "didn't understand."**

For instance, participant wrote the following responses for each column.

Theory	Parameters	Coviddesc
"The data seems to go up and down"	"I thought that maybe as more people are using Twitter currently it could be higher but will drop after lockdowns release"	"I thought it would make it go higher"

In this case, the participant's responses indicate that they likely misunderstood what the data we provided was about. They are assuming that more Twitter users = higher life satisfaction (because the life satisfaction data was based on twitter data), but that's not what we were asking them to predict!

<u>0 No issue responses</u> refer to responses where the participants responded in good faith and did not noticeably misunderstand the task. **NOTE**: rows can be flagged as "no issue" even if only one of the columns was responded to, so long as the response doesn't come across as problematic.

Supplementary Appendix 3. Coding Instructions for Forecasting Method

Teams could select multiple methods for their forecasts, including intuition, theory, simulation-based, data-driven, or others. For the chief analyses, we were interested in teams that relied on data-inclusive vs. data-free methods. Therefore, to cross-validate forecasting methods provided by teams, and to obtain markers of data-inclusive vs. data-free methods, two research assistants independently reviewed submission justifications and rationales. The following steps describe the process used to code participant submissions in terms of the type of forecasting method:

1. **Intuition** – These are forecasts where the participant made their predictions by relying solely on their intuition.
   - Participants who made such forecasts relied on their expert intuitions to estimate how they expected the variable to change over the next year.
   - Example 1: The participant indicates that they "expected the difference between democrats and republicans to become larger toward November" because of the election.
   - Example 2: The participant predicts that the holidays would lead to an increase in life satisfaction, followed by a decrease after the New Year.

2. **Theory** – These were forecasts where the participant indicated that their predictions were guided by a particular theory.
   - Example 1: The participant mentions that their predictions were based on "hedonic adaptation" theory, which predicts a return to baseline after a significant event such as the pandemic.
   - Example 2: Participant mentions that their model is based on the "family stress model," which predicts that great social disruption and economic fallout leads to significant drops in life satisfaction.

3. **Data-driven** – These were forecasts where the participant indicated that they used one or more data-driven forecasting methods.
   - Participants may have indicated they used one or more methods of forecasting that can vary in complexity.
   - These methods ranged from simple mean of the data provided to complex autoregressive moving average (ARIMA) models or econometric models.
4. **Hybrid** – These were forecasts that used **Data-driven** methods **AND** either **Intuition** and/or **Theory.**
   - If a participant indicated that they used both Theory and Intuition, but **NOT** Data, coders were instructed to mark it as Theory.

Because the number of participants who explicitly mentioned theory was underpowered (*n* = 9 in Tournament 1), we collapsed this category with intuition-based judgments.

Supplementary Appendix 4. Coding Instructions for Model Complexity

Model complexity was coded on a 1-3 scale, indicating the complexity of the forecasting methods used by participants to develop their forecasts:

1. **Simple forecasting method**
    a. Simple forecasting methods referred to forecasting methods that did not account for any additional parameters or variables, such as regression to the mean, Intuition-based forecasts, and/or simple Drift models.
2. **Moderate forecasting method**
    a. Moderate forecasting methods referred to forecasting methods that used 1-3 additional parameters to develop their forecasts, such as univariate time series forecasting models, Holt-Winters seasonal corrections, & auto-regressions with time lags.
3. **Complex forecasting method**
    a. Complex forecasting methods refer to forecasting methods that use more than 3 additional parameters to develop their forecasts, such as ARIMA and dynamic econometric models.

Coders were instructed to consider the number of parameters teams used. They were also instructed to base their judgment on the method participants provided in their rationales (e.g., ARIMA) in cases where a participant mentioned a complex method but did not list any (or few) additional parameters. In case multiple models were mentioned, coders identified the model teams indicated they used for the actual forecast itself.

Supplementary Appendix 5. Coding Instructions for Updating Justifications

Coders were provided with the following instructions:

"You will be presented with a data set containing participant responses with regards to the theory, methods, and parameters they considered in their forecast. All responses are from teams that had submitted a forecast 6 months prior, and who chose to update their forecasts after receiving a comparison of how their forecasts had compared so far to the actual results for the domains they predicted.
Your task is to review their responses and determine whether they meet the criteria for the following categories of justification:

1. **Updated based on data received:** The team indicates they updated their data based on the updated data set we provided them.
   - To be coded for this category, the participants need to explicitly mention that they updated their response in response to the new data being provided.
   - E.g., "Updating my estimates based on the mean"
2. **Updated due to theoretical insights:** The team indicates they updated their data based on an explicitly mentioned theory.
   a. To be coded for this category, the participants must explicitly mention a theory as the reason for their updated forecast.
3. **Updated due to consideration of external events:** The team indicates they updated their data based on events that have occurred in the 6 months since their original forecast.
   a. To be coded for this category, the participants must explicitly mention an external event as the reason for updating their forecast.
   b. E.g., "After testing for seasonality and discontinuity, we used the mean of new values after the onset of COVID-19 pandemic and racial protests in the USA."

Each category is coded in a separate column as either a 1 (Justification present) or 0 (Justification not present). **Please note**, a response can be coded for multiple categories, in which case it would be marked as a 1 in each of the relevant columns."