nature portfolio

Corresponding author(s): Igor Grossmann

Last updated by author(s): Dec 8, 2022

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 κ = .70 unweighted, κ = .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 κ = .70 unweighted, κ = .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	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging