Multi-trait diversity of online groups improves geo-political forecasting accuracy as a function of group size

Niccolò Pescetelli*[1,2], Alex Rutherford[1,2], and Iyad Rahwan[1,2]

[1]Center for Humans and Machines, Max Planck Institute for Human Development, 94 Lentzeallee, Berlin, DE 14195
[2]Media Lab, Massachusetts Institute of Technology, 75 Amherst Street, Cambridge, MA 02139

Many modern interactions happen in a digital space, where automated recommendations and homophily can shape the composition of groups interacting together and the knowledge that groups are able to tap into when operating online. Digital interactions are also characterized by different scales, from small interest groups to large online communities. Here, we manipulate the composition of online groups based on a large multi-trait profiling space to explore the causal link between group composition and performance as a function of group size. We asked volunteers to search information online under time pressure and measured individual and group performance in forecasting real geo-political events. Our manipulation affected the correlation of forecasts made by people after online searches. Group composition interacts with group size so that diversity benefits individual and group performance proportionally to group size. Aggregating opinions of modular crowds composed of small independent groups achieved better results than using non-modular ones. Finally, we show differences existing among groups in terms of disagreement, speed to convergence to consensus forecasts and within-group variability in performance. The present work sheds light on the mechanisms underlying effective collaboration in digital environments.

group diversity | forecasting | judgment aggregation | group size
Correspondence: *pescetelli@mpib-berlin.mpg.de*

Introduction

Understanding how people collect information about world events, and discuss this knowledge with others online to form shared opinions is a crucial and timely research question. In the past decade, there have been widespread concerns that search engines and news filtering algorithms may contribute to the formation of clusters of individuals with highly correlated information and poorly diversified news sources (1–3). Little is known about the exact mechanisms underlying personalization but content is often provided by clustering users on highly dimensional feature spaces, along shared variables (demographics, geo-location, social network, tastes and past behavior) (4–8). Furthermore, people sharing traits are more likely to cluster together in online communities, a phenomenon known as homophily (9, 10). One question is whether recommendation algorithms and homophily can impact the ability of online groups to collectively search and use online information to form accurate predictions, especially under high time pressure and uncertainty—namely when the opportunities for rational debates are scarce (11, 12).

In this paper, we manipulate the composition of online groups and their size/modularity (see Supplementary information §1-2). Both factors are expected to affect the amount and independence of information that a group can tap into. We measure individual and group performance as Brier errors in forecasting real geo-political events (Table 1), a task with high ecological validity that challenges experts and professional intelligence analysts. These problems are characterized by high degrees of uncertainty, correlated information between judges, dependence on multiple indicators (*e.g.*, economics, politics, social unrest, etc.), and, importantly, time criticality (*i.e.*, there are huge costs associated with making the correct prediction too late).

Diversity is a highly heterogeneous construct touching several disciplines (13–16). From an informational stand point, psychologists have recognized the importance of group diversity for information independence, group performance, resilience to group biases, complex thinking, creativity and exploration of large solution spaces (17–27). The approach used in psychology is aimed at studying single dimensions of diversity (*e.g.*, skill, age, race (22, 27, 28)). Contrary to this, we are here interested in the effects that sorting people based on a large multi-feature space (Figure 1a) can have on the information diversity that a group can forage online (see Supplementary Information for a full list of features considered here). We note that demographics, cognitive and personality traits can be easily inferred from digital traces, and used to customize searches and recommend content (29–32). Although some of these features (like demographics) are known to psychologists not to affect information diversity per se (33, 34), they may do so in an online environment that maps inter-individual differences into information access. The more distant two people are on an arbitrarily large profiling space, the less likely they might be to belong to the same online information bubble. Given the difficulty of disentangling the causal contributions of group composition on performance, we here employ an experimental design, where half of the sample (core segment) is randomly assigned to interact with the rest 25% most similar (inner segment) or 25% most dissimilar (outer segment) individuals in the sample (20, 22, 23, 27) (Figure 1b-c). We used mean Euclidean distance on profiling space as a measure of similarity, but notice that this measure was strongly correlated with standard

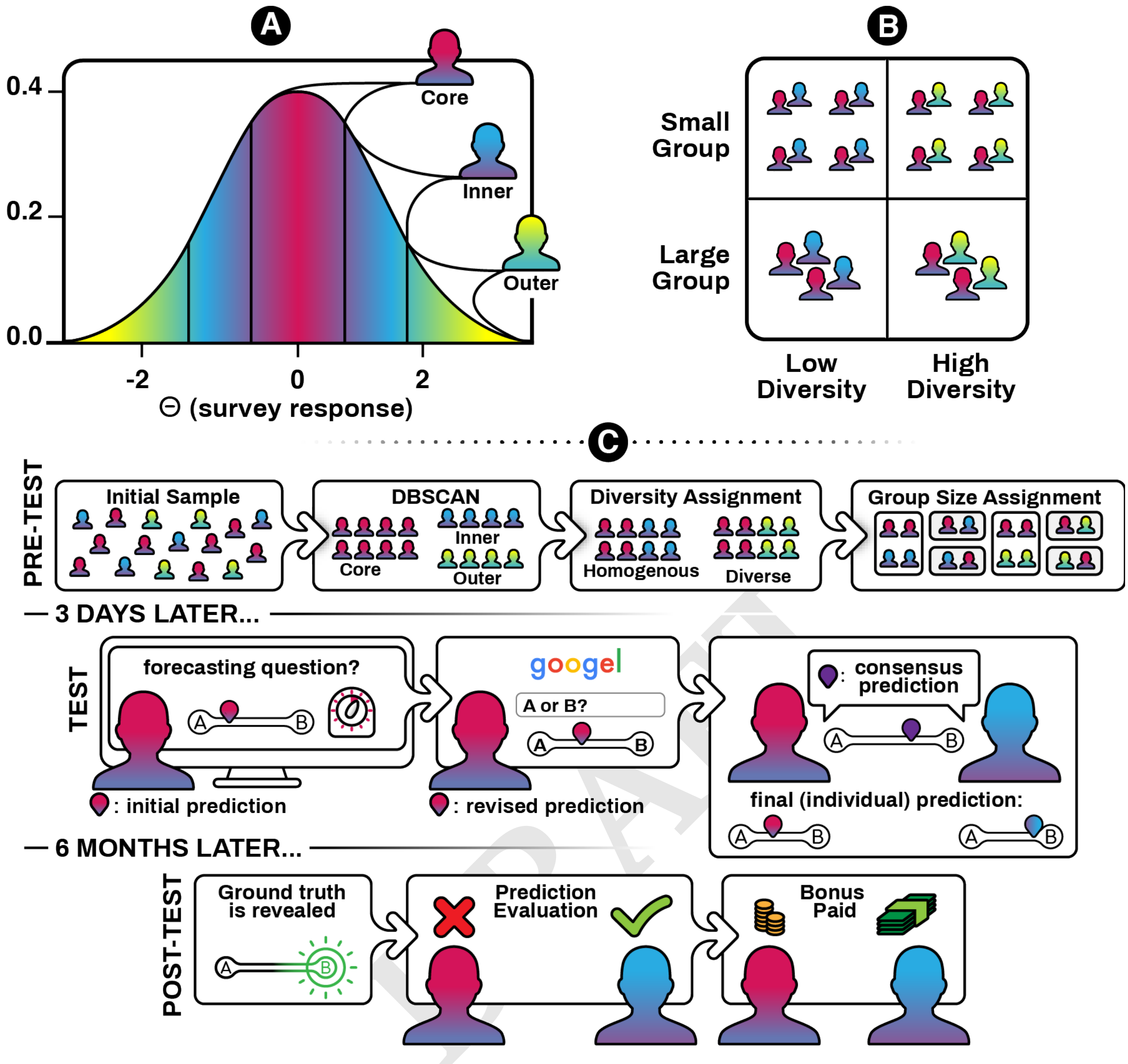

Fig. 1. Experimental design. (a) One dimensional representation of the partitioning of the Θ space by the DBSCAN algorithm. In reality, $\Theta \in \mathbb{R}^D$, where D is the number of dimensions considered ($D = 29$) (b) 2x2 design with factors: diversity (low vs. high) x modularity (low vs. high). Low vs. high diversity manipulation was achieved by matching the core participants to either the inner segment participants (low diversity condition) or the outer segment (high diversity condition). (c) Experimental procedure. At pre-test time (upper row), participants are administered a battery of surveys that are used to cluster them into a core, inner and outer segments (DBSCAN). Core participants are then randomized to a diversity and modularity condition. At test time, they answer eight forecasting problems first alone (Stage 1-2) and then within their groups (Stage 3).

deviance, another popular measure of diversity with multi-dimensional input ($r: .92, p < .001$) (Figure S13).

Orthogonally to diversity, we randomized the size and modularity of the online collective. As the scale of online collaboration widely varies (from small interest groups to large online communities) we want to characterize the effects of group composition as a function of size. Manipulating group size or the number of groups interrogated can have positive effects on group performance, by removing paths through which errors can spread (35–40). Smaller groups are more likely to maximize accuracy in environments characterized by inter-judgment correlations thanks to their inherent noise and greater exploratory behavior (39, 41–45). Furthermore, aggregating information from multiple smaller interacting groups performs better than traditional wisdom-of-crowd because it insulates the aggregate from correlated errors (35). In other words, rather than interrogating one single large crowd ($M = 1$), greater accuracy is obtained by dividing the large crowd into smaller but independent (*i.e.*, non communicating) groups ($M > 1$). We call this feature modularity. Modularity maintains information diversity (across groups) in spite of herding (within groups). However, the study by Navajas et

Individual Forecasting Problems (IFPs)	Truth revealed	Ground truth
1. Before 1 August 2018, will the Moroccan government and the Polisario Front meet for official negotiations over Western Sahara?	2018-08-03	0
2. Before 8 September 2018, will Poland, Estonia, Latvia, or Lithuania accuse Russia of intervening militarily in its territory without permission?	2018-09-10	0
3. Before 8 September 2018, will Saudi Arabia announce that it is ending the blockade of Yemen's Hudaydah port?	2018-09-10	0
4. Will Fidesz and KDNP win 133 or more seats in Hungary's upcoming parliamentary election?	2018-04-11	1
5. Will a Loya Jirga convene in Afghanistan before 8 September 2018?	2018-09-10	0
6. Will any NATO member invoke Article 4 or Article 5 before 8 September 2018?	2018-09-10	0
7. Will the Council of the European Union make an Article 7.1 determination against a member state before 8 September 2018?	2018-09-10	0
8. Will Turkish President Recep Tayyip Erdoğan experience a significant leadership disruption by 31 August 2018?	2018-09-04	0

Table 1. Individual forecasting problems. All IFPs were formulated within the IARPA HFC tournament, and thus represent independent decision-problems. Ground truths were revealed by the IARPA HFC tournament (hence also independently from experimenters' biases) and on the dates specified above (YYYY-MM-DD format). Ground truths are represented on the right column: 0 = the event did not occur; 1 = the event did occur. Question order was randomized for each group. Distribution of forecasts across questions and signal detection theoretical analysis of response bias is provided to show that the results cannot be explained by a general tendency for low probabilities (Figure S8-9).

al. (35) was performed on estimation tasks, where crowds are known to perform well (46). Whether the same results generalize to more complex real-world problems is unknown.

After sorting people into groups of different sizes and composition, participants were asked to give for each forecasting problem an initial guess (initial forecast). Then they were asked to revise it after privately browsing online (revised forecast), and after debating with others online (private final forecast and group consensus forecast). A pre-registration of our hypotheses is available via OSF. At the individual level, we expected alignment of opinions and improved accuracy due to online browsing and social influence. At the aggregate level, we expected group diversity and modularity to positively affect aggregate performance. No predictions were made regarding the direction of their interaction. Exact analyses were not pre-registered. Aggregation followed the same procedure described in (35). Small groups (∼5 people) were approximately the square root of large groups (∼25 people), cf. (36).

The closer (more similar) individuals were on the profiling space the more correlated their forecasts became after online searches. Group diversity benefited individual and aggregated performance and interacted with group size so that large groups benefited from it more than smaller ones. Analysis of forecasts distributions and exploratory linguistic analysis of chat data showed slower consensus building, greater disagreement, and greater variance in group members' performance impacting large diverse groups less negatively than small ones. We also find that forcing individuals to reach a consensus as opposed to simply being exposed to social information benefits their ability to forecast future events. These findings inform how social interaction online can affect real-life problem solving in complex information environments. We discuss these results in light of the recent literature on collective behavior in ecology and social science.

Results

Multi-dimensional profiling Exploratory analyses were ran to characterize our multi-trait diversity measure. Trait diversity correlated with information diversity only after (but not prior) online browsing. After browsing, larger Euclidean distance along the profiling space Θ between pairs of individuals was inversely related to the correlation coefficient of the forecasts made by the same two individuals ($initial: r = 0.12, p = 0.38; Revised: r = -0.39, p = 0.006; Final: -0.056, p < 0.001$). This indicates greater alignment of beliefs proportionally to individual similarity as a function of online browsing.

A principal component analysis was ran to characterize post-hoc the multi-trait distribution of our sample. Trait variation in our population was highly structured, about five components explained about 90% of the variance (Figure S13), suggesting most trait dimensions were redundant or showed little variation. Principal components correlated with ethnic-cultural and socio-political variability in our sample (Figure S14-16). The structure of participants segmentation was already visible on a low-dimensional principal component projection. This result confirmed that core participants were more similar (along the principal components) to participants belonging to the inner segment than to participants belonging to the outer segment (Figure S17). A parallel analysis (Figure S18) suggested to retain eight principal components, reported in Supplementary information. No principal component was trivially related with opinion diversity or performance (Figures S22-23).

Individual-level performance For each forecast, a Brier error score (range 0-2) was computed according to Equation 1. Distributions of individual and aggregated errors are reported in Supplementary material (Figure S2). Errors were larger (worse performance) for initial ($\beta = 0.62, SE =$

$0.09, t = 6.88, p < 5.81e-12$), revised ($\beta = 0.69, SE = 0.08, t = 7.77, p < 7.73e-15$) and final ($\beta = 0.23, SE = 0.09, t = 2.39, p = 0.01$) forecasts compared to consensus forecasts (Figure 2a), indicating an overall forecast improvement over repeated judgments (Table 2A-S3). Against our pre-registered hypotheses, initial forecasts were numerically but non-significantly better than revised forecasts. Both initial and revised forecasts however were worse than following forecasts ($\beta s < -0.38, SEs < 0.09, ts < -5.12, ps < 2.94e-07$), confirming our pre-registered hypothesis of an accuracy improvement due to social interaction (47). Final and consensus forecasts contained the same socially acquired information and were made in random order. Surprisingly, errors were smaller for the consensus than the final forecast. This difference suggests that forcing consensus (rather than simple social exposure) can improve individual forecasting accuracy.

We conducted an exploratory analysis on the effects that diversity (reference: Low) and group size (reference: Large) assignment had on individual forecasting accuracy (Table 2B-S4). Initial and revised forecasts were not affected by our manipulation and were thus excluded from this analysis.

Notice that at the individual level, we can only test whether interacting in small or larger groups has an effect on forecasting error, given that modularity is a group-level feature (see Supplementary information §2). A model with an interaction term was superior to one without, notwithstanding the added complexity ($df = 8, \chi^2 = 7.63, \chi^2 df = 1, p = 0.005$). Working in diverse groups marginally predicted better individual performance ($\beta = -0.37, SE = 0.20, t = -1.83, p = 0.06$). Participants in homogeneous small groups performed non significantly worse their counterparts in homogeneous larger groups ($\beta = -0.20, SE = 0.20, t = -0.99, p = 0.31$). The beneficial effect of diversity on individual performance was positively affected by group size, suggesting that individual interaction with diverse peers was more beneficial in large than small groups ($\beta = 0.82, SE = 0.29, t = 2.85, p = 0.004$) (Figure 2b). The same interaction was found when using average multi-trait distance rather than categorical group assignment as a measure of diversity, (Table S5, Figure S3).

Group-level performance In forecasting like in democratic decisions, aggregated individual judgments are more informative than individual ones. At the aggregate level, we can now ask whether modularity and hierarchical aggregation can improve forecasting accuracy (35, 36). For each group, we computed an aggregate forecast by taking the median forecast in the group for each forecast type. By definition, we have only one group per diversity treatment in the non-modular condition ($M = 1$), but multiple subgroups in the modular condition ($M > 1$). Thus, aggregating judgments in the high modularity condition proceeded by aggregating forecasts in each group first, and then aggregating aggregates(35). An exploratory analysis, showed that consensus forecasting errors were lower than both initial ($\beta = 0.68, SE = 0.22, t = 2.97, p = 0.002$) and revised ($\beta = 0.59, SE = 0.23, t = 2.60, p = .009$) errors, suggesting a benefit of social interaction (Table 2C-S8). The advantage of consensus over final forecasts disappeared at the aggregate level ($\beta = -0.12, SE = 0.29, t = -0.43, p = .66$) (Figure 3a).

Our main hypotheses consisted in analyzing the effect of group assignment on aggregated forecasting errors during the social exchange. A model with fixed effects for diversity, modularity and an interaction between the two provided better fit than one without interaction ($df = 7, \chi^2 = 6.10, \chi^2 df = 1, p = 0.01$). As predicted, aggregate forecasts from diverse groups were better than aggregate forecasts from homogeneous groups ($\beta = -0.56, SE = 0.23, t = -2.39, p = 0.01$) (baseline: large, Table 2D-S9). Also as predicted, aggregated forecasts obtained from smaller/modular groups were better than from larger/non-modular groups ($\beta = -0.82, SE = 0.26, t = -3.10, p = 0.001$) (baseline: homogeneous). Finally, we found an interaction between diversity and modularity whose direction we did not predict ($\beta = 0.93, SE = 0.38, t = 2.43, p = 0.01$), indicating that the beneficial effect of diversity on aggregate forecasting accuracy was significantly greater in large groups over smaller groups (Figure 3b).

Disagreement, consensus reaching and performance variability. To understand why diversity interacted with group size, we performed three main exploratory analyses. First, we analyzed the distribution of forecasts produced by each group in different questions (Figure S2). In particular, we were interested in the disagreement between participants' estimates (diversity of opinions in (50)), namely the dispersion (standard deviation) of the forecast distribution within a group. A greater standard deviation suggests more conflicting views and thus more conflicting evidence for the group to resolve when trying to reach a consensus under time pressure. Compared to initial forecasts, disagreement was lower in final forecasts ($\beta = -4.41, SE = 1.18, t = -3.72, p < .001$) and higher in revised forecasts ($\beta = 5.06, SE = 1.18, t = 4.27, p < .001$), suggesting (surprisingly) an increase in the spread of opinions after online information search and (unsurprisingly) opinion alignment after social interaction (Table S14). We found no main effects of diversity ($\beta = -0.48, SE = 2.36, t = -0.20, p > .8$) or group size ($\beta = -3.51, SE = 1.80, t = -1.94, p > .05$). However, diversity interacted with group size suggesting that it had a smaller effect on disagreement in large groups compared to small ones ($\beta = 7.11, SE = 2.60, t = 2.73, p = .006$). Residual disagreement remained even after people had the chance to come to a consensus, as observed in final forecasts (Figure 4a).

Our second analysis, suggests that online information foraging affected within-group variability in performance. Larger variability indicates that a group contains members who are very accurate (on average across the eight IFPs) and members who are quite poor. Performance variability is typically associated with reduced collective intelligence ((51, 52)). In the initial stage people's accuracy was similar to each other (around 0.1-0.2 standard deviations of Brier scores), but variability increased in small diverse groups after online information foraging. This effect was not as nearly as pronounced for small homogeneous groups and large groups

(A) Individual forecasting error as a function of forecast type

Effect	Estimate	Fitted Brier score	SE	t	p
Intercept	-2.14224	0.1173915	0.24230	-8.841	**$<$ 2e-16**
Initial	0.62237	0.2187395	0.09040	6.884	**5.81e-12**
Revised	0.69532	0.2352946	0.08947	7.772	**7.73e-15**
Final	0.23849	0.1490093	0.09979	2.390	**0.0169**

(B) Individual forecasting error as a function of Diversity and Group size

Effect	Estimate	Fitted Brier score	SE	t	p
Intercept	-1.96631	0.139972	0.30877	-6.368	**1.91e-10**
Final	0.20997	0.1726759	0.07814	2.687	**0.00720**
Diverse	-0.37285	0.1189339	0.20278	-1.839	*0.06595*
Small	-0.20011	0.1413602	0.20094	-0.996	0.31932
Diverse:Small	0.82956	0.2231896	0.29025	2.858	**0.00426**

(C) Aggregated forecasting error as a function of forecast type

Effect	Estimate	Fitted Brier score	SE	t	p
Intercept	-1.8387	0.1590198	0.2508	-7.331	**2.29e-13**
Initial	0.6815	0.3143683	0.2293	2.972	**0.00296**
Revised	0.5999	0.2897372	0.2301	2.607	**0.00913**
Final	-0.1281	0.1398955	0.2964	-0.432	0.66557

(D) Aggregated forecasting error as a function of Diversity and Modularity

Effect	Estimate	Fitted Brier score	SE	t	p
Intercept	-1.76627	0.1709691	0.33428	-5.284	**1.27e-07**
Final	-0.06877	0.15960632	0.15360	-0.448	0.65434
Diverse	-0.56382	0.09082084	0.23514	-2.398	**0.01649**
Modular	-0.82268	0.07010727	0.26515	-3.103	**0.00192**
Diverse:Modular	0.93267	0.10137943	0.38254	2.438	**0.01477**

Table 2. Generalized mixed-effects models on individual and aggregated errors. Table of analysis on forecasting errors (in Brier scores) for individual (A-B) and aggregated measures (C-D) and as a function of forecast type (A-C) and condition (B-D). Baselines for each factor: consensus, homogeneous, large/non-modular. The effect of final forecasts on individual errors (A) and the effects of diversity and the interaction between diversity and modularity (D) did not survive a Bonferroni correction. Boldface: p<.05; Italics: p<.10. Tables B-C represent exploratory analyses. Hypotheses in tables A and D were preregistered. All analyses were also repeated with binarized accuracy (Tables S10-13) and logit link function (Table S16). For convenience, all tests refer to two-sided hypotheses and were calculated with the lmerTest package in R (48)

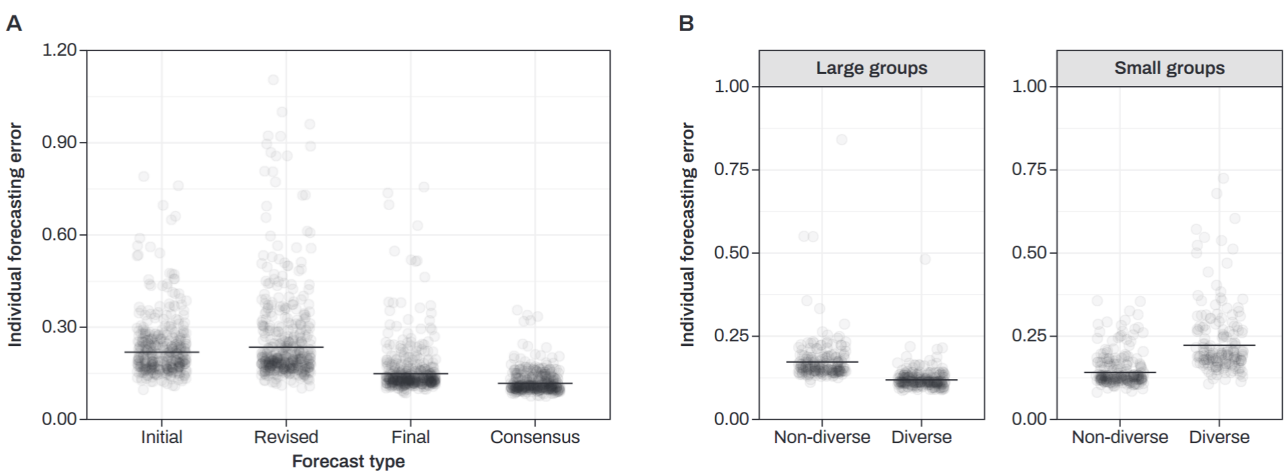

Fig. 2. Individual-level analysis. (a) Partial residuals plot showing the effect of forecasting type on individual forecasting error (measured in Brier scores). Lower numbers represent higher accuracy. Solid lines represent model fit. (b) Partial residuals plot showing the effect of diversity and group size on individual forecasting error (expressed in Brier scores). Solid lines represent model fit. Notice that, for visualization purposes, the graphs has been plotted onto the original error scale rather than log scale as in the fitted GLMM. Thus, large residuals should not cause concern (49). See Figure S10 when using a logit link. Source data are provided as a Source Data file.

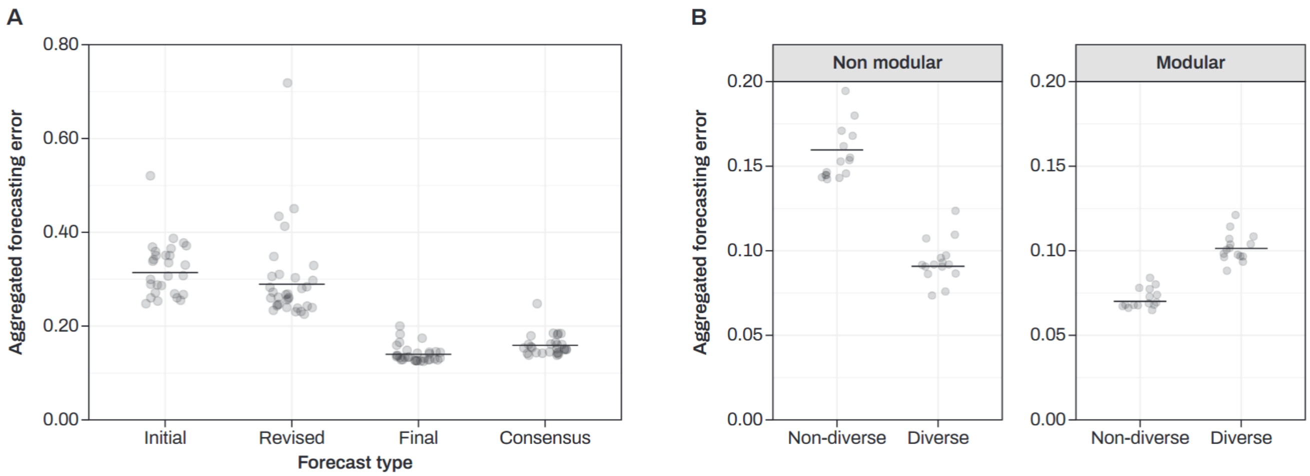

Fig. 3. Group-level analysis. Individual forecasts were aggregated for each forecast type, first within each group and then across groups in each treatment. (a) Partial residuals plot showing the effect of forecasting type on aggregated forecasting error (measured in Brier scores). Lower numbers represent higher accuracy. Solid lines represent model fit. (b) Partial residuals plot showing the effect of diversity and modularity on aggregated forecasting error. Solid lines represent model fit. Notice that the graphs have been plotted onto the original error scale. See Figure S11 when using a logit link. Source data are provided as a Source Data file.

(Figure 4b), suggesting that browsing selectively negatively impacted small diverse groups.

A third factor we investigated was whether our manipulation affected the process of consensus reaching through online deliberation (see Supplementary Information §5-6). We manually labelled forecast estimates mentioned by participants during the deliberation phase and fitted a model representing convergence of these estimates to the consensus forecast. Group diversity decreased consensus reaching times ($\beta = -0.31, SE = 0.12, t = -2.55, p = .01$, baseline: large). Also small groups showed quicker consensus reaching than large ones ($\beta = -0.46, SE = 0.10, t = -4.68, p < .001$, baseline: homogeneous) (Table S15). A positive interaction between the two factors indicated that speed in consensus reaching observed in diverse groups decreased as a function of smaller group size ($\beta = 0.69, SE = 0.17, t = 4.004, p < .001$) (Figure S6-7).

Methods

Procedure The study was approved by MIT Institutional Review Board. Participants (N=193) gave informed consent before joining the study. Three days before test (pre-test), participants answered a battery of demographic, cognitive and personality questions that was used to map them on a multi-dimensional space Θ. We used an unsupervised clustering algorithm (DBSCAN) to label participants as belonging to the center mass of the distribution (core segment) or its tail (inner and outer segments, Figure 1a). This structure was already visible on a low-dimensional projection of participants on the first two principal components of the data (Figure S17).

We manipulated group diversity (low vs. high) and crowd modularity (low vs. high) (Figure 1b). Core participants ($\sim 50\%$ of our initial sample) were randomly assigned (a) to work with either close (inner segment, $\sim 25\%$ of our sample) or distant (outer segment, $\sim 25\%$ of our sample) participants on the feature space, and (b) to work in small ($\sim$5 people) or large ($\sim$25 people) groups (Figure 1c). During the experiment (test phase), participants answered 8 individual forecasting problems (IFPs), randomly selected from a larger pool of binary real geopolitical forecasting problems released within IARPA's Hybrid Forecasting Competition and unresolved (*i.e.*, whose solution was unknown) at the time of the experiment. The exact problems selected were not preregistered. For each IFP, participants went through three timed consecutive stages. During stage one, participants answered a binary forecasting problem (Table 1) and had to enter an Initial private forecast off the top of their heads (initial). During stage two, they had to search relevant information online, using their browser, and enter a revised private forecast (revised). Finally, during the third and last stage, participants discussed in real time their views using an inbuilt chat (Figure 1c). During this stage, participants had to agree on a joint forecast (consensus) as well as giving their final private forecast (final). Notice that although consensus forecasts in a group had to be the same final forecasts could differ, thus allowing us to capture residual disagreement existing between group members after interaction had taken place. Participants were rewarded both for their time and - about six months later (post-test) when the ground truths were revealed - for accurate predictions. Performance was evaluated using Brier scores, a quadratic error score used in forecasting for its proper scoring properties, *i.e.*, a scoring rule incentivizing honest responding. For a binary question, a Brier score is computed as:

$$b = (o-p)^2 + (\bar{o}-\bar{p})^2 \qquad \textbf{(1)}$$

where p represents the predicted event probability (range $[0,1]$) and o is the indicator variable for the observed event (0: the event happened; 1: the event did not happen). $\bar{p}$ and $\bar{o}$ represent complementary probabilities. A Brier score of 0 represent a fully predicted event (*i.e.*, no uncertainty), while a

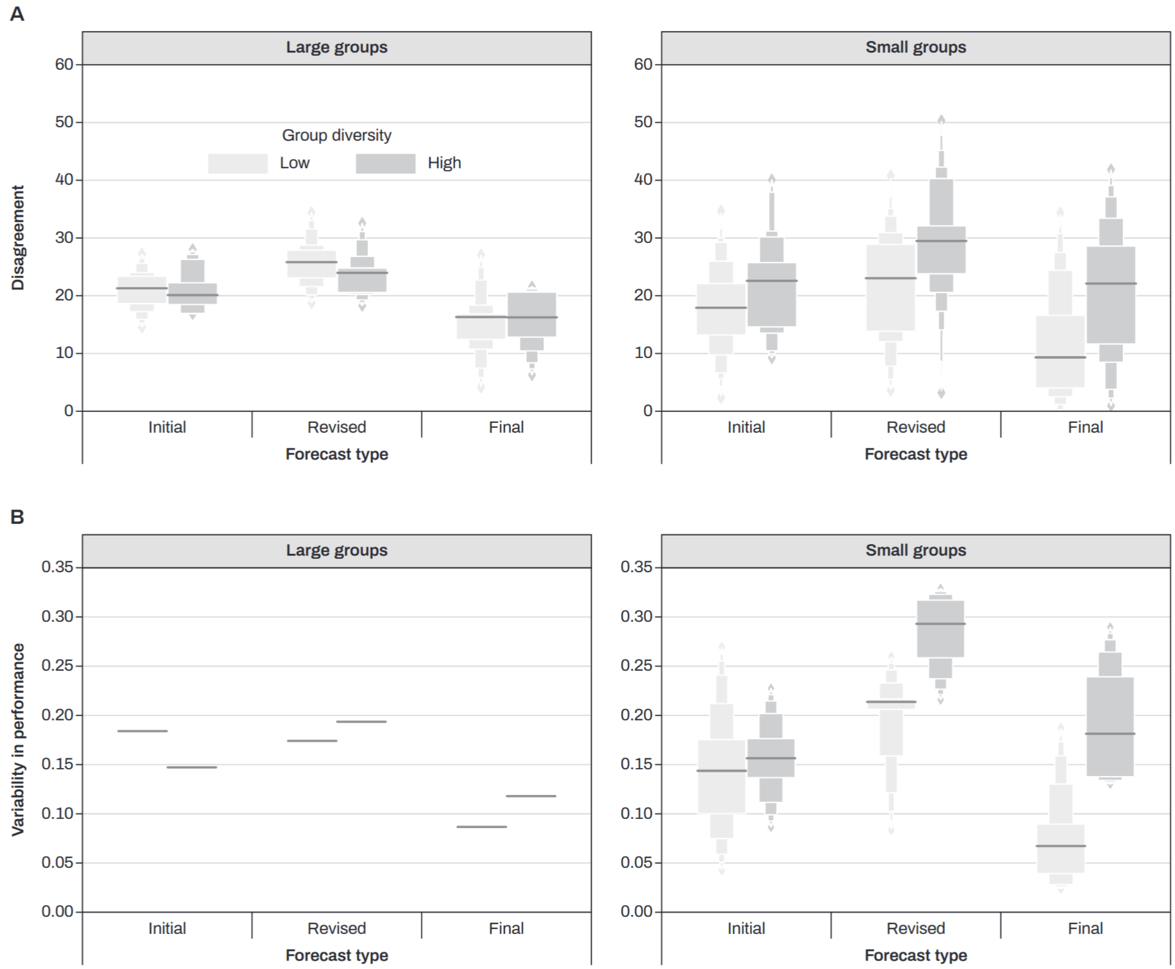

Fig. 4. Disagreement and variability in performance. (a) Distributions of opinion disagreement as a function of forecasting stage, group trait diversity and group size. Opinion disagreement is calculated as the standard deviation over group members' forecasts. (b) Performance variability as a function of forecasting stage, group trait diversity and group size. Performance variability is the standard deviation over average individual performance in a group. Larger values indicate that a group contains members who are very good and members who are quite poor (on average across the eight IFPs). Notice that a single value of performance variability exists for large groups, but not for small groups (m=6 and m=4 for small low and high diversity groups respectively). Notice also that for both panels consensus forecasts were removed because, by definition, they did not produce meaningful variation in these measures. Box areas correspond to distribution ideal tail areas of .50, .25, .125, .0625 (53). Source data are provided as a Source Data file.

Brier score of 2 represents a gross forecasting error (the forecaster predicted with absolute confidence the event would occur and it did not, or viceversa). Notice that Brier scores measure second-order accuracy, meaning that they punish over- (and under-)confidence rather than number of incorrect binary judgments. An improvement in Brier score represents a more precise probabilistic forecast, which might not necessarily reflect how often a participant is right (first order accuracy). For these reasons, Brier scores represent the standard in forecasting. (47, 54, 55).

Analyses Errors were fitted with multi-level generalized linear mixed-effects models (GLMM) with Gaussian log link function. The results are robust across alternative link functions, like probit and logit (Table S16). All analyses, unless specified, were limited only to participants who fell in the core segment (*i.e.*, test participants), as these were the only ones to whom the randomization procedure applied. This allows us to draw causal inferences on the effect of our manipulation, as all core participants were equal in expectation. Our main analyses corresponding to our preregistered hypotheses are reported in Table 2a and d. They included at the individual level the effect of forecast type, and the aggregate level the effect of group diversity and size assignment. To provide a full picture, we complement the main analyses with the effect of the manipulation on individual errors (Table 2b) and the effect of forecast type on aggregate errors (Table 2c).

Also according to our preregistered hypotheses, we analyzed within-group disagreement at each stage of the experiment (Table S14). Disagreement was defined as the standard deviation of the forecast within a group, broken down by forecast type and condition. We also run a set of exploratory analyses on chat data, aimed at understanding how individuals integrated private information to reach a consensus within their group (see Supplementary material §5-6).

Statistics and Reproducibility The experiment was repeated only once. A pilot experiment had been previously discarded (data never analysed) due to a bug in the web aplication.

Discussion

In this study, we experimentally manipulated the diversity and the modularity of online collectives collaboratively performing a real-life complex forecasting task. We found that sorting groups based on a composite measure of diversity—including demographic, relational and cognitive indicators—affected the correlation of beliefs of people only after people were asked to gather information online and interact with others to forecast the future. Both social interaction and the need to reach an internal consensus via deliberation improved people's forecasting accuracy. Collaborating in diverse groups was beneficial for people's individual ability to forecast the future, proportionally to group size (Figure 2). When aggregating judgments together using a simple median, this translated into an advantage of diverse groups and modular groups, and an interaction between diversity and group size (Figure 3). We explored the mechanisms underlying this interaction with a range of exploratory analyses (Figure 4).

The widespread use of automated content recommendation paired with people's tendency to interact with others who share similar characteristics is thought to create insulated online information bubbles. There is growing concern that this tendency might have negative long term consequences on political and democratic institutions, as citizens form partial or inaccurate representations of the world. Although we cannot answer these important questions with our study, we tried to characterize the effect that interacting with peers who differ along an arbitrary large profiling space has on the forecasting accuracy achieved by in-expectation-identical people (core segment participants) as a function of group size. We provided preliminary evidence that the ability of an online collective to solve complex geo-political forecasting tasks, under conditions of uncertainty and time pressure, may be coupled with their digital ecosystem. People's shared traits did not predict a priori how correlated their beliefs about world events were. Instead belief coupling happened only after they interacted with their unique information silos via their web browsers. Forecasts became correlated only after online browsing, and proportionally to people's similarity on our multi-trait profiling space. In other words, our operationalization of trait-similarity had measurable effects on the online information a group could tap into. This is in contrast with offline settings, where trait diversity does not directly impact information diversity (20, 22, 23, 26, 33, 34). The use of an experimental methodology bypasses the limitations of observational approaches, strengthening causal inference (22, 23, 27, 56). Trait similarity in our experiment largely captured participants' variability along interpretable ethnic-cultural and socio-political variables (Figures S13-18). Arguably, these features affect political judgments and the type of content that a person is likely to retrieve online. Although these findings suggest possible causal pathways between trait similarity and the effects described, they also raise worries that these dimensions may be used by search engines to skew information retrieval during online searches. This effect was not among our pre-registered hypotheses so we warn caution in overinterpreting this finding. Future studies should attempt a replication.

Our findings also suggest the importance of diversity in online settings characterized by large collectives. Given the difficulty and domain specificity of the questions in our experiment, increasing diversity may have increased the chance that at least one of the participants in a group could, for example, recall what a Loya Jirga is and make an informed guess. This effect would be more pronounced in a large group than a small group. To illustrate this, imagine asking a group of scientists this question: "Is *Campephilus principalis* likely to become an invasive species throughout Australia in the next 20 years?". If we select a discipline at random, and then make large or small groups they would be unlikely to know what *Campephilus principalis* is and would guess Yes with some probability greater than zero. Now, if we compose groups of scientists randomly chosen across disciplines, a small group does not do much better than a group from a single discipline because the odds of containing an ornithologist remains low. However, the odds of stumbling upon an ornithologist increase with group-size and a finite number of academic disciplines. If there happens to be an ornithologist, they can trivially identify the answer to this question as No (this species, also called Ivory-Billed Woodpeckers, is largely believed to be extinct). Similarly for political questions, imagine we have a set of questions from across a large range of countries or cultures, all of which are obviously unlikely to anyone with domain-knowledge. Diversity would improve forecasting in large, but not small groups, because large groups have an increased chance of containing an expert. Critically, because the probability of the events is low (Table 1), Brier error will be high in anyone without domain knowledge that assumes the events have closer to equal probability of occurring. Although this logic nicely explains the beneficial effect of diversity observed in large groups, it lacks explanatory power in other respects. First, it does not explain why we observed a symmetrical effect in small groups instead of no effect at all (Figures 2b and 3b). Second, it does not explain why differences among groups largely emerged after the revision and social stages rather than during stage one. Finally, it is unclear why performance variability remained similar between large diverse and homogeneous groups, notwithstanding a supposedly different concentration of domain-experts (Figure 4). Thus, although these statistical considerations are certainly relevant, technological (individuals interacting with their search engines) and social (individuals interacting with each other) aspects are also an important part of the story. Importantly, alternative measures of diversity and more theory-driven profiling should be considered in the future to address these concerns. For the scope of our paper however, the specific implementation of group diversity was not as important as its functional value in influencing information foraging and error distributions in online groups. Characterizing measures of diversity is a research field in its own right. We recognize that our method is not perfect and caution should be used when trying to generalize our results.

Investigating collective decisions under extreme conditions is highly informative. Many decisions faced by intelligence analysts as well as normal people everyday are characterized by weak signal, uncertainty, time pressure or short collective attention, namely all conditions under which rational deliberation is least effective (11, 12, 57). The specific forecasting problems asked in the task were a random subsample of forecasting problems that were selected by a national forecasting tournament (Hybrid Forecasting Competition) to be a representative sample of geo-political forecasting. They required domain knowledge that participants were unlikely to possess prior to online browsing. This feature also served a precise design purpose. The specificity of the forecasting problems ensured that group discussions were driven by the content that was collectively retrieved online rather than biased by what participants knew in advance. Group members had only a short amount of time to forage for relevant online content. The ability of a group to collectively search relevant information in parallel was thus, arguably, more important than the ability of each individual to search any piece of information thoroughly. Finally, another thing to notice is that most events did not occur (Table 1). This is not uncommon in forecasting. Rare events are often the most consequential and difficult to predict, as the covid-19 pandemic shows. Being able to predict rare events resides at the heart of accurate forecasting (47, 58). In these circumstances, an unspecific bias towards deeming events unlikely to occur would generally pay off, and generate few highly consequential mistakes. To rule out the confound of an unspecific bias, we first ran a signal detection analysis that indicated that people did not show any initial bias toward uncritically deeming events as rare (Figure S9). Thus, it is unclear why an unspecific tendency toward answer low probability (confidently believing the events were unlikely) would emerge from online browsing or social interaction. Social interaction is known to extremize initially held individual opinions, a phenomenon known in psychology as risky-shift (59). Thus, if anything one would expect social interaction in our experiment to pull initial predictions toward 0% and 100% symmetrically. Instead, group discussions seemed to adjust initial predictions intentionally towards the correct response. Furthermore, the unspecific bias explanation does not account for the interaction between group diversity and group size observed. Manual labelling of chat conversations revealed that about half of people in each group had at least some knowledge about each topic, and conversations mainly revolved around evidence in favor or against each option. Although it is difficult to disentangle whether domain-specific knowledge was due to prior beliefs or online browsing, the former explanation is unlikely due to initial forecasts being distributed around chance (Figure S9). For this reason we concluded that the observed accuracy improvement was more likely due to online browsing and group deliberation, rather than an unspecific bias towards reducing probability.

In line with recent work in collective behavior, we find that when decision-makers are not independent group accuracy can benefit from a reduced group size and increased modularity (36, 39, 41, 42, 60). Research in social learning (61) has shown that group outcomes are affected by a complex interplay among several factors, including learning strategies, task complexity, modularity and network structure. The present study showed how two factors that independently reduce correlated errors, namely diversity and modularity, can interact in unexpected ways (17, 35, 36). To characterize this novel interaction, we described information aggregation using a range of exploratory analyses, such as within-group disagreement (Figure 4a), convergence speed to consensus forecast (Supplementary material §6) and performance variability among group members (Figure 4b). The latter is often a pre-requisite for good group performance in the literature on collective intelligence (51, 52, 62).
Notwithstanding the value of these results, we would like to raise a word of caution. In particular, as specified in our preregistration, we had no expectations on the direction of the interaction between diversity and group size before testing our model. Similarly, many analyses were exploratory in nature and cannot be used to draw final conclusions. Future studies will need to address whether the result can be replicated. Speculatively however, our results suggest that, given the difficulty in reducing homophily and self-assortativity in large online crowds, one might try instead to increase their modularity. Crucially, we stress the importance of addressing this debate also on ethical grounds. Here, utilitarian and deontological approaches must be reconciled to inform practitioners and businesses (63).

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract number 2017-17061500006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The authors would like to thank Zhaozheng Alice Jin for her contribution to the development of the experimental web application; Dr. Nick Obradovich for his support throughout this work; and Dr. Hans van Dijk for sharing his data and insights.

Replicability

Data Availability. Research data supporting the findings of this study have been deposited in Open Science Framework. Pescetelli, N., Rutherford, A., & Rahwan, I. (2020, July 6). Multi-trait diversity improves forecasting accuracy in large but not small online groups. Data can be retrieved using the permanent link: `osf.io/wb538`. A Reporting Summary for this article is available as a Supplementary Information. Source data are provided with this paper.

Code Availability. Code to replicate analysis and figures supporting the findings of this study have been deposited in Open Science Framework. Pescetelli, N., Rutherford, A.,

& Rahwan, I. (2020, July 6). Multi-trait diversity of online groups improves geo-political forecasting accuracy as a function of group size. Data can be retrieved using the permanent link: osf.io/wb538.

Preregistration material. Pre-registration material is available via AsPredicted.org: https://aspredicted.org/9m6df.pdf.

Authors contributions

Conceptualization: NP; Data curation: NP; Formal Analysis: NP and AR; Funding acquisition: IR; Investigation: NP; Methodology: NP; Project administration: NP; Resources: NP, AR, IR; Software: NP and AR; Supervision: AR and IR; Validation: NP and AR; Visualization: NP and AR; Writing – original draft: NP and AR; Writing – review and editing: NP, AR, IR.

Competing interests

The authors declare no conflicts of interest.

Corresponding author

Niccolo Pescetelli: pescetelli [at] mpib-berlin.mpg.de

Bibliography

1. Eli Pariser. *The Filter Bubble: What The Internet Is Hiding From You.* Penguin, London, 2011.
2. Ronald E. Robertson, David Lazer, and Christo Wilson. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 955–965, New York, New York, USA, 2018. ACM Press. ISBN 9781450356398. doi: 10.1145/3178876.3186143.
3. Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 8 2015. ISSN 0027-8424. doi: 10.1073/pnas.1419828112.
4. A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proc. of the 16th Int.Conf. on World Wide Web*, page 271–280, 2007.
5. Bracha Shapira and Boaz Zabar. Personalized search: Integrating collaboration and social networks. *Journal of the American Society for Information Science and Technology*, 62(1):146–160, 1 2011. ISSN 15322882. doi: 10.1002/asi.21446.
6. Qiaozhu Mei and Kenneth Church. Entropy of search logs. In *Proceedings of the international conference on Web search and web data mining - WSDM '08*, page 45, New York, New York, USA, 2008. ACM Press. ISBN 9781595939279. doi: 10.1145/1341531.1341540.
7. Amruta Joshi, Abraham Bagherjeiran, and Adwait Ratnaparkhi. User Demographic and Behavioral Targeting for ContentMatch Advertising. In *Data Mining and Audience Intelligence for Advertising(ADKDD 2011) in conjunction with SIGKDD'11*, page 57, 2011.
8. Maria Mellor. Why is TikTok creating filter bubbles based on your race? *Wired*, 2020.
9. Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. Me, My Echo Chamber, and I. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 823–831, New York, New York, USA, 2018. ACM Press. ISBN 9781450356398. doi: 10.1145/3178876.3186130.
10. Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 8 2001. ISSN 0360-0572. doi: 10.1146/annurev.soc.27.1.415.
11. Hugo Mercier. The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences*, 20(9):689–700, 9 2016. ISSN 13646613. doi: 10.1016/j.tics.2016.07.001.
12. Hélène Landemore. *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton University Press, 2013.
13. Alejandro Portes and Erik Vickstrom. Diversity, Social Capital, and Cohesion. *Annual Review of Sociology*, 37(1):461–479, 8 2011. ISSN 0360-0572. doi: 10.1146/annurev-soc-081309-150022.
14. Nancy DiTomaso, Corinne Post, and Rochelle Parks-Yancy. Workforce Diversity and Inequality: Power, Status, and Numbers. *Annual Review of Sociology*, 33(1):473–501, 8 2007. ISSN 0360-0572. doi: 10.1146/annurev.soc.33.040406.131805.
15. Daan van Knippenberg and Michaéla C. Schippers. Work Group Diversity. *Annual Review of Psychology*, 58(1):515–541, 1 2007. ISSN 0066-4308. doi: 10.1146/annurev.psych.58.110405.085546.
16. James Habyarimana, Macartan Humphreys, Daniel N Posner, M Jeremy, and Jeremy M Weinstein. Why Does Ethnic Diversity Undermine Public Goods Provision ? *American Political Science Review*, 101(4):709–725, 2007.
17. Scott E. Page. *The Difference How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies.* Princeton University Press, 2007. ISBN 9780691138541.
18. Stefan Krause, Richard James, Jolyon J. Fariac, Graeme D. Ruxtond, and Jens Krause. Swarm intelligence in humans: diversity can trump ability. *Animal Behaviour*, 81(5):941–948, 2011. doi: 10.1016/j.anbehav.2010.12.018.
19. James Surowiecki. *The Wisdom of Crowds. Why the Many are Smarter than the Few.* Little, Brown Book Group, London, abacus edition, 2004.
20. Samuel R. Sommers. On racial diversity and group decision making: Identifying multiple effects of racial composition on jury deliberations. *Journal of Personality and Social Psychology*, 90(4):597–612, 2006. ISSN 1939-1315. doi: 10.1037/0022-3514.90.4.597.
21. Norbert L. Kerr and R. Scott Tindale. Group Performance and Decision Making. *Annual Review of Psychology*, 55(1):623–655, 2 2004. ISSN 0066-4308. doi: 10.1146/annurev.psych.55.090902.142009.
22. Denise Lewin Loyd, Cynthia S. Wang, Katherine W. Phillips, and Robert B. Lount. Social Category Diversity Promotes Premeeting Elaboration: The Role of Relationship Focus. *Organization Science*, 24(3):757–772, 6 2013. ISSN 1047-7039. doi: 10.1287/orsc.1120.0761.
23. Sheen S. Levine, Evan P. Apfelbaum, Mark Bernard, Valerie L. Bartelt, Edward J. Zajac, and David Stark. Ethnic diversity deflates price bubbles. *Proceedings of the National Academy of Sciences*, 111(52):18524–18529, 12 2014. ISSN 0027-8424. doi: 10.1073/pnas.1407301111.
24. Ilan Yaniv. Group diversity and decision quality: Amplification and attenuation of the framing effect. *International Journal of Forecasting*, 27(1):41–49, 1 2011. ISSN 01692070. doi: 10.1016/j.ijforecast.2010.05.009.
25. Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science (New York, N.Y.)*, 330(6004):686–8, 10 2010. ISSN 1095-9203. doi: 10.1126/science.1193147.
26. Elizabeth Mannix and Margaret A. Neale. What Differences Make a Difference? *Psychological Science in the Public Interest*, 6(2):31–55, 10 2005. ISSN 1529-1006. doi: 10.1111/j.1529-1006.2005.00022.x.
27. Anthony Lising Antonio, Mitchell J. Chang, Kenji Hakuta, David A. Kenny, Shana Levin, and Jeffrey F. Milem. Effects of Racial Diversity on Complex Thinking in College Students. *Psychological Science*, 15(8):507–510, 8 2004. ISSN 0956-7976. doi: 10.1111/j.0956-7976.2004.00710.x.
28. Jürgen Wegge, Carla Roth, Barbara Neubach, Klaus-Helmut Schmidt, and Ruth Kanfer. Age and gender diversity as determinants of performance and health in a public organization: The role of task complexity and group size. *Journal of Applied Psychology*, 93(6):1301–1313, 2008. ISSN 1939-1854. doi: 10.1037/a0012680.
29. M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110 (15):5802–5805, 4 2013. ISSN 0027-8424. doi: 10.1073/pnas.1218772110.
30. D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational Social Science. *Science*, 323(5915):721–723, 2 2009. ISSN 0036-8075. doi: 10.1126/science.1167742.
31. Anikó Hannák, Piotr Sapieżyński, Arash Molavi Khaki, David Lazer, Alan Mislove, and Christo Wilson. Measuring Personalization of Web Search. *arXiv*, 6 2017. doi: 1706.05011.
32. Kevin Granville. Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens, 2018.
33. Stephanie de Oliveira and Richard E. Nisbett. Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences*, 115(9):2066–2071, 2 2018. ISSN 0027-8424. doi: 10.1073/pnas.1717632115.
34. Hans van Dijk, Marloes L. van Engen, and Daan van Knippenberg. Defying conventional wisdom: A meta-analytical examination of the differences between demographic and job-related diversity relationships with performance. *Organizational Behavior and Human Decision Processes*, 119(1):38–53, 9 2012. ISSN 07495978. doi: 10.1016/j.obhdp.2012.06.003.
35. Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, 2(2):126–132, 2 2018. ISSN 2397-3374. doi: 10.1038/s41562-017-0273-4.
36. Albert B. Kao and Iain D. Couzin. Modular structure within groups causes information loss but can improve decision accuracy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1774):20180378, 6 2019. ISSN 0962-8436. doi: 10.1098/rstb.2018.0378.
37. David Lazer and Allan Friedman. The Network Structure of Exploration and Exploitation. *Administrative Science Quarterly*, 52(4):667–694, 12 2007. ISSN 0001-8392. doi: 10.2189/asqu.52.4.667.
38. Daniel Barkoczi and Mirta Galesic. Social learning strategies modify the effect of network structure on group performance. *Nature Communications*, 7:13109, 10 2016. ISSN 2041-1723. doi: 10.1038/ncomms13109.
39. Wataru Toyokawa, Andrew Whalen, and Kevin N. Laland. Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, 3(2):183–193, 2 2019. ISSN 2397-3374. doi: 10.1038/s41562-018-0518-x.
40. Ethan Bernstein, Jesse Shore, and David Lazer. How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences*, 115(35):8734–8739, 8 2018. ISSN 0027-8424. doi: 10.1073/pnas.1802407115.
41. Mirta Galesic, Daniel Barkoczi, and Konstantinos Katsikopoulos. Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision*, 5(1):1–15, 1 2018. ISSN 2325-9973. doi: 10.1037/dec0000059.
42. A. B. Kao and I. D. Couzin. Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings of the Royal Society B: Biological Sciences*, 281(1784):20133305–20133305, 4 2014. ISSN 0962-8452. doi: 10.1098/rspb.2013.3305.
43. Lingfei Wu, Dashun Wang, and James A. Evans. Large teams develop and small teams

disrupt science and technology. *Nature*, 566(7744):378–382, 2 2019. ISSN 0028-0836. doi: 10.1038/s41586-019-0941-9.

44. Nicolas Fay, Naomi De Kleine, Bradley Walker, and Christine A. Caldwell. Increasing population size can inhibit cumulative cultural evolution. *Proceedings of the National Academy of Sciences*, page 201811413, 3 2019. ISSN 0027-8424. doi: 10.1073/pnas.1811413116.
45. M. A. Kline and R. Boyd. Population size predicts technological complexity in Oceania. *Proceedings of the Royal Society B: Biological Sciences*, 277(1693):2559–2564, 8 2010. ISSN 0962-8452. doi: 10.1098/rspb.2010.0452.
46. Jens Krause, Graeme D. Ruxton, and Stefan Krause. Swarm intelligence in animals and humans. *Trends in Ecology & Evolution*, 25(1):28–34, 1 2010. ISSN 01695347. doi: 10.1016/j.tree.2009.06.016.
47. Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown Publishers, New York, 2015. ISBN 9780804136693.
48. Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 2017. ISSN 1548-7660. doi: 10.18637/jss.v082.i13.
49. Patrick Breheny and Woodrow Burchett. Visualization of Regression Models Using visreg. *The R Journal*, 9(2):56–71, 2017.
50. Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America*, 108(22):9020–5, 5 2011. ISSN 1091-6490. doi: 10.1073/pnas.1008636108.
51. Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. Optimally interacting minds. *Science (New York, N.Y.)*, 329(5995):1081–5, 8 2010. ISSN 1095-9203. doi: 10.1126/science.1185718.
52. Justin Kruger and David Dunning. Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of personality and social psychology*, 77(6):1121–1134, 1999.
53. Heike Hofmann, Hadley Wickham, and Karen Kafadar. Letter-Value Plots: Boxplots for Large Data. *Journal of Computational and Graphical Statistics*, 26(3):469–477, 7 2017. ISSN 1061-8600. doi: 10.1080/10618600.2017.1305277.
54. Philip E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton, 2006.
55. Stephen M Fleming, Brian Maniscalco, Yoshiaki Ko, Namema Amendi, Tony Ro, and Hakwan Lau. Action-Specific Disruption of Perceptual Confidence. *Psychological science*, 2014. ISSN 1467-9280. doi: 10.1177/0956797614557697.
56. Samuel R. Sommers, Lindsey S. Warp, and Corrine C. Mahoney. Cognitive effects of racial diversity: White individuals' information processing in heterogeneous groups. *Journal of Experimental Social Psychology*, 44(4):1129–1136, 7 2008. ISSN 00221031. doi: 10.1016/j.jesp.2008.01.003.
57. Philipp Lorenz-Spreen, Bjarke Mørch Mønsted, Philipp Hövel, and Sune Lehmann. Accelerating dynamics of collective attention. *Nature Communications*, 10(1):1759, 12 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09311-w.
58. Nassim Nicholas Taleb. *The black swan: The impact of the highly improbable*. Penguin, London, 2008. ISBN B00FK8YZ5I.
59. Serge Moscovici and Marisa Zavalloni. The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, 12(2):125–135, 1969.
60. Thomas N. Wisdom, Xianfeng Song, and Robert L. Goldstone. Social Learning Strategies in Networked Groups. *Cognitive science*, 37:1383–1425, 2013. doi: 10.1111/cogs.12052.
61. Rachel L. Kendal, Neeltje J. Boogert, Luke Rendell, Kevin N. Laland, Mike Webster, and Patricia L. Jones. Social Learning Strategies: Bridge-Building between Fields. *Trends in Cognitive Sciences*, 22(7):651–665, 7 2018. ISSN 13646613. doi: 10.1016/j.tics.2018.04.003.
62. Ralf H J M Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Patricia A Carney, Andy Bogart, Giuseppe Argenziano, Iris Zalaudek, and Max Wolf. Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1601827113–, 7 2016. ISSN 1091-6490. doi: 10.1073/pnas.1601827113.
63. Hans van Dijk, Marloes van Engen, and Jaap Paauwe. Reframing the Business Case for Diversity: A Values and Virtues Perspective. *Journal of Business Ethics*, 111(1):73–84, 11 2012. ISSN 0167-4544. doi: 10.1007/s10551-012-1434-z.