Learning From Aggregated Opinion

Kerem Oktar[1], Tania Lombrozo*[1], and Thomas L. Griffiths*[1,2]

[1]Department of Psychology, Princeton University

[2]Department of Computer Science, Princeton University

Author Note

This is a pre-publication version of the paper. The final published version is available at: https://doi.org/10.1177/09567976241251741

Kerem Oktar https://orcid.org/0000-0002-0118-5065

We have no known conflicts of interest to disclose.

All pre-registrations, materials, analysis scripts, and data are available at https://osf.io/dexuy.

We are grateful to Shlomi Sher, Kenny Easwaran, and our anonymous reviewers for valuable feedback on this research. This research project and related research were made possible with the support of the NOMIS Foundation.

Correspondence concerning this article should be addressed to Kerem Oktar, Dept. of Psychology, Princeton University, Princeton, NJ 08540. Email: oktar@princeton.edu

*These authors share co-senior authorship on this paper.

Abstract

The capacity to leverage information from others' opinions is a hallmark of human cognition. Past research has thus investigated how we learn from others' testimony. Yet a distinct form of social information—aggregated opinion—increasingly guides our judgments and decisions. We investigated how people learn from such information by conducting three experiments with participants recruited online within the United States (N=886) comparing the predictions of three computational models: a Bayesian solution to this problem that can be implemented by a simple strategy for combining proportions with prior beliefs, and two alternatives from epistemology and economics. Across all studies, we found the strongest concordance between participants' judgments and the predictions of the Bayesian model, though some participants' judgments were better captured by alternative strategies. These findings show that people draw systematic inferences from aggregated opinion, often in line with a Bayesian solution, and lay the groundwork for future research.

Keywords: disagreement, opinion, Bayesian inference, judgment, belief, heuristics

Research Transparency Statement

General Disclosures Conflicts of interest: We have no known conflicts of interest to disclose. Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article. Ethics: This research complies with the Declaration of Helsinki (2023), and all experiments were approved by the [Author Institution] Institutional Review Board (IRB). Computational reproducibility: All pre-registrations, data, and analysis scripts are available in our open-access OSF repository for any parties interested in reproducing our results; see https://osf.io/dexuy/?view_only=c921857784424142aae36164b1a0f60f. Any deviations from the pre-registered analyses are explicitly identified in the manuscript and supplementary materials.

Study One Disclosures. Preregistration: The hypotheses, key analyses, and methods were pre-registered on 05/12/2022 (see https://aspredicted.org/L3D_G4Q) prior to data collection on 5/24/2022.

Study Two Disclosures. Preregistration: The hypotheses, key analyses, and methods were pre-registered on 10/13/2022 (see https://aspredicted.org/DKH_CWM) prior to data collection on 10/24/2022.

Study Three Disclosures. Preregistration: The hypotheses, key analyses, and methods were pre-registered on 11/06/2023 (see https://aspredicted.org/PHZ_XDB) prior to data collection on 11/07/2023.

Learning From Aggregated Opinion

Much of our knowledge about the world is grounded in others' testimony (Rabb et al., 2019). This requires distinctive socio-cognitive mechanisms highly attuned to others' expertise, social standing, and intentions (e.g., Harris et al., 2018; Mercier & Sperber, 2011). For example, sensitivity to the reliability of others' testimony emerges early in childhood (Langenhoff et al., 2022), and both children and adults track who is likely to know what (Lutz & Keil, 2002; Wilkenfeld et al., 2016). However, we are increasingly called upon to make inferences about the world not from the testimony of a few known informants, but from the aggregated opinions of many unknown individuals: For example, we need to learn about the quality of products on online marketplaces through the aggregated reviews of previous customers (Hayes et al., 2021), and we need to learn about the popularity of electoral candidates through polls that aggregate the opinions of thousands (Stoetzer et al., 2022).

Decades of work in psychology (e.g., Kruglanski, 2004), politics (e.g., Huckfeldt et al., 2004), and economics (e.g., Golman et al., 2016) have shed light on the general problem of how people respond to others' opinions. However, work on the more specific problem of how people draw inferences from *aggregated* opinions offers conflicting results. When learning about novel issues from the aggregated opinions of a few correlated sources, people seem to *overweight* others' opinions (Desai et al., 2022; Enke & Zimmermann, 2019; Yousif et al., 2019). When drawing inferences about real-life controversies (e.g., climate change) from disagreeing millions, people seem to *underweight* others' opinions (Oktar & Lombrozo, 2022; see also; Hartman et al., 2022; Iyengar et al., 2019). In other cases, people's inferences seem to accord with Bayesian expectations (Orchinik et al., 2023; Orticio et al., 2022; Stoetzer et al., 2024).

Are people simply bad at consistently drawing accurate inferences from aggregated opinion? Prior work fails to answer this question, as it has not experimentally isolated the inference process conducted formal model comparisons across opinion distributions to

assess whether, when, and why people go wrong. We address this gap by proposing a Bayesian model of inference from aggregated opinion and presenting three behavioral studies that isolate these inferences and compare them to the Bayesian model and to alternative formalizations from epistemology (Easwaran et al., 2016) and economics (Romeijn & Atkinson, 2011).

Formalizing Opinion Inference: Bayes, UPCO, Competence

Imagine learning about a poll of N strangers concerning some issue S—for example, whether the incumbent is the leading candidate in an election. For simplicity, we assume that we only learn whether these N people agree or disagree with S (generating binary opinion samples). We will denote these samples $x_1, x_2, ..., x_N$—with a total of X_1 agreeing with S, and X_0 disagreeing—and call the full vector of samples X. Below, we introduce three models that generate inferences from such data (see Supplementary Materials A for mathematical details; our OSF repository for model implementations).

Bayesian Analysis. Our goal is to use aggregated opinions, X, to infer the probability that the statement S is true, $P(S)$. For simplicity, we will use θ to denote the value of $P(S)$. Bayesian inference combines prior beliefs about θ (denoted $p(\theta)$) with a likelihood function that connects observations of opinions to inferences about θ (denoted $P(X|\theta)$).[1] With these two components, Bayes' rule specifies that the optimal inference (i.e., the posterior, $p(\theta|X)$) is given by:

$$p(\theta|X) = \frac{P(X|\theta)p(\theta)}{\int_0^1 P(X|\theta)p(\theta)d\theta}\,. \tag{1}$$

To make predictions using the Bayesian model, we need to specify how opinions relate to truth and what people already know about S. If people treat informants as providing independent pieces of information and assume that the distribution of opinion corresponds to the probability with which S is true, the likelihood takes on a simple form called the

[1] As θ is a continuous quantity, we use $p(\theta)$ to denote a probability density on θ, and use $P(\cdot)$ for probability mass functions.

binomial likelihood (whereby $P(x_i = 1|\theta) = \theta$ and $P(x_i = 0|\theta) = (1 - \theta)$). These assumptions may be relaxed to capture more complex inferences, as they are frequently violated in reality (see, e.g., Xie & Hayes, 2022, and Supplementary Materials A).

Combining the binomial likelihood with a uniform prior (i.e., absent additional information; $p(\theta) = 1$ for all θ) yields a Beta distribution with the following mean (Laplace, 1774):

$$\mathbb{E}[\theta|X] = \frac{X_1 + 1}{N + 2}\,. \tag{2}$$

In Experiments 1 and 2, 'Bayes' will correspond to this update rule. We describe how we incorporate informative priors to this model in Experiment 3. Intuitively, Bayes is sensitive to the *proportion* of opinions that support S, being a linear function of the proportion $\frac{X_1}{N}$ and the uninformed guess $\frac{1}{2}$: $\frac{X_1+1}{N+2} = \frac{X_1}{N}\frac{N}{N+2} + \frac{1}{2}\frac{2}{N+2}$. Our Bayesian analysis thus justifies the use of a simple strategy—combining priors and proportions—and does not require people to explicitly apply Bayes' rule or engage in complex probabilistic calculations. We return to this point in the General Discussion.

UPCO. There are many alternative rules for drawing inferences from opinions. One such rule, termed UPCO (short for 'updating on the credences of others') is a multiplicative combination of the opinions of every individual in a group (Easwaran et al., 2016). UPCO is a heuristic that mimics Bayesian inference in some cases and respects the results of Condorcet's Jury Theorem (Condorcet, 1785). It is given by:

$$P(S|x_0, ..., x_n) = \frac{\prod_0^n x_i}{\prod_0^n x_i + \prod_0^n (1 - x_i)}\,. \tag{3}$$

Intuitively, UPCO is sensitive to the *absolute margin* of opinion that supports S. For example, the margin in the case where 15 people believe S is true and 5 think it is false is 10. Thus, [15T,5F] leads to a similar prediction to [110T,100F]. Note that in the case of binary opinion, x_i needs to be mapped onto a continuous estimate; we found small deviations from the midpoint to fit data best (e.g., .51; see Supplementary Materials A).

Competence. Another recently proposed formalism aims to estimate the competence of informants (reflected in a reliability rate, r) from the distribution of opinion

itself, and use this estimate to inform predictions about S (Romeijn & Atkinson, 2011):

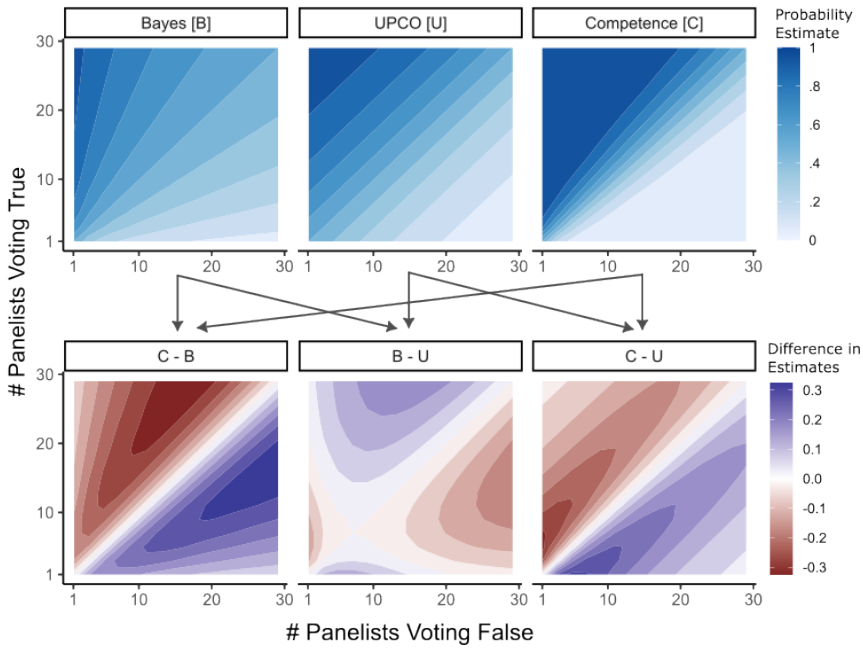
$$P(S|X_1, X_0) = \frac{(N+1)!}{X_1!X_0!} \int_0^{.5} r^{X_1}(1-r)^{X_0} dr\,. \tag{4}$$

We call this algorithm 'Competence.' Intuitively, Competence interpolates between UPCO and Bayes: it is sensitive to both the margin and the proportion, and converges quickly.

These models make convergent predictions for some opinion distributions (e.g., when a panel of people are split on an issue), and divergent predictions on others, with variation across pairs of models in which distributions maximally differentiate predictions (see Figure 1).

Figure 1

Model Predictions Across Opinion Distributions



Note. The top row illustrates model predictions as a function of the number of people in a panel voting True and False on an given issue. The bottom row illustrates pairwise differences in model predictions. Each point corresponds to a particular opinion distribution (e.g., 5 voting true, 15 voting false), and the color indicates the probability that a model assigns to that point (top row) or differences in model predictions for that point (bottom row). Note that the contour plot discretizes continuous model predictions into bins for ease of visual comparison.

Experiment 1: Comparing Models Through Discriminative Points

To empirically compare human inferences with the predictions of these models, we conducted three experiments. In our first experiment, participants made judgments about the truth of 18 unknown statements on the basis of others' opinions in a game-show setting.

Method

Participants

Following a power analysis with pilot data (see Supplemental Materials E), we recruited 133 adults (46 male, 84 female, 3 other, mean age = 34) on Prolific in exchange for monetary compensation ($1 for a 5-minute study). Participation across all studies was restricted to users currently residing in the United States with an approval rating $\geq 98\%$ on at least 100 tasks. Repeat participation within and across studies and pilots was restricted using the Prolific platform. Seventeen participants were excluded from analyses based on pre-registered exclusion criteria (completing the experiment too quickly or failing a comprehension check). This study's design, hypotheses, and key analyses were preregistered; see https://aspredicted.org/L3D_G4Q. All three experiments were approved by the [Author Institution] Institutional Review Board (IRB).

Materials and Procedure

In this study, participants were asked to make a series of 18 truth judgments integrating the conflicting opinions of varying panels of informants in a game-show. Importantly, participants drew inferences about propositions that were masked (e.g., participants read: "For this question, $[N]$ members of the audience were chosen as members of the jury", but were not be told what the question in fact was), and formed their judgments purely on the basis of the distribution of others' opinions.

Participants first read a description of the game-show that specified how the jury was selected randomly from the audience and answered a basic comprehension check (see Supplementary Materials B for details). They then encountered 18 trials with the following

measure (where the variables in square brackets were replaced with trial-specific numbers):

> For this question, [N] members of the audience were chosen as members of 'the jury.' [X_1] of them thought that the statement was 'true,' and [X_0] thought that the statement was 'false.' How likely do you think it is that the correct answer was 'true'?

Participants provided these judgments on a slider scale from 'Completely Impossible' [0] to 'Definitely True' [100].

Three aspects of this paradigm are worth noting for their role in controlling central features of inferences from aggregated opinion. First, the game show setting provides us with an opportunity to designate participants' goals in the task, decoupling inferences from motivations. Second, we are able to specify the distribution, reliability, and interdependence of opinion that participants learn from, decoupling inferences from assumptions about informants. Finally, the use of masked propositions brackets the role of priors, allowing us to focus on the role of evidence from aggregated opinion (in Experiment 3, we re-introduce a role for priors). Varying these features experimentally enables systematic exploration of the large space of inferences from aggregated opinion (see Almaatouq et al., 2022). In our studies, we explored inferences when priors are weak (Experiments 1 & 2) and informative (Experiment 3), informants are independent and randomly sampled, and the inference context is purely focused on accuracy.

We generated 18 trials in this paradigm that maximally discriminated between our three models. To choose the opinion distributions (e.g., 12 agree, 17 disagree) used in these trials, we computed model predictions for every possible opinion distribution (for juries with < 100 members). These models converge on some points, and diverge on others (see Figure 1). For example, when a large jury overwhelmingly votes 'False,' all three models predict low probability of truth. Other cases (e.g., small, split juries) are more likely to induce differing predictions. We identified the opinion distribution that led to the most divergent predictions for each pair of models, which led to three maximally discriminative

points (i.e., opinion distributions).

We then used these three key points to generate 15 other points by sampling 5 additional opinion distributions per model with the same predicted probabilities. For instance, if the point [1,11] most clearly distinguished UPCO from Bayes, and UPCO predicted .40 for that point, we would sample 5 other points in opinion space (e.g., [3,13]) for which UPCO also predicted .40. Half of the trials were inverted to have more judges voting false than true and reverse scored in later analyses. After providing these 18 judgments, participants answered demographic questions and were debriefed.

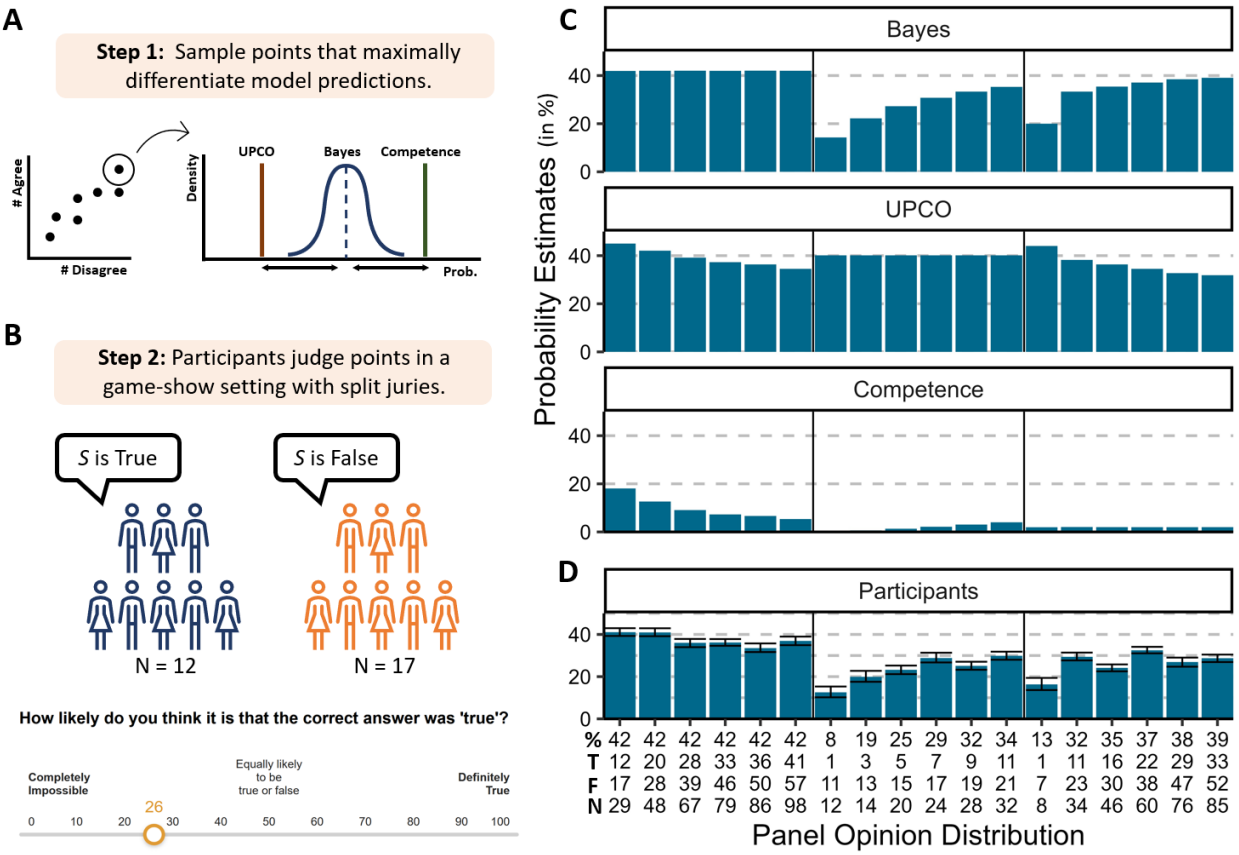
Results

We had three main predictions. Based on past work showing that people often approximate Bayesian inferences (e.g., Griffiths & Tenenbaum, 2006), we first predicted that the Bayesian model's estimates would be significantly correlated with mean participant judgments across our 18 key points. Second, we predicted that the correlation between model and human judgments would be strongest for the Bayesian model. Finally, beyond correlations at the population level, we also predicted that the Bayesian model would capture individual judgments better than the heuristic models. That is, we expected that the mean correlation across participants' judgments and the predictions of the three models would be highest for Bayes. Our data support all three predictions (see Figure 2).

In support of our first prediction, the Bayesian model's predictions were highly correlated with participant judgments, $r = .92, p < .001$. The heuristic models were more weakly correlated (for UPCO: $r = -.08, p = .75$; for Competence, $r = .79, p < .001$). Partially in line with our second prediction, the difference in correlation coefficients was statistically significant for Bayes and UPCO ($t = 4.51, p < .001$) but not for Bayes and Competence ($t = 1.35, p = .18$).

Figure 2

Comparison of Model Predictions and Mean Participant Judgments in Experiment 1.

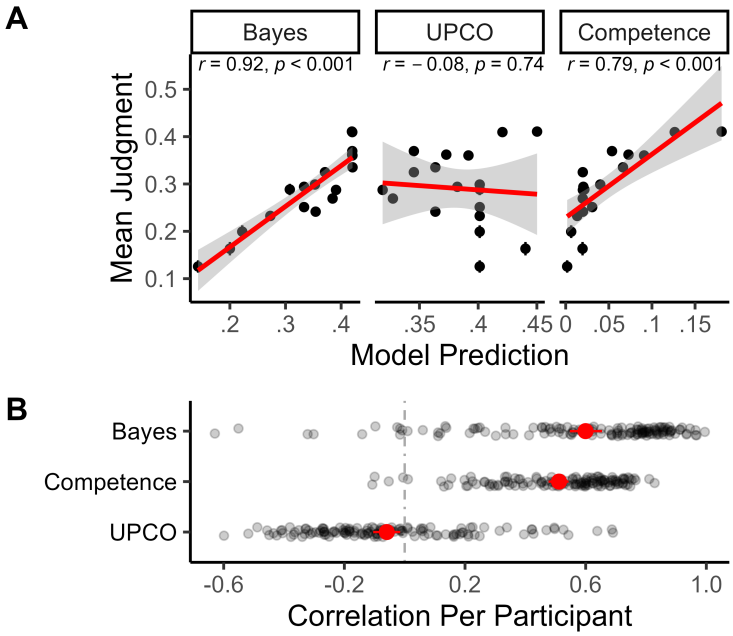


Note. **A** depicts the stimulus selection procedure. For each pair of models, we calculated the opinion distribution (e.g., 12 agree, 17 disagree) that led to the most divergent predictions, and used these 3 key points to sample the remaining 15 points. **B** illustrates our experimental task. Participants were told that randomly chosen juries of 2 to 100 members evaluated trivia statements, and reached diverging conclusions—with opinion distributions matching the stimuli described in **A**—and made inferences about the truth of these statements. These points ranged from juries of 8 to 98 members. **C** shows model predictions, where Bayes corresponds to the binomial posterior mean. **D** is a plot of mean participant judgments across points (horizontal axis shows opinion distributions, with the top row indicating the % voting True (%), the second row indicating the # voting True (T), the third row indicating the # voting False (F), and the last row showing the total # of judges (N); error bars show bootstrapped 95% confidence intervals).

We found stronger evidence at the level of individual participants: Mean correlations between each individual participant's judgments and model predictions were significantly higher for Bayes than UPCO ($t(132) = 13.68, p < .001$), and for Bayes than Competence ($t(132) = 4.19, p < .001$), as revealed by paired t-tests of the correlation coefficients (see Figure 3). Comparing absolute differences between predictions and judgments supports similar conclusions (see Supplementary Materials C).

Figure 3

Correlations between Human Judgments and Model Predictions from Experiment 1.



Note. **A** shows correlations across mean participant estimates and model predictions (shaded area indicates 95% confidence interval). **B** plots the correlations between *each participant's judgments* and model predictions (red point shows the mean).

We can also ask which model best predicts the judgments of each individual participant. Conducting this exploratory analyses revealed that $\sim 75\%$ of participants are best captured by the Bayesian model in terms of correlations, and $\sim 80\%$ in terms of deviations. The rest of participants are split between UPCO and Competence, with Competence performing better for correlations than deviations (see Supplementary

Materials, Figure 2).

Discussion

The results of Experiment 1 suggest that Bayes is the best predictor of people's judgments for stimuli that most strongly discriminate between our models. However, there are important limitations to this analysis.

First, though we sampled informative points, we sampled just a few (our 18 points comprise $\sim 1.4\%$ of the space of opinion distributions for juries up to 50 members; see Supplementary Figure 3). A consequence of this sparse sampling is that it is hard to know the extent to which our results generalize—while Bayes performed well for our stimuli, it could fail to accurately characterize inferences in cases of moderate agreement, for instance.

Second, our analysis compared the three models against each other. Beyond finding the best model, however, our goal is to best characterize people's inferences, and it is possible that a combination of our models could outperform individual predictions. To conduct more sophisticated model comparisons, we need greater statistical power.

Experiment 2: Densely Sampling Opinion Space

In Experiment 2, we extended our prior findings by densely sampling opinion distributions for juries of up to 60 members. This allowed us to investigate model performance across the full space of opinion distributions and provided enough power to conduct fine-grained comparisons of our models with ensembles.

Method

Participants

Following a power analysis with pilot data (see Supplemental Materials E), we recruited 458 adults (204 male, 242 female, 12 other, mean age = 38) as in Experiment 1 for a slightly longer study ($1.20 for a 6-minute study). 108 participants were excluded from analyses based on pre-registered criteria (33 for completing the experiment too quickly, 72 for failing an attention check—though including these participants does not

change conclusions). This study's design, hypotheses, and analyses were preregistered; see https://aspredicted.org/DKH_CWM.

Materials and Procedure

Participants completed the same task as in Experiment 1; providing inferences about statements purely on the basis of the opinions of a jury, as participants were not shown the statements themselves.

Instead of rating 18 key points, however, they were randomly assigned 20 of 225 possible stimuli. Each of the 20 stimuli corresponded to a particular arrangement of the jury in the game show, from a split two-person jury to a split 58-person jury (i.e., 29 think it is false, 29 think it is true). The 225 stimuli were densely sampled to cover a quarter of all points from [1,1] to [29,29] in a square grid (alternatively skipping a point on each jury, e.g., the first four points are ([1,1], [1,3], [3,1], [3,3]; see Supplementary Figure 4).

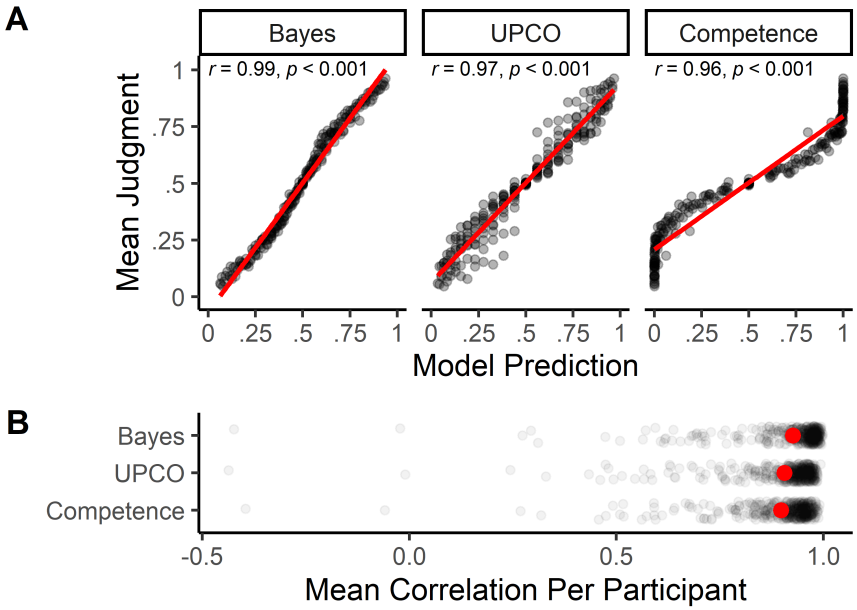
Results

Our data in this study show even stronger concordance between participants' judgments and the predictions of the three models (see Figure 4). That the correlations in this study are much higher should not be surprising: Whereas we picked 18 highly discriminative points in Experiment 1, here we are also including 'easier' points on which models agree. This reduces the power to differentiate across our models from correlations.

Importantly, this richer data set enables us to investigate the relationship between participants' judgments and the structure of model predictions in much finer detail than we can observe through correlations. We can compare—across all 225 opinion distributions—the relationship between model predictions and mean participant judgments. This constitutes a much more stringent test of our models: Whereas correlations ask whether models capture the pattern of inferences, comparing model predictions in the original space allows us to ask whether models provide accurate quantitative predictions of inference (see Figure 5).

Figure 4

Correlations between Human Judgments and Model Predictions from Experiment 2.

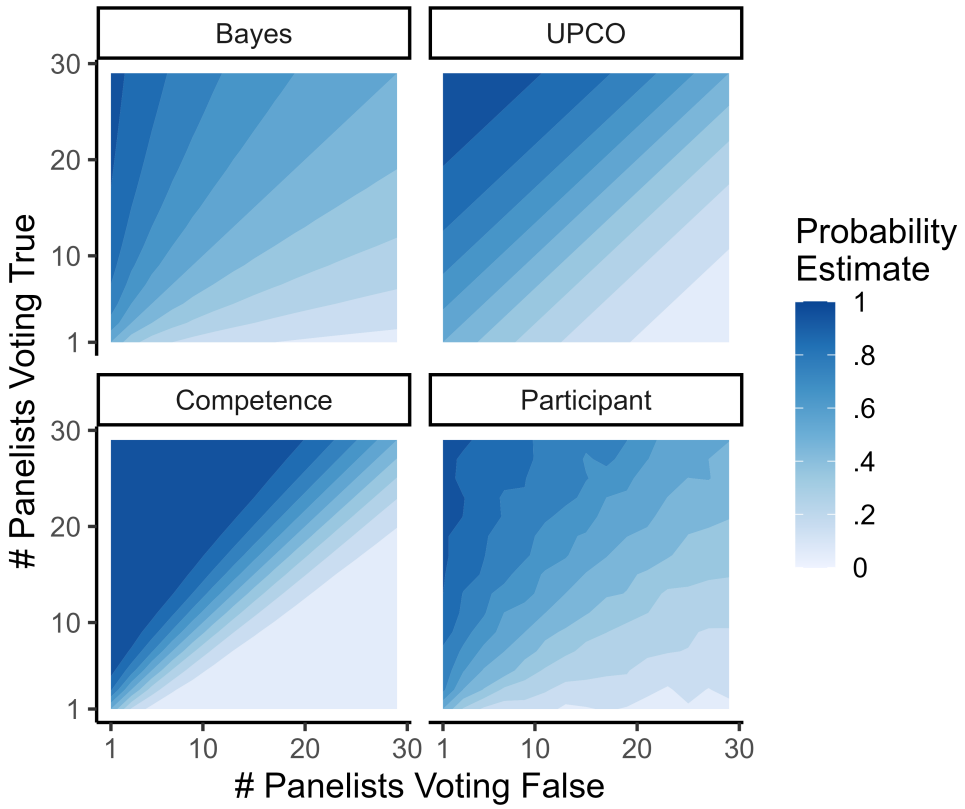


Note. **A** shows correlations across mean participant estimates and model predictions (shaded area indicates 95% confidence interval). **B** plots the correlations between *each participant's judgments* and model predictions (red point shows the mean).

To compare the relative performance of each model, we therefore adopted a more stringent predictive test than comparing correlations. We pre-registered analyses comparing the Akaike Information Criterion (AIC) scores of three separate linear regressions—each regression predicting mean participant judgments from model predictions across the 225 points. Note that AIC is an estimator of prediction error that penalizes flexibility (i.e., models with more parameters but the same predictive performance have worse AIC scores). This penalization is key, as we are not just interested in which individual model provides the best predictions, but are also interested in comparing the performance of model ensembles, which combine multiple models to generate predictions. The penalization explains why ensembles, which have more parameters than individual models, would not necessarily have better AIC scores, even if they made more accurate in-sample predictions.

Figure 5

Human Judgments and Model Predictions Across Opinion Distributions in Experiment 2.



Note. This contour plot shows model predictions and participant judgments across all opinion distributions for juries up to 60 members. The Participant plot shows a linear interpolation of mean judgments on the square grid. Visually, Bayes corresponds most closely to human judgments.

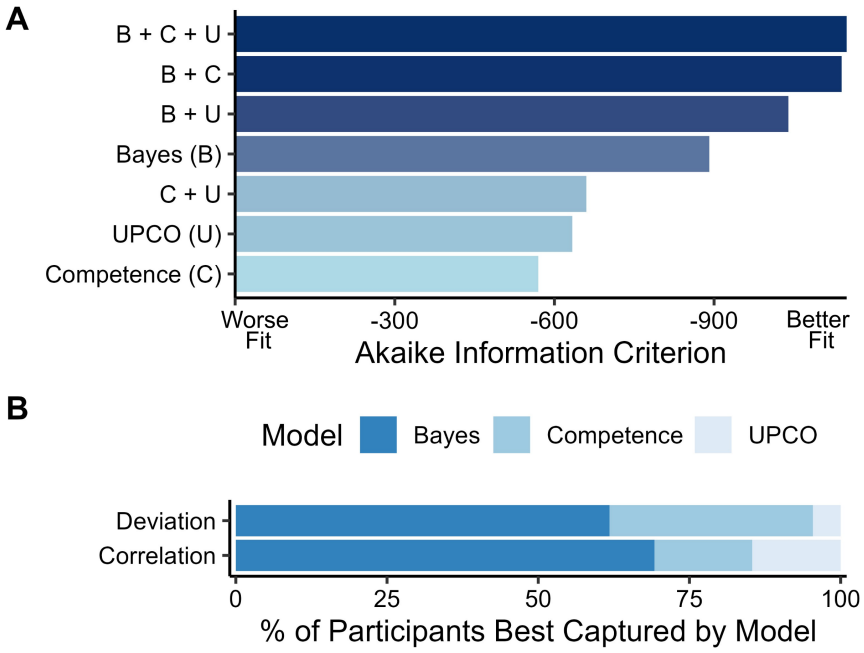
Our first hypothesis was that the Bayesian model would be the best predictor of judgments (i.e., have the lowest AIC score), and our analyses support this: Bayes once again best predicted mean judgments across opinion distributions (see Figure 6 for comparisons of AIC scores, Supplementary Figure 5 for a regression table).

To investigate whether our models were better at explaining judgments cumulatively than individually, we compared the AIC scores of all ensembles (i.e., linear combinations of models) with the scores of individual models. We were agnostic about whether the additional models would yield improvements in performance over the Bayesian model. This

analysis revealed that ensembles outperform individual models—with the addition of Competence conferring a larger boost in prediction than UPCO (see Figure 6, subplot **A**).

Figure 6

Model and Ensemble Comparisons from Experiment 2.



Note. **A** shows AIC values for regressions predicting judgments from model predictions and linear combinations of models (since smaller AIC values indicate better fit, the horizontal axis has been flipped). Bayes is the best individual predictor, but is outperformed by ensembles. **B** shows which model best captures inferences at the level of individual participants.

One explanation for why ensembles outperform individual models is that there may be heterogeneity across participants' inference strategies. While Bayes best predicts judgments *overall*, some participants' behavior could be better captured by heuristic models. To investigate this possibility, we conducted an exploratory participant-level analysis where we calculated the proportion of participants whose responses were best captured by the three models. Using the same participant-level correlation and deviation analyses from Experiment 1, we found that UPCO and Competence were better predictors of a quarter of our participants' judgments (see Figure 6, subplot **B**). These analyses thus

reveal substantial heterogeneity in people's inferences.

Discussion

The densely sampled stimuli in Experiment 2 extend our prior results by showing that Bayes not only performs well for points that differentiate between our models, but also generalizes well to the broader space of opinion distributions. Moreover, our analyses reveal that people use a diversity of inference strategies.

Experiment 3: Integrating Prior Beliefs with Opinions

In our final study, we conducted a more stringent test of the Bayesian model by investigating (a) whether it continues to be the best predictor when there is variation in priors, and (b) how it performs when participants evaluate real claims. To address these questions, we elicited participants' prior beliefs about 18 trivia statements, provided them with aggregated opinions, and investigated their updated judgments. The Bayesian model predicts people's judgments should vary with the prior probabilities assigned to statements (see Supplementary Materials G).

Method

Participants

Following a power analysis with pilot data (see Supplemental Materials E), we recruited 295 adults (163 male, 126 female, 6 other, mean age = 42) as in Experiment 2 for a slightly longer study ($2.00 for a 10-minute study). Sixty participants were excluded from analyses based on pre-registered criteria (52 for completing the experiment in an unrealistic amount of time, 8 for failing an attention check). This study's design, hypotheses, and analyses were preregistered; see https://aspredicted.org/PHZ_XDB.

Materials and Procedure

Participants completed a task similar to Experiment 1. Instead of rating masked statements, however, each participant was assigned a random pairing of 18 actual trivia statements with the panel opinion distributions used in Experiment 1. All participants

thus evaluated all 18 trivia items and panel distributions, but they received differing subsets of all possible trivia and panel combinations ($18 \times 18 = 324$ possible combinations). The order of presentation was randomized both within the first set of prior judgments and the later set of updated judgments.

To ensure that our stimuli spanned multiple domains and confidence levels, we sampled trivia from the General Knowledge Norms dataset (GKN; S. K. Tauber et al., 2013), which provides updated information on the original set of 300 general-information questions from Nelson and Narens (1980). These questions span a wide variety of domains, including history, sports, art, geography, literature, and entertainment, and the GKN includes a variety of measures, such as people's confidence in their beliefs about each item. We uniformly sampled 18 questions that evenly spanned the range of confidence to use in our study. This resulted in a diverse set of statements, some of which participants would have very strong priors on (and hence would be able to easily identify as true or false), and some of which participants would lack informative priors on (and hence would not be confident). Using this diverse set increases the odds that the predictive comparisons of our models would generalize across issues on which people have weak and strong priors.

To generate true and false items, we used data in the General Knowledge Norms to rank questions by prior confidence. Starting with the most well-known item, we generated false answers to every other statement in the ranking, and the rest were paired with correct answers, resulting in 9 true and 9 false items. Some of the false answers were generated to be obviously false (e.g., that venison is the name of ox meat—when it is the name of deer meat), whereas others were false in more subtle ways (e.g., that Bismarck is the name of a German battleship sunk in World War 1—when it sunk in World War 2). This variance prevented falsity from being confounded with the strength of prior knowledge: if all statements were obviously false, even otherwise unknown trivia items would become obvious (e.g., if the warship item read 'Bismarck is the name of the Saudi Arabian battleship sunk in World War 1,' it would have become a trivially easy item).

The final set covered statements that are both well known (e.g., 'POPEYE IS THE NAME OF THE COMIC STRIP CHARACTER WHO EATS SPINACH TO INCREASE HIS STRENGTH'), partially known (e.g., 'CANBERRA IS THE CAPITAL AUSTRALIA'), and relatively unknown ('DITHERS IS THE LAST NAME OF DAGWOOD'S MAID IN THE COMIC STRIP "BLONDIE"'; see Supplementary Materials F for a list of all statements).

Participants first provided their prior truth judgments for the statements ("Based on your prior knowledge, how likely do you think it is that this statement is true?") on a truth slider scale from 'Definitely False' [0] to 'Definitely True' [100]. They then provided their confidence in their prior judgments ("(...) How confident are you in this response?") from 'Not at all' [0] to 'Extremely' [10]. They then learned about the panel's opinion distribution and provided their final judgments using the same truth scale:

> In this round of Know-off, the following statement was shown to the contestant: [X]. For this question, [A] members of the audience were chosen as members of the jury. [B] of them thought that the statement was true. [C] of them thought that the statement was false. How likely do you think it is that the correct answer was true?

The 18 true/false combinations for the jury (corresponding to [A], [B], and [C] above) were taken from Experiment 1. The order in which statements were presented was randomized both within each experiment (across the prior and posterior measures), and across participants. Participants then provided demographic information.

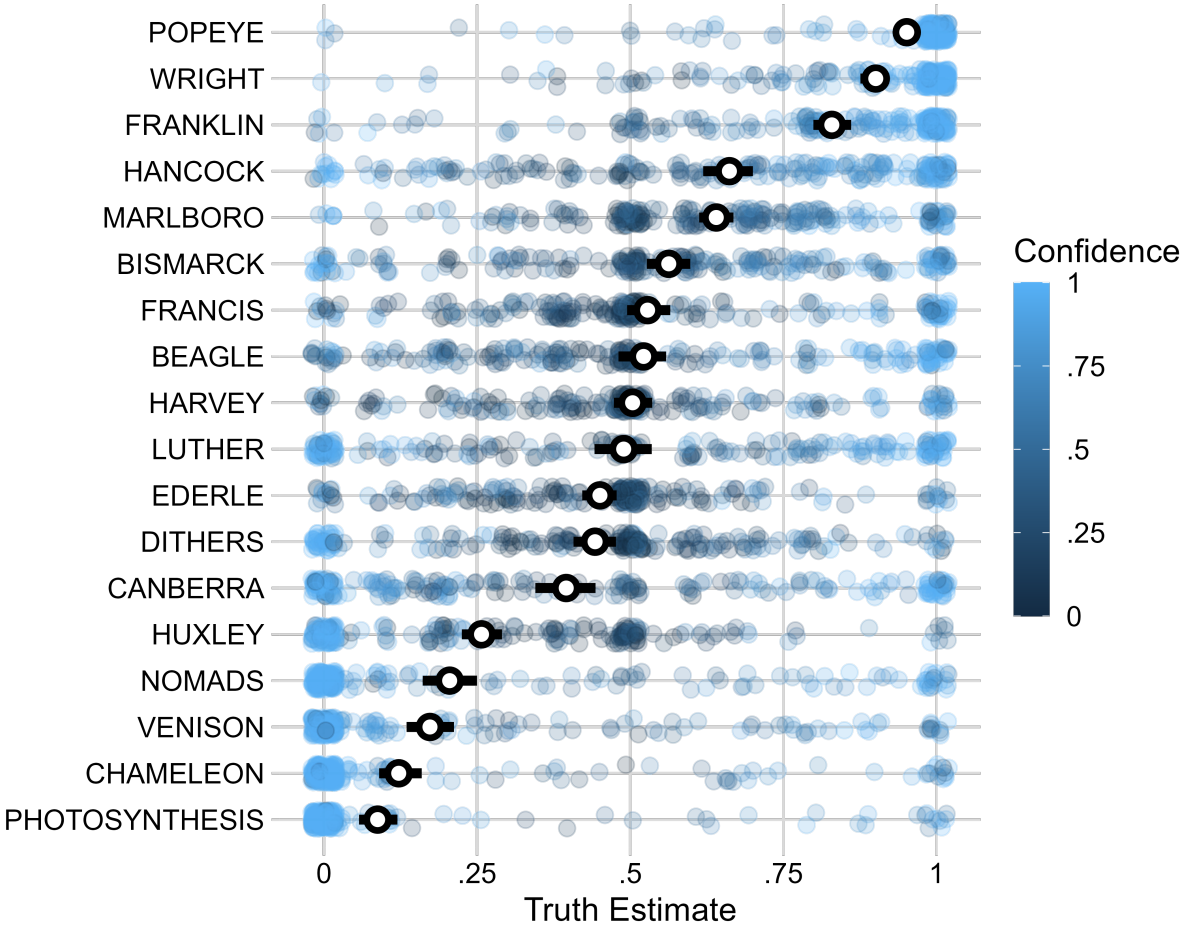
Note that we need to infer participants' prior distributions from their judgments to fit the Bayesian model—we describe our pre-registered method for fitting these distributions in Supplementary Materials G, and also provide details on the fitting of all three models. Importantly, we explicitly emphasized that the jury members deliberated independently in our instructions and comprehension checks for this study (see Supplementary Materials F; Experimental Preamble).

Results

We first verified that the trivia we sampled spanned a range of prior beliefs, from statements generally known to be true or false to highly uncertain statements. The distribution of judgments indicates that we successfully sampled a diverse set of items (see Figure 7).

Figure 7

Prior Truth Judgments for Trivia Statements Used in Experiment 3.



Note. Participants' truth judgments for the chosen trivia statements shows that our stimuli span truth inferences (from definitely false to definitely true) as well as confidence levels (from not at all confident to extremely confident). Points were slightly jittered to display overlapping data; black points show mean truth inferences with bootstrapped 95% confidence intervals. Each row plots responses to a particular item; Supplementary Materials F contains further details on these items.

We then investigated which models best predicted participants' responses by conducting the same AIC-based comparisons of individual models and ensembles used in Experiment 2. We opted to conduct our analysis across all judgments because most participants received unique opinion/statement combinations, which reduces the informativeness of mean comparisons. As with our prior studies, we hypothesized that the Bayesian model would best predict judgments, but were agnostic with regards to whether ensembles would outperform individual models.

Our results support this hypothesis, as the Bayesian model had the best individual predictive performance. Moreover, we obtained largely the same ordering of model performance as Experiment 2, with ensembles boosting the performance of the Bayesian model (see Figure 8 for comparisons of AIC scores and Supplementary Figure 11 for a plot of all judgments).

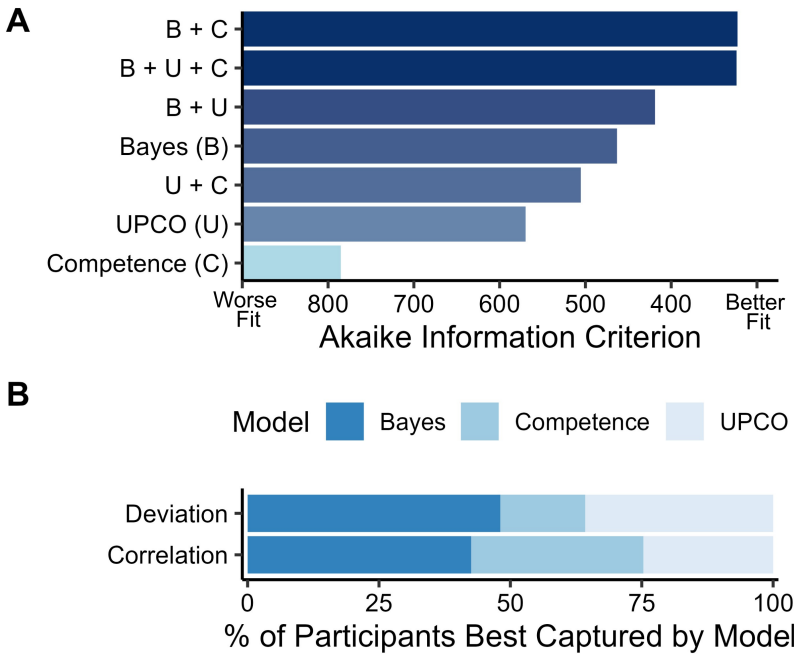
Note that the AIC scores in Figure 8 are estimates from a fixed-effects analysis with the three models as predictors. Importantly, including random intercepts for items and participants in a more comprehensive mixed-effects model does not change the predictive ordering of models or any conclusions (see Supplementary Figures 12 and 13 for both regression tables and a discussion of our pre-registered analysis plan).

Beyond replicating comparisons across our three models, these data allowed us to investigate how the Bayesian model would compare to purely proportion- and prior-based heuristics. In our previous studies, the Bayesian model's predictions were difficult to disentangle from proportion-based heuristics, due to the absence of prior information. We hypothesized that participants' inferences would be better characterized by the Bayesian model than heuristic strategies that do not integrate prior knowledge with the data.

Our results support this prediction. The proportion heuristic (AIC $\approx$ 1722) and the prior heuristic (AIC $\approx$ 922) under-performed relative not just to the Bayesian model (AIC $\approx$ 410) but also to UPCO (AIC $\approx$ 497) and Competence (AIC $\approx$ 507) across the same regression-based comparisons (lower AIC values indicate better predictive performance).

Figure 8

Model and Ensemble Comparisons from Experiment 3.



Note. **A** shows AIC values for regressions predicting judgments from model predictions and linear combinations of models (since smaller AIC values indicate better fit, the horizontal axis has been flipped). Bayes is the best individual predictor, but is outperformed by ensembles. **B** shows which model best captures inferences at the level of individual participants.

We also hypothesized that the Bayesian model would best characterize individual participants' inferences. Our data support this hypothesis as well, with Bayes best capturing 43% of participants' estimates in terms of correlations, and 48% in terms of deviations. In particular, Bayes outperforms the other models in cases of large updates (see Supplementary Figure 14). Yet there is a decrease in the proportion of participants best characterized by Bayes from Experiment 2, which suggests that more participants utilize heuristics for the more complex judgments in Experiment 3.

Discussion

This experiment shows that the Bayesian model best predicts inferences even when participants must integrate prior beliefs with aggregated opinion. Once again, the best

performing model is an ensemble enriched by other models due to heterogeneity across participants' strategies.

General Discussion

Past work has emphasized that we learn much of what we know from others—often through local, social interactions (Harris et al., 2018), and increasingly through exposure to the aggregated views of many others (through likes, polls, and reviews; Kozinets et al., 2010). Recent work has investigated learning from aggregated opinion and reached diverging conclusions—people seem radically insensitive to aggregated opinion in some cases, and overly-sensitive in others. To reconcile these findings, we proposed a Bayesian model of learning from aggregate opinion and conducted three experiments to compare the predictions of this model with human judgments. Across all experiments, we found the strongest concordance between participants' judgments and the predictions of the Bayesian model over two tested alternatives, though many participants utilized alternative strategies, and models performed best in predicting aggregate judgments. Importantly, the predictive success of the Bayesian model does not necessarily indicate that people are performing Bayesian inference, as this model results in a simple strategy for combining priors and proportions. The Bayesian analysis does explain, however, why people might follow this strategy: it corresponds to a reasonable statistical inference (for further discussion, see Marr, 1982; S. Tauber et al., 2017).

These results raise important questions. First, how can people's inferences be best characterized by a Bayesian model, when prior work has found that people over-or under-weight aggregated opinion? Our formal analysis highlights three factors as potential explanations for this discrepancy: prior beliefs about controversies, the reliability of disagreeing parties, and dependency structures across informants. For instance, strong prior beliefs about controversies (e.g., climate change) can lead people to persist in their views amid disagreement, and hence appear insensitive to aggregated opinion (Orchinik et al., 2023). On the other hand, prior belief in the independence of news sources can lead

people to draw strong conclusions from aggregated opinion (Desai et al., 2022). Relatedly, the complexity of the task may matter: We found that more participants relied on alternative heuristics in Experiment 3, potentially due to the increased demands of integrating prior information with opinion data. Whether people are well-calibrated can therefore only be adjudicated on a case-by-case basis, through modeling and measurement of factors such as informant reliability (Landrum et al., 2015) and dependency (Whalen et al., 2018). Relatedly, future work could explore departures from the assumption that the distribution of opinion directly maps on to the probability of truth, and cross-cultural and contextual variance in inferences.

Second, do our results generalize to real-world opinion aggregation problems? The answer depends on the extent to which a particular problem matches our task. For instance, the results of Experiment 3 suggest that the Bayesian model will perform well in predicting how people utilize aggregate opinions in actual trivia contests, as our task is isomorphic to the inference problem posed in this case (e.g., the 'Ask the Audience' lifeline in *Who Wants to Be a Millionaire?*). When it comes to real-life controversies such as abortion, however, generalizability is an open question, given differing assumptions about the reliability (Hartman et al., 2022) and dependence (Judd & Park, 1988) of disagreeing others, as well as variation in the aggregation and presentation of opinions (Fisher et al., 2018). Moreover, the psychology of controversy is clearly shaped by factors beyond the epistemic—such as inferences about an issue's subjectivity (Oktar & Lombrozo, 2022). Our experimental paradigm can be adapted to investigate the influence of many such factors.

With widening rifts of opinion corroding the foundations of many democracies across the globe, elucidating when and why we learn from controversy (i.e., polarized aggregated opinion)—or fail to do so—is becoming an increasingly important goal. Our work advances this aim with both methodology and theory. Methodologically, our approach demonstrates how formal theories can be combined with experimental data to systematically investigate questions about mass opinion. Theoretically, we identify key

factors underlying inferences from aggregated opinion, and show that in simple cases many people can draw Bayesian inferences across opinion distributions. This suggests that persistence amid controversy is grounded in rich inferences about reliability, dependence, and beyond, rather than an inability to draw inferences from aggregated opinion.

References

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2022). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, 1–55.

Condorcet, J. A. N. d. C. (1785). *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix*. Imprimerie Royale Paris.

Desai, S. C., Xie, B., & Hayes, B. K. (2022). Getting to the source of the illusion of consensus. *Cognition*, *223*, 105023.

Easwaran, K., Fenton-Glynn, L., Hitchcock, C., & Velasco, J. D. (2016). Updating on the Credences of Others: Disagreement, Agreement, and Synergy. *Philosophers' Imprint*, *16*, 1–39.

Enke, B., & Zimmermann, F. (2019). Correlation neglect in belief formation. *The Review of Economic Studies*, *86*(1), 313–332.

Fisher, M., Newman, G. E., & Dhar, R. (2018). Seeing stars: How the binary bias distorts the interpretation of customer ratings. *Journal of Consumer Research*, *45*(3), 471–489.

Golman, R., Loewenstein, G., Moene, K. O., & Zarri, L. (2016). The preference for belief consonance. *Journal of Economic Perspectives*, *30*(3), 165–88.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, *69*(1), 251–273.

Hartman, R., Hester, N., & Gray, K. (2022). People see political opponents as more stupid than evil. *Personality and Social Psychology Bulletin*, 01461672221089451.

Hayes, B. K., Wisken, A., & Cruz, N. (2021). Explaining the popularity bias in online consumer choice. *Journal of Experimental Psychology: General*, *150*(10), 2185–2191.

Huckfeldt, R., Johnson, P. E., & Sprague, J. (2004). *Political Disagreement: The Survival of Diverse Opinions within Communication Networks*. Cambridge University Press.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, *22*(1), 129–146.

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, *54*(5), 778–788.

Kozinets, R. V., De Valck, K., Wojnicki, A. C., & Wilner, S. J. (2010). Networked narratives: Understanding word-of-mouth marketing in online communities. *Journal of Marketing*, *74*(2), 71–89.

Kruglanski, A. W. (2004). *The psychology of closed mindedness*. Psychology Press.

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109–111.

Langenhoff, A. F., Engelmann, J. M., & Srinivasan, M. (2022). Children's developing ability to adjust their beliefs reasonably in light of disagreement. *Child Development*.

Laplace, P. S. (1774). Mémoire sur la probabilité de causes par les évenements. *Mémoire de l'académie royale des sciences*.

Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, *73*(4), 1073–1084.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt; Co., Inc.

Mercier, H., & Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 338–368.

Oktar, K., & Lombrozo, T. (2022). Mechanisms of belief persistence in the face of societal disagreement. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*.

Rabb, N., Fernbach, P. M., & Sloman, S. A. (2019). Individual Representation in a Community of Knowledge. *Trends in Cognitive Sciences*, *23*(10), 891–902.

Romeijn, J.-W., & Atkinson, D. (2011). Learning juror competence: A generalized condorcet jury theorem. *Politics, Philosophy & Economics*, *10*(3), 237–262.

Stoetzer, L. F., Leemann, L., & Traunmueller, R. (2022). Learning from polls during electoral campaigns. *Political Behavior*, 1–22.

Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the nelson and narens (1980) norms. *Behavior Research Methods*, *45*, 1115–1143.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410.

Whalen, A., Griffiths, T. L., & Buchsbaum, D. (2018). Sensitivity to shared information in social learning. *Cognitive Science*, *42*(1), 168–187.

Wilkenfeld, D. A., Plunkett, D., & Lombrozo, T. (2016). Depth and deference: When and why we attribute understanding. *Philosophical Studies*, *173*, 373–393.

Xie, B., & Hayes, B. (2022). Sensitivity to evidential dependencies in judgments under uncertainty. *Cognitive Science*, *46*(5), e13144.

Yousif, S. R., Aboody, R., & Keil, F. C. (2019). The illusion of consensus: A failure to distinguish between true and false consensus. *Psychological Science*, *30*(8), 1195–1204.