Journal Title
XX(X):1–14
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Mitigating Biases in Collective Decision-Making: Enhancing Performance in the Face of Fake News

Axel Abels[1,2,3] and Elias Fernandez Domingos[1,2,3] and Ann Nowé[2,3] and Tom Lenaerts[1,2,3,4]

Abstract

Individual and societal biases significantly impair the efficacy of human advisors, leading to judgment errors that disproportionately affect protected groups. This paper investigates how such biases influence the ability of individuals to discern false from true news headlines, particularly those involving sensitive characteristics. We collected a comprehensive dataset based on participants' responses to these headlines to examine the impact of biases on their judgments. Our analysis identifies consistent patterns of individual bias and demonstrates that demographic factors, headline categories, and the manner in which information is presented significantly influence human judgment errors. Although the "wisdom of the crowd" concept suggests that collective diversity might counteract individual biases, our findings indicate that prevalent societal beliefs often dominate, leading to collective errors rather than wisdom. To counteract these effects, we propose the adoption of adaptive aggregation algorithms over simplistic aggregation strategies. Using our dataset as a benchmark, we validate the effectiveness of this approach. Our results not only show enhanced performance but also highlight the interactions between the emergence of collective intelligence and the mitigation of participants' biases.

Keywords

collective intelligence, wisdom of crowds, judgment aggregation, bias, fake news, machine learning

Significance Statement: Cognitive biases like confirmation bias impair judgment in critical areas such as healthcare, law enforcement, and online moderation. For instance, these biases can exacerbate the spread of fake news, particularly about sensitive groups, by favoring information that confirms pre-existing beliefs. While the "wisdom of the crowd" promises to harness diverse opinions for superior outcomes, we show it can fail when widely held beliefs permeate into aggregates. Importantly, we demonstrate how adaptive machine learning algorithms can counteract these biases. By adjusting through a limited set of user interactions, our approach significantly enhances the quality and reliability of collective decision-making. These findings suggest a promising path toward harnessing the true potential of crowds, thereby improving fairness and accuracy in crucial decision-making processes.

Cognitive biases are systematic errors in judgment and decision-making resulting from cognitive limitations, individual preferences, and inappropriate heuristics (Tversky and Kahneman 1974). When human groups deliberate, individuals tend to transmit these biases to others, which may lead the group as a whole to make sub-optimal choices (Janis 2008). In settings wherein group decisions play a pivotal role — from medical diagnostics to crowdsourcing — an understanding of these biases, how they affect the integrity and efficiency of collective decisions, and how they can be countered is paramount.

Existing literature underscores the profound influence of stereotypes, biases, and decision-making across multiple facets of life, from choices in recruitment (Rooth 2010) to judiciary decisions (Lin et al. 2020; Dressel and Farid 2018). Tools and studies like the Stereotype Content Model (Fiske et al. 2002; Cuddy et al. 2009) and Implicit Association Tests (Greenwald et al. 1998; Greenwald and Krieger 2006) offer crucial insights into measures of biases. However, there remains a considerable scope for further exploration, particularly in understanding how the emergence of collective intelligence interacts with efforts to mitigate bias.

Pooling human expertise can help mitigate individual errors stemming from biases such as overconfidence (Harvey 1997), biases with temporal effects (Baddeley and Hitch 1993; Jones and Sugden 2001), or illusory correlations (Pohl 2004). Works such as those of Condorcet (1785), Rawls (1971), and Surowiecki (2005) provide foundational insights into how collective decision-making can enhance outcomes compared to individual efforts. Building on these concepts, modern algorithms (e.g., (Luo et al. 2018; Abels et al. 2023a,b)) address the complexities of aggregating possibly

[1]Machine Learning Group, Université Libre de Bruxelles, Boulevard du Triomphe, CP 212, Brussels, 1050, Belgium
[2]AI Lab, Vrije Universiteit Brussel, Pleinlaan 2, Brussels, 1050, Belgium
[3]FARI Institute, Université Libre de Bruxelles - Vrije Universiteit Brussel, Brussels, 1000, Belgium
[4]Center for Human-Compatible AI, UC Berkeley, 2121 Berkeley Way, Berkeley, CA 94720, USA

Corresponding author:
Axel Abels, Machine Learning Group, Université Libre de Bruxelles, Boulevard du Triomphe, CP 212, Brussels, 1050, Belgium
Email: axel.abels@ulb.be

dissenting opinions by dynamically adapting to varying levels of expertise. While those works provide theoretical performance guarantees, their empirical validation largely relies on simulations with synthetic experts (Pang et al. 2018; Abels et al. 2020, 2023a,b). Consequently, it remains unclear whether the underlying assumptions apply to human experts, raising questions about their effectiveness in real-world applications.

In addition to investigating how biases influence individual judgment, we therefore leverage the collected data to fill this gap in experimental validation. We specifically assess whether adaptive aggregation algorithms (EXP4 (Auer et al. 2002), MetaCMAB (Abels et al. 2023a), and ExpertiseTree (Abels et al. 2023b)) enhance collective intelligence and if their impact on biases differs from that of non-adaptive strategies such as majority votes.

Our experimental design[*], inspired by crowdsourced fact checking (Sethi 2017; Wei et al. 2022), involves human participants tasked with discerning the authenticity of news headlines, particularly those that involve sensitive groups (see Figure 1). In the experiment, participants were presented with an equal number of genuine and modified headlines, obtained by altering the involved sensitive group. Without being informed of the ratio of authentic to altered headlines, participants were asked to rate the likelihood of each headline being genuine. By assessing whether participants' error rates are affected by demographic factors, by the sensitive group, or by how information is presented we can evaluate the prevalence of various biases.

To evaluate the effectiveness of (non-)adaptive aggregation strategies, responses were used in simulated collective decision-making tasks (see Methods). For each simulation, we sampled uniformly $N \in \{2, 4, 6, ..., 36\}$ participants who answered the same set of headlines. Their answers to each headline were iteratively fed into the studied aggregation algorithms. These algorithms then aggregated the responses from this subset to reach a collective decision. After each decision, the true nature of the headlines—whether genuine or altered—was revealed to the algorithms. This feedback loop (see Figure 1) allows the algorithms to adapt their aggregation policy to improve performance on subsequent headlines[†].

To summarize, our contributions are

- A dataset containing participant responses to a collection of news headlines involving sensitive characteristics, providing a rich source for analyzing human biases in decision-making.
- A comprehensive exploration of these responses, focusing on potential biases and the presence of stereotyping.
- An analysis of the performance of collective decision-making algorithms, with an emphasis on the interaction between bias mitigation and the emergence of collective intelligence.

The following section provides a detailed examination of our results. Overall, they not only contribute to the understanding of bias in human decision-making but also underscore the valuable role that machine learning can play in improving collective decision-making processes.

Results

Dissecting Participant Biases

To explore the extent to which collective decision-making systems can reduce biases, we begin by analyzing the biases exhibited by participants.

Demographic Differences in Performance We first investigate whether the participants' demographics correlate with differences in performance across headline categories and find the following, as illustrated by Figure 2.

There is a significant difference in accuracy between men and women, especially when responding to gender-related headlines. Men exhibited a lower average accuracy (0.524) compared to women (0.554), with a statistically significant difference ($\beta = 0.0303, p = 0.01, ci = [-0.053, -0.007]$). Participants' performance also varied significantly with age, particularly when interpreting headlines about different age groups. Individuals younger than 35 years were more accurate in identifying altered headlines about age (0.535 accuracy) compared to those older than 35 years (0.511 accuracy), with a significant difference ($\beta = 0.024, p = 0.013, ci = [0.005, 0.043]$). Unlike gender and age, we did not observe significant differences in accuracy based on participants' ethnic backgrounds. Both majority and minority ethnic groups performed similarly across various headline types ($|\beta| \leq 0.013, p \geq 0.384$).

Framing Effect We next investigated the framing effect — how the presentation of information shapes responses (Plous 1993) — by comparing participant reactions to original versus altered headlines. We hypothesize that, in the absence of framing effects, if a headline is believed to be true, its altered version would likely be deemed false, and vice versa. Deviations from this expectation indicate a framing effect, suggesting that people respond to more than just the headline's content.

Using the Mann-Whitney U test, we tested whether the response distribution for a real headline differed from its altered counterpart. A total of 44% of headlines were found to induce a framing effect ($p < 0.05$). To further dissect this phenomenon, Figure 3a illustrates the relationship between average responses to original and altered headlines.

Group Biases Next, we contrasted responses between groups within the same category. For example, are responses to headlines reporting positive outcomes for men different from those reporting positive outcomes for women?

Figure 4a shows that participants are more prone to errors when evaluating altered headlines which report positive outcomes for white people, and real headlines which report negative outcomes for white people.

A similar analysis as a function of age, shown in Figure 4b reveals that the disparity in error between real and altered headlines is consistent across sentiments. In particular,

[*]Preregistered in (Abels et al. 2023c), see SI Section *Preregistration Discussion* for a comparison of this paper to the preregistration.

[†]Headlines are not repeated within a simulation. Thus, each decision is based on a learning history that excludes the current headlines, akin to the distinction between training and test sets in supervised learning. Specifically, at time t, the quality of a decision is evaluated based on the model's training on data from times 1 to $t-1$, testing the model's efficacy on previously unseen data.

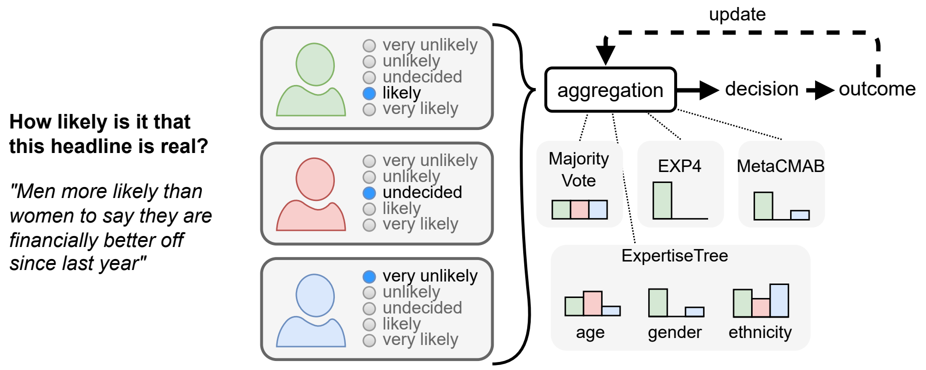


Figure 1. Overview of the Collective Decision-Making problem studied in this work. Participants (identified here by the colors green, red and blue) in the experiment were presented with a sequence of headlines and were asked to estimate the likelihood that they were true. These diverging opinions are then aggregated to reach a collective decision. The aggregation is iteratively optimized by comparing it to a ground truth. Different aggregations weigh opinions differently, see Methods for algorithmic details. Example weights for the three participants are given for each approach. A majority vote values all group members uniformly, resulting in even bars. EXP4 (Auer et al. 2002) selects a single participant (i.e., weight is concentrated on a single participant) whose opinion is followed. MetaCMAB (Abels et al. 2023a) distributes weights more evenly, correlating them with performance to enhance the collective decision-making process. ExpertiseTree (Abels et al. 2023b) similarly distributes weights, but will learn distinct models for each headline category if this proves beneficial.

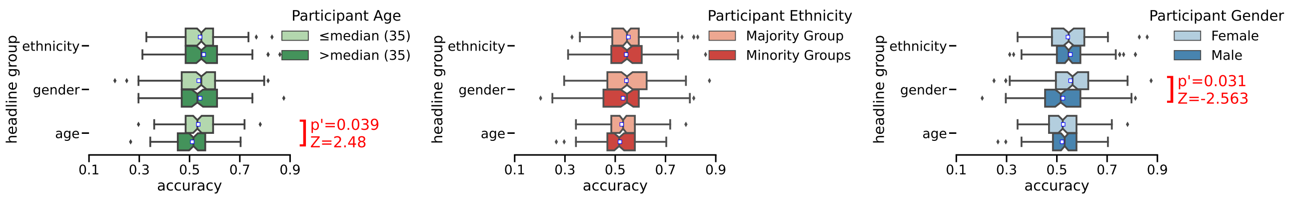


Figure 2. Participants' accuracy as a function of their demographic information and the type of headline they respond to. The figures, from left to right, depict the accuracy in responding to each headline category in function of three demographic variables: age group (threshold at the median participant age of 35), ethnicity (partitioned into majority and minority ethnicities, see Figure 7), and gender. The GEEs fitting reveals statistically significant differences (Bonferroni adjusted p-values $p' < 0.05$) in age and gender related questions. Specifically, older participants showed slightly lower accuracy in responding to age-related questions, and male participants exhibited lower accuracy in responding to questions related to gender.

it shows that participants are strong at identifying real headlines, but weak at identifying altered headlines.

Finally, the analysis by gender, given in Figure 4c reveals that errors are relatively consistent across all but one group; headlines altered to be negative for Males.

Influence of Aggregation Algorithms on Collective Decision-Making Processes

Here we compare the performance of various aggregation algorithms: selecting a participant at random (random member), a Weighted Majority Vote (WMV), EXP4 (Auer et al. 2002), MetaCMAB (Abels et al. 2023a), and ExpertiseTree (Abels et al. 2023b). Each approach is included in the study for its unique contribution to understanding and improving collective decision-making. We provide a description of these algorithms, as well as a more detailed justification for their inclusion in the methods section. We focus first on how these algorithms influence the overall decision quality compared to individual participants.

Accuracy Figure 5a shows overall accuracy as a function of group size for these algorithms and compares it to the best group member (identified in hindsight). Figure 5b gives terminal regret values, i.e., the difference in performance between the best group member and the algorithm for the final decision. Based on Wilcoxon tests (see Methods) with an $\alpha = 0.05$ significance threshold, the performance of all methods is significantly different, except for the terminal regret of EXP4 and the WMV.

Round-by-Round Analysis of Instantaneous Regret To further analyze these dynamics, we analyzed the round-by-round performance for group sizes of 4 and 36 participants, specifically focusing on instantaneous regret (Figure 5c-d). This metric measures the performance gap between each algorithm and the best performing group member at specific time steps. In particular, a negative regret indicates performance surpassing the single best group member, indicative of the emergence of collective intelligence. As it collects more experiences, ExpertiseTree approaches this threshold, and ultimately surpasses it for larger groups (compare $N = 4$ and $N = 36$, Figure 5c-d).

Win Percentage An alternative way of evaluating the prevalence of collective intelligence is by measuring how often collective decision-making algorithms surpass the performance of the single best group member. Figure 5e provides this comprehensive view of how different algorithms fare against individual participants. While all methods tend to outperform a majority of group members, MetaCMAB and the ExpertiseTree approach consistently outperform all but 2 (for MetaCMAB) or 1 (for ExpertiseTree) group members. For example, for group size

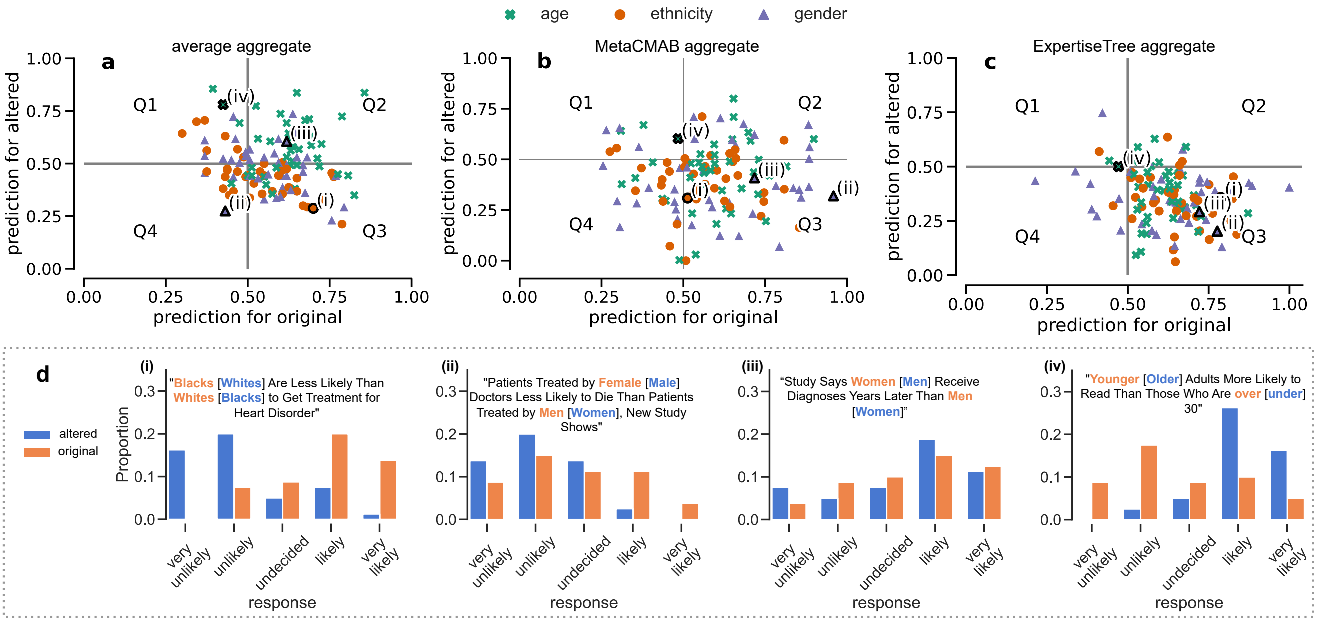


Figure 3. Effects of framing in participants' responses. **a-c**, the quadrants represent false stereotypes (Q1), positive framing effects (Q2), common knowledge (Q3), and negative framing effects (Q4). **a**, Each point represents a distinct headline; the x-coordinate displays the average response to its original form, and the y-coordinate shows the average response to its altered form. **b**, The x and y coordinates indicate MetaCMAB's model predictions for the original and altered headlines, respectively. **c**, In a similar manner, x and y coordinates reveal the predictions of ExpertiseTree's model for the original and altered headlines, respectively. **d**, Response distributions for highlighted points (i), (ii), (iii) and (iv) in panels **a-c** are given as histograms. Each point's headline and response distribution are given in its original (orange), and [altered] (blue) form.

$N = 36$, ExpertiseTree exceeds the performance of the top group member in 45% of cases (as shown in the bottom left cell of the relevant heatmap), and surpasses the second-best member in 56% of simulations (second leftmost cell). On the other hand, MetaCMAB generally matches the performance of the second-best member (with a win ratio of 0.40) and less often outperforms the top member, achieving a win ratio of 0.3.

Collective Decision-Making Models Mitigate Biases

In this section we compare and contrast the prevalence of biases in aggregations through either an average or adaptive methods MetaCMAB and ExpertiseTree. We restrict the analysis of EXP4 to the supplementary material as we found it did not improve over the performance of a simple average.

Reduced Framing Effects Our hypothesis posits that advanced collective decision-making algorithms, such as MetaCMAB and ExpertiseTree, should show a notable reduction in framing effects. This would manifest as decreased densities in Quadrants Q1 (False Stereotypes), Q2 (Positive Framing Effect), and Q4 (Negative Framing Effect) of the framing effect analysis.

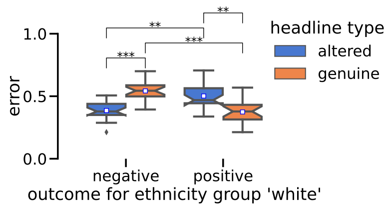
We evaluated this hypothesis by analyzing the aggregated predictions of these models (see Figure 3b for MetaCMAB and Figure 3c for ExpertiseTree). These plots indicate a reduced proportion of headlines in Q1 ($19 \to 8$) and an increased proportion of headlines in Q4 ($15 \to 25$) for MetaCMAB, as well as a reduced proportion of headlines in Q1 ($19 \to 4$) and Q2 ($34 \to 11$) and an increased proportion in Q3 ($52 \to 97$) for ExpertiseTree.

Reduced Group Biases We now evaluate whether different algorithms are predictive of different error rates conditioned on the category. This evaluation involves comparing error rates from three sources: average participant predictions, MetaCMAB predictions, and ExpertiseTree predictions. We summarize here the results and provide full GEE tables in the Supplementary Information.

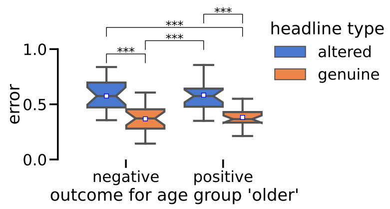
The GEE analysis for averaged responses highlights significant interactions across all variables with respect to prediction error. The baseline error (intercept) was estimated at $\beta = 0.579$ ($p < 0.001$, $ci = [0.536, 0.623]$). Negative effects were observed for headlines classified by ethnicity ($\beta = -0.135, p < 0.001, ci = [-0.204, -0.067]$) and gender ($\beta = -0.093$, $p = 0.049$, $ci = [-0.186, -0.001]$), with significant interactions noted between headline class and alteration status for both categories.

For MetaCMAB's predictions, the extent of bias was reduced. The baseline error was lower ($\beta = 0.431$, $p < 0.001$, $ci = [0.377509, 0.484283]$), and while the ethnicity headline class maintained a significant effect ($\beta = -0.087$, $p = 0.021, ci = [-0.160649, -0.012963]$), the gender effect was not significant. MetaCMAB's error is however significantly worse for genuine headlines than for altered headlines ($\beta = -0.079, p = 0.023, ci = [-0.148, -0.011]$). The interaction effect of genuine headlines also remained significant in relation to ethnicity ($\beta = 0.132$, $p < 0.001$, $ci = [0.088, 0.179]$).

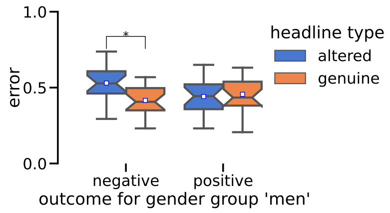
In contrast, the GEE analysis for ExpertiseTree's predictions showed that none of the variables had a significant effect on the prediction errors. The intercept was $\beta = 0.319$ ($p < 0.001$, $ci = [0.287, 0.351]$), and all other parameters concerning headline classes (both ethnicity and



(a) Headlines involving ethnicity



(b) Headlines involving age



(c) Headlines involving gender

Figure 4. Error rates categorized by sensitive attributes in headlines. Brackets indicate that there are significant differences (*, **, and *** indicating respectively p-values below 0.05,0.01, and 0.001) between pairs of boxplots as identified by Kruskal-Wallis H-tests (ethnicity: $H = 33.995$, $p < 0.001$, $df = 3$, $\eta^2 = 0.408$, age: $H = 34.167$, $p < 0.001$, $df = 3$, $\eta^2 = 0.410$ gender: $H = 10.421$, $p = 0.0153$, $df = 3$, $\eta^2 = 0.098$) followed by Dunn's tests (see Methods).

gender) and their interaction with altered status were non-significant (e.g., ethnicity headline class: $\beta = -0.053$, $p = 0.108$, $ci = [-0.118, 0.0117]$).

A comparison of the splits learned by ExpertiseTree (see Figure 6) shows its trees grow deeper as the group size increases. In particular, for large group sizes the ExpertiseTree learns distinct models for all headline categories.

Discussion

Our investigation into collective decision-making, biases, and the effectiveness of adaptive methods, reveals several key insights with implications for the field of machine intelligence.

Collective Intelligence through Online Machine Learning With no prior knowledge of the expertise of the group members, efficient collective decision-making systems need to learn to surpass the performance of the single best group member. When group members provide honest confidence estimates, these can be used to weigh their input. In our experiment, participants were given the option to express strong confidence (e.g., 'very likely'), weak confidence (e.g., 'likely'), or a lack of knowledge ('undecided'). If participant responses were properly calibrated, i.e., their expressed confidence aligned with their expected accuracy, previous work indicates that the resulting confidence-weighted majority vote (WMV) would consistently outperform the single best group member (Grofman et al. 1983). As Figure 5a illustrates, however, this is not the case: the WMV consistently performs worse than the single best group member.

To tackle this confidence bias, various online machine learning approaches have been proposed to supplement participants' self-reported confidence with algorithmically determined weights. These weights are iteratively refined as more information regarding the participants' accuracy is gathered, thereby enhancing the system's ability to adjust its aggregation over time.

We compared here three different approaches: EXP4, MetaCMAB, and ExpertiseTree. We find that MetaCMAB and ExpertiseTree, which act on an aggregate over all group members, perform significantly better than EXP4, which is designed to select the single best decision-maker in a group. In fact, as the group sizes increase, their accuracy converges towards the level achieved by the best-performing group member.

While ExpertiseTree typically does not outperform the top individual participant overall, its potential becomes more evident in later stages of the decision-making process. As shown in Figure 5b, this algorithm ultimately surpasses the best participant's performance in larger groups ($N \geq 8$). Therefore, although initial uncertainty limits overall performance, these results suggest that this algorithm eventually facilitate the emergence of collective intelligence. In contrast, while a WMV can achieve collective intelligence in specific scenarios (homogeneous performance levels and low correlation between group members, or properly calibrated confidence, see (Abels et al. 2023a)), participant responses in this context do not satisfy these necessary conditions.

Results in terms of win percentages (Figure 5e) confirm that the WMV and EXP4 algorithms consistently underperform in comparison with the top-performing group member. In contrast, algorithms emphasizing collective intelligence (MetaCMAB and ExpertiseTree) frequently outperform the single best group member.

Our findings contribute to the understanding of how machine learning algorithms can be utilized to improve the quality of collective decisions. By integrating a diverse set of individual opinions and dynamically adjusting to the varying levels of expertise, these algorithms enable the emergence of collective intelligence, which is crucial in contexts where decision-making involves complex, sensitive issues.

Overall, our findings align with prior results on the performance of these algorithms established on synthetic expertise: the use of adaptive algorithms enables a dynamic and responsive collective decision-making process. Although they initially are on par with a WMV, MetaCMAB and ExpertiseTree quickly improve their performance by adjusting the importance assigned to group members. Notably, for groups of 36, MetaCMAB and ExpertiseTree begin to surpass the best group member after 42 and 24 headlines, respectively. This demonstrates that these algorithms quickly adapt to the collective pool of knowledge, often surpassing the performance of the most skilled individual in larger groups. This collective intelligence is achieved by leveraging diverse inputs without prior knowledge about participants' expertise.

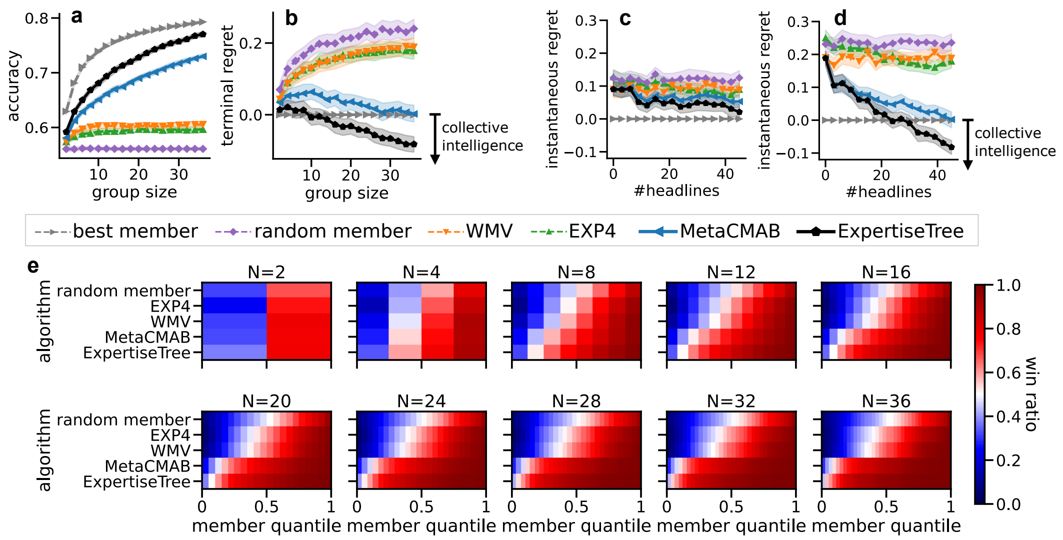


Figure 5. Performance as a function of the number of participants. Shaded areas around the lines in panels **a-d** represent 95% confidence intervals. **a**, displays the accuracy (i.e., the proportion of correctly identified headlines, see Methods) of each algorithm for each group size. **b**, shows, for each group size, the terminal regret (see Methods): the difference in performance between the best group member and the algorithm for the final decision. Negative values in **b** show when the collective outperforms the single best group member. Panels **c** and **d**, depict for, respectively, groups of size 4 and 36 the instantaneous regret as learning progresses — the difference in performance between the best group member and the algorithm for the decision at that time (see Methods). These two plots show at which point algorithms surpass the single best group member. **e**, presents the algorithmic improvement on expertise for different algorithms and different group sizes (N). Within each heatmap, group members are ranked by accuracy, and the x-axis indicates the member's performance quantiles. In particular, the lower the quantile, the more accurate the group member. For each row within a heatmap, each cell represents the proportion of cases for which the respective algorithm surpasses the corresponding group member's performance. In particular, the left-most cell in each diagram corresponds to the best group member. Non-zero values for this cell suggest the emergence of collective intelligence.

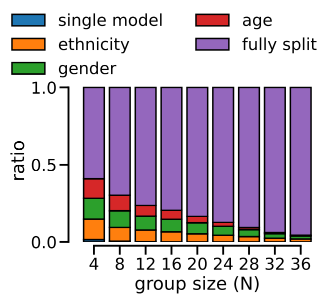


Figure 6. Prevalence of splits learned by ExpertiseTree as a function of the number of members in a group. ExpertiseTree can learn to either i) not split, ii) split off the ethnicity category, iii) split off the gender category, iv) split off the age category, v) split all categories. For (ii) — and analogously for (iii) and (iv) — this implies that one model is learned for all headlines relating to ethnicity, while another model is learned for both the gender and age category.

Headline Categories and Implicit Biases Based on prior research (Greenwald and Krieger 2006; Craig and Richeson 2016), we expected to observe differences in participants' responses to various headline categories, such as gender, ethnicity, or age. For instance, individuals might be less inclined to question stereotypes concerning age groups but more cautious when it comes to stereotypes associated with ethnic groups. Our analysis of how participants respond to different headline categories yielded notable trends in skepticism. In particular, the quadrants of Figure 3a align with four types of beliefs:

- (Q1, False Stereotypes) Stereotypes skew judgment, leading to consistent misjudgment of these headlines that align with preconceived notions.
- (Q2, Positive Framing Effect) Headlines are perceived as true regardless of alterations, indicating a lower skepticism induced by their phrasing or content.
- (Q3, Common Knowledge) Headlines matching widespread beliefs are accurately judged, irrespective of their alteration.
- (Q4, Negative Framing Effect) Converse of Q2; these headlines are consistently deemed false, pointing towards inherent skepticism or bias in their presentation.

Interestingly, headlines related to age were predominantly found in Q2 and rarely in Q4. On the contrary, ethnicity-related headlines were absent from Q2. This suggests participants generally exhibited reduced skepticism towards age-related headlines, contrasting with a heightened skepticism towards ethnic headlines. This pattern persisted even when headlines were altered to focus on different sensitive groups (see Supplementary Information); responses for a headline whose subject is changed from an ethnic group to an age group will be less skeptical. This suggests that this skepticism (or lack thereof) is not a feature of the specific headlines, but rather of the sensitive group it contains. This finding aligns with research by Nosek et al. (2007), suggesting that people are more inclined to associate age with positive or negative concepts than ethnicity. This could explain the greater likelihood of believing age-related headlines.

When detailed by subcategory, our results (in particular Figure 4a) suggest that participants in the experiments were more skeptical of headlines which report negative outcomes for white people.

These patterns can also be understood in terms of confirmation bias, the tendency for people to favor information that confirms their pre-existing beliefs or biases (Oswald and Grosjean 2004). This phenomenon, notably discussed in studies of crowdsourcing (Gemalmaz and Yin 2021; Draws et al. 2021), offers further insights into participants' responses. For example, if individuals hold implicit biases that view certain groups in a specific light, they are more likely to believe headlines that confirm these biases, and dismiss those that challenge them.

Demographic Differences in Performance Results on the Implicit Association Test (IAT) (Greenwald and Krieger 2006) suggest that most groups, with the exception of African-Americans, displayed an implicit bias against African-Americans. Similarly, Craig and Richeson (2016) found stigmatized groups can form coalitions, in the sense that members of a stigmatized group might be more sensitive to the discrimination of other stigmatized groups. We found that performance disparities between demographic groups predominantly appeared in dimensions relevant to those groups (see Figure 2). For instance, differences in gender or age influenced the accuracy of responses to headlines pertaining to those specific categories. This suggests that personal identity factors play a crucial role in shaping perceptions and biases.

Mitigating Bias in Collective Decision-Making through Machine Intelligence We demonstrated that while individual responses are prone to various biases, including those related to demographic factors, adaptive algorithms can effectively counteract these biases.

Considering the enhancements in performance shown by MetaCMAB and ExpertiseTree compared to individual predictions, we investigated if these improvements also indicate a reduction in group bias, i.e., whether the group (e.g., gender, ethnicity, or age) for which a prediction is given impacts the predictor's error.

Our analysis reveals that individuals often display biased performance. When individual responses are pooled through an average, these biases tend to permeate, resulting in biased aggregations. In contrast, adaptive aggregations acquired through for example MetaCMAB or ExpertiseTree should mitigate these biases.

The ExpertiseTree algorithm, in particular, showed a significant reduction in framing effects and group biases compared to individual responses or simple aggregation methods like the weighted majority vote.

Unlike other methods, ExpertiseTree's performance was not significantly affected by changes in headline categories, suggesting individual biases were effectively mitigated.

What is more, the distribution of points in Figure 3b suggests a decrease in false stereotypes and a reduced positive bias for MetaCMAB, but a similar prevalence of negative framing effects. In contrast, we found less populated biased quadrants for ExpertiseTree. This supports its robustness in mitigating false stereotypes and framing effects, making it a valuable tool in creating unbiased, balanced decision-making processes.

In summary, our research underscores the importance of understanding and addressing biases in human advice to enhance the performance of collective decision-making. The use of advanced machine learning algorithms, such as MetaCMAB and ExpertiseTree, offers a promising avenue for achieving this goal. By effectively aggregating diverse human opinions and dynamically adjusting to various levels of expertise and bias, these algorithms pave the way for collectively intelligent, fair, and effective decision-making processes.

Methods

In order to ground our results in a realistic problem, we presented participants with a fake news problem. In doing so, we are emulating the use of human fact checkers as a tool to combat the spread of fake news on social media. Our aim is thus to evaluate how laypeople perform at this task, what factors influence their choices, and how they could be enhanced by a collective decision-making system. Participants were presented with a series of headlines for which they were asked to express a likelihood of the headline being real. In order to elicit possibly biased responses, we focused on headlines which involved sensitive groups.

In particular, we collected headlines from (Kulkarni 2018; Mazumder et al. 2014; Med 2023) based on the following criteria:

- The headline (implicitly or explicitly) contrasts two sensitive groups (e.g., *"Men more likely than women to say they are financially better off since last year"*)
- The headline should present a clear negative or positive outcome. For example, it is not clear whether *"Poll: Kanye more popular with whites than nonwhites"* is positive or negative, but *"African-Americans, Hispanics, dying at faster rate of fentanyl overdoses than whites: analysis"* is clearly a negative outcome for African-Americans and Hispanics.

Participants were presented with a mix of such headlines: 50% were unaltered, while the other 50% had their sensitive group swapped.

Alterations were made by swapping sensitive groups with their complementary group. For example, "Men" $\leftrightarrow$ "Women", "Older people" $\leftrightarrow$ "Younger people", or "African-Americans" $\leftrightarrow$ "White Americans".

Each headline was characterized by a sensitive group (thus whether it concerned a gender group, an ethnic group, or an age group), by a sentiment (positive or negative), and by a truth value (whether the headline was real or altered).

Headlines for each participant were selected to represent a balanced mix of these features. The number of headlines required to achieve this balance can be computed as:

$$\begin{aligned}&(\text{\# of sensitive groups} \times \text{\# of sentiments})\\&\qquad\times(\text{\# of outcomes} \times \text{\# of truth values})\\&\qquad\qquad= (3 \times 2) \times (2 \times 2) = 24.\end{aligned}$$

Each treatment, or set of questions, should therefore contain a multiple of 24 headlines.

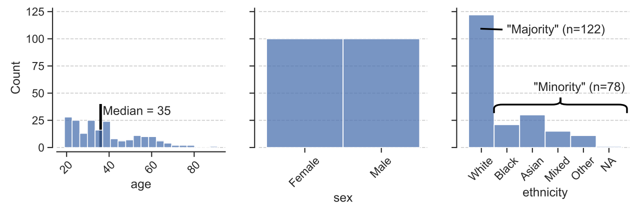


Figure 7. Demographic distribution of participants. Age categories are established by splitting participants according to the median age of 35. Ethnicity categories, Majority and Minority, are assigned according to whether participants belong to the most frequent group (i.e., White). NA = Not Available

Moreover, to understand the influence of presentation, every headline was included in both its real and altered versions, distributed across different treatments.

Consequently, our finalized dataset encompasses 5 sets of questions (=treatments), with each set containing 48 headlines. These 240 headlines contain 40 instances for each of the 6 sensitive groups. By distributing these over 5 treatments, we ensured each participant saw 8 real and 8 altered headlines for every sensitive group. Note that a positive sentiment towards one group implies a negative sentiment towards its complementary group. For example, *"Across Age Groups, Whites Fared Worse in Employment Rates"* is both negative for whites and implicitly positive for African-Americans. We thus have 16 real (8 positive and 8 negative) and 16 altered (again 8 positive and 8 negative) headlines per group.

We sequentially presented the headlines to participants and asked them to assess the likelihood of each being real. For each headline, participants were given 5 choices; {"very unlikely", "unlikely", "undecided", "likely", "very likely"}.

Participants and Data Collection

We recruited 200[‡] participants via the Prolific crowdsourcing platform (Palan and Schitter 2018). While the distribution of participants by gender was balanced, balancing participants by ethnicity or by age is more challenging through Prolific. We therefore did not impose restrictions on participation by ethnicity or by age, except that all participants are ≥ 18 years old. As a result, participants were predominantly white ($N_{\text{white}} = 122$, $N_{\text{non-white}} = 78$, see Figure 7), reflecting demographics of the sampled population.

Participants were paid a flat sum for participation at a rate of 6£ an hour. Median completion time was 8 minutes 58 seconds. Participants were uniformly assigned to one of the 5 treatments (=set of 48 headlines). Only participants who completed the questionnaire were included in our analysis. Consequently, our dataset contains a total of $5 \times 40 \times 48 = 9600$ responses.

Along with responses, we captured the following metrics: the headline's sensitive group, the order in which responses were given, participants' demographic data (age, gender, and ethnicity), headline sentiment (positive or not), its authenticity (altered or not), and response time.

Evaluating Performance

To assess the quality of participants, we gauge the accuracy of their estimates against actual outcomes. Define $p_n(h)$ as participant n's response for headline h. Following experimental evidence on how humans interpret these categories (Wintle et al. 2019), we map the responses "very unlikely", "unlikely", "undecided", "likely", and "very likely" to respectively 0, 0.25, 0.5, 0.75, and 1.

The true value of h is denoted as $y(h)$, with $y(h) = 0$ for altered headlines and 1 otherwise. Participant error is then quantified in terms of absolute difference: $\epsilon(h, n) = |p_n(h) - y(h)|$. Conversely, the accuracy is the complement of the error: $a(h, n) = 1 - \epsilon(h, n)$, and takes as value 1 if participant n predicts the headline's class with high confidence.

In this setting, bias can be assessed by observing whether participants are more likely to associate specific outcomes (either negative or positive) with certain groups. Additionally, differences in accuracy or error rates based on the sensitive group featured in the headline can also serve as indicators of bias. For example, participant errors for headlines reporting positive outcomes for white people could be significantly larger than those same participants' errors for headlines reporting negative outcomes for white people. Such differences are explored in Figure 4 and its related discussion.

Collective Decision-Making Process

While it is typically hard to properly estimate the performance of participants a priori, online learning algorithms allow us to learn how to aggregate expertise as we encounter new cases. We outline this decision-making setting in Algorithm 1.

Algorithm 1 Collective Decision-Making for Headline Selection

```
Initialize learner (e.g., EXP4 (Auer et al. 2002), MetaCMAB (Abels et al. 2023a), or ExpertiseTree (Abels et al. 2023b)) with initial aggregation policy
for t = 1, 2, ..., T do
    Participants observe the set of headlines for round t
    Each participant rates the headlines on how likely they are to be real
    Aggregate their ratings to select a headline
    Collect reward of r_t = 0 if chosen headline was altered, 1 otherwise
    Use collected penalty to update aggregation policy for t + 1
end for
```

To focus on individual biases, we designed our experiment without any interaction between participants, ensuring that the collected responses were independent. Following this, we sampled subgroups from the participant pool to evaluate the performance of the different online learning algorithms.

Specifically, we bootstrap, for every group size N and for each treatment (set of headlines), 1000 subsets of N participants with which we simulate this collective decision-making process for each of the methods we now describe.

[‡] Additionally, 26 participants who did not complete the questionnaire were excluded from this analysis.

Chosen Algorithms Several methods have been developed to improve decision-making with expert[§] advice. Among them, we select four pivotal approaches that span from fundamental to state-of-the-art, providing a comprehensive view of online learning algorithms for collective decision-making. These approaches include: a simple majority vote serving as a baseline for comparison; EXP4 (Auer et al. 2002), which uses observed performance to gradually pinpoint the single best participant; MetaCMAB (Abels et al. 2023a), an online learning approach that enables more collective decisions by integrating diverse opinions; and ExpertiseTree (Abels et al. 2023b), which tailors aggregation strategies to specific contexts.

As illustrated in Figure 1, a key differentiator among these methods is their strategy for assigning weight to participants' opinions, ranging from equal weighting to complex, performance- and context-based adjustments. Crucially, the more advanced approaches (MetaCMAB and ExpertiseTree) learn to weight participants based on their responses, not only relative to the truth, but also relative to each other. These methods not only prioritize contributions from the highest performers, resembling strategies that rely on the collective input of empirically proven subsets of participants (Mannes et al. 2014), but they also incorporate mechanisms to offset the influence of highly polarized groups of experts. By recognizing and adjusting for common errors among participants, these approaches can reduce the negative effects of polarized groups, whose insights, though potentially insightful, overlap significantly with those of others (Abels et al. 2023a). Accounting for correlation among participants in this way is somewhat similar to approaches which explicitly contrast individual's contributions with that of the group (Budescu and Chen 2015). The main distinction between those alternative approaches and the ones we test in this study is the inclusion of exploration techniques which prevent premature reliance on a limited group of experts.

This selection of approaches is intentionally diverse, covering both traditional and state-of-the-art methods to highlight the spectrum of possibilities for aggregating advice. From the benchmark simplicity of majority voting to the nuanced, context-sensitive aggregation offered by ExpertiseTree, each method contributes uniquely to our understanding of effective decision-making in varied scenarios. A more detailed description of these chosen methods and the rationale behind their selection is provided in the supplementary matieral.

Measures of Collective Decision-Making Performance Because we consider an online learning setting, we evaluate the performance on two metrics. First, on overall accuracy, i.e., the proportion of rounds for which an optimal decision was made. And secondly, on final performance, i.e., the expected performance at the end of training. In addition, we can analyze how performance changes as a function of the number of participants.

Given our objective of enhancing the probability of choosing authentic headlines, our goal becomes maximizing the *average reward* defined as $\bar{\mathcal{R}}_T = \frac{1}{T}\sum_{t=1}^{T} r_t$, where $r_t = 1$ if the chosen headline is genuine and 0 otherwise.

We denote by $\bar{\mathcal{R}}_T^n$ the average reward we would obtain if we were to select participant n and choose at each round the headline which they rate highest. Maximizing over the set of participants, we can in hindsight select the optimal participant as being $n^* = argmax_{n=1}^{N} \bar{\mathcal{R}}_T^n$.

To facilitate comparison to the single best participant, we can juxtapose an algorithm's performance against the best participant's performance. Specifically, for each round, we compute the instantaneous regret:

$$R_t = r_t^{n^*} - r_t \tag{1}$$

Of particular interest is the instantaneous regret at time $t = T$, i.e., R_T, which we refer to in the main text as terminal regret.

If our performance is lower than the best participant, this metric averages out positively; otherwise, it is negative. Thus, observing negative regret suggests the presence of collective intelligence.

Code and Data availability

Code and data, including participant responses, necessary to reproduce our results are available at respectively https://doi.org/10.5281/zenodo.10804450 and https://doi.org/10.5281/zenodo.10794209.

Statements and Declarations

Ethical considerations

Due to the sensitive nature of the questions, participants were informed ahead of the questionnaire that "Some of the questions presented in this study involve sensitive characteristics such as gender and ethnicity. Please be advised that exposure to scenarios involving these characteristics may potentially make you feel discomfort.". In addition, ethical approval was granted on February 16th 2023 by the host institution's ethics board before data collection.

Consent to participate

To ensure ethical compliance, informed consent was obtained from all participants through the user interface. This consent process involved clearly explaining the purpose of the study, the types of data that would be collected, how the data would be used, and the participants' rights.

Consent for publication

Participants were asked to consent to their responses being published as part of a dataset accompanying any research results.

Declaration of conflicting interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article

[§] The term "expert" suggests a high level of competence among group members, however this is not required. Algorithms for this setting are designed to mitigate the influence of weaker advisors.

Funding

A.A. was supported by a FRIA grant (nr. 5200122F) by the National Fund for Scientific Research (F.N.R.S.) of Belgium. T.L. is supported by the F.N.R.S. project with grant number 40007793, the Service Public de Wallonie Recherche under grant n°2010235–ARIAC by DigitalWallonia4.ai. A.A., T.L. and A.N. benefit from the support of the Flemish Government through the AI Research Program. T.L. and A.N. acknowledge the support by TAILOR, a project funded by EU Horizon 2020 research and innovation program under GA No 952215. E.F.D. is supported by an F.N.R.S Chargé de Recherche position, grant number 40005955.

References

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157): 1124–1131, 1974.

Irving L Janis. Groupthink. *IEEE Engineering Management Review*, 36(1):36, 2008.

Dan-Olof Rooth. Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, 17(3):523–534, 2010.

Zhiyuan 'Jerry' Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. The limits of human predictions of recidivism. *Science advances*, 6(7):eaaz0652, 2020.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878, 2002.

Amy JC Cuddy, Susan T Fiske, Virginia SY Kwan, Peter Glick, Stéphanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, et al. Stereotype content model across cultures: Towards universal similarities and some differences. *British journal of social psychology*, 48(1):1–33, 2009.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

Anthony G Greenwald and Linda Hamilton Krieger. Implicit bias: Scientific foundations. *California law review*, 94(4):945–967, 2006.

Nigel Harvey. Confidence in judgment. *Trends in cognitive sciences*, 1(2):78–82, 1997.

Alan D Baddeley and Graham Hitch. The recency effect: Implicit learning with explicit retrieval? *Memory & Cognition*, 21:146–155, 1993.

Martin Jones and Robert Sugden. Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50:59–99, 2001.

Rüdiger Pohl. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press, 2004.

Marquis de Condorcet. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, page 1785, 1785.

John Rawls. Atheory of justice. *Cambridge (Mass.)*, 1971.

James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pages 1739–1776. PMLR, 2018.

Axel Abels, Tom Lenaerts, Vito Trianni, and Ann Nowé. Dealing with expert bias in collective decision-making. *Artificial Intelligence*, 320:103921, 2023a.

Axel Abels, Tom Lenaerts, Vito Trianni, and Ann Nowe. Expertise trees resolve knowledge limitations in collective decision-making. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 79–90. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/abels23a.html.

Kunkun Pang, Mingzhi Dong, Yang Wu, and Timothy M Hospedales. Dynamic ensemble active learning: A non-stationary bandit with expert advice. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2269–2276. IEEE, 2018.

Axel Abels, Tom Lenaerts, Vito Trianni, and Ann Nowé. How expert confidence can improve collective decision-making in contextual multi-armed bandit problems. In *International Conference on Computational Collective Intelligence*, pages 125–138. Springer, 2020.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Axel Abels, Elias Fernández Domingos, Ann Nowé, and Tom Lenaerts. Be the expert, Feb 2023c. URL https://osf.io/bs8tp/?view_only=80c8e9ddc622425289a1d404d8aa0836.

Ricky J Sethi. Crowdsourcing the verification of fake news and alternative facts. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 315–316, 2017.

Xuan Wei, Zhu Zhang, Mingyue Zhang, Weiyun Chen, and Daniel Dajun Zeng. Combining crowd and machine intelligence to detect false news on social media. *Mis Quarterly*, 2022.

Scott Plous. *The psychology of judgment and decision making.* Mcgraw-Hill Book Company, 1993.

Bernard Grofman, Guillermo Owen, and Scott L. Feld. Thirteen theorems in search of the truth. *Theory and Decision*, 15 (3):261–278, Sep 1983. ISSN 1573-7187. doi: 10.1007/BF00125672.

Maureen A Craig and Jennifer A Richeson. Stigma-based solidarity: Understanding the psychological foundations of conflict and coalition among members of different stigmatized groups. *Current Directions in Psychological Science*, 25(1): 21–27, 2016.

Brian A Nosek, Frederick L Smyth, Jeffrey J Hansen, Thierry Devos, Nicole M Lindner, Kate A Ranganath, Colin Tucker Smith, Kristina R Olson, Dolly Chugh, Anthony G Greenwald, et al. Pervasiveness and correlates of implicit attitudes and stereotypes. *European review of social psychology*, 18(1):36–88, 2007.

Margit E Oswald and Stefan Grosjean. Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 79, 2004.

Meric Altug Gemalmaz and Ming Yin. Accounting for confirmation bias in crowdsourced label aggregation. In *IJCAI*, pages 1729–1735, 2021.

Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 9, pages 48–59, 2021.

Rohit Kulkarni. A Million News Headlines, 2018. URL https://doi.org/10.7910/DVN/SYBGZL.

Sahisnu Mazumder, Bazir Bishnoi, and Dhaval Patel. News headlines: What they can tell us? In *Proceedings of the 6th IBM Collaborative Academia Research Exchange Conference (I-CARE) on I-CARE 2014*, pages 1–4, 2014.

Medscape. https://www.medscape.com, 2023.

Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

Bonnie C Wintle, Hannah Fraser, Ben C Wills, Ann E Nicholson, and Fiona Fidler. Verbal probabilities: Very likely to be somewhat more confusing than numbers. *PLoS One*, 14(4): e0213522, 2019.

Albert E Mannes, Jack B Soll, and Richard P Larrick. The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276, 2014.

David V Budescu and Eva Chen. Identifying expertise to extract the wisdom of crowds. *Management science*, 61(2):267–280, 2015.

Jean Antoine Nicolas et al. *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix. Par m. le marquis de Condorcet,..* de l'Imprimerie Royale, 1785.

Algorithms

The problem we study in this work can be formalized as a problem of bandits with expert advice (Auer et al. 2002).

A straightforward approach in this setting is to perform static aggregation of participant opinions. In particular, decisions can be made by acting on an arithmetic mean of all opinions, resulting in a **majority vote**. Such votes are most effective when expertise is homogeneous and participants' errors are uncorrelated (Nicolas et al. 1785; Grofman et al. 1983). Participants however often differ in their performance, and can show varying degrees of correlation due to shared experiences or stereotypes. Both factors degrade the performance of the simple average. By giving all group members an equal amount of influence, we dilute the knowledge of more accurate experts. The majority vote serves as a fundamental baseline for evaluating more sophisticated algorithms. Its simplicity and widespread use in decision-making scenarios make it a valuable point of comparison.

If information about the experts' expected performance is available, votes can instead be weighted as a function of this performance estimate. The exponential weighting method **EXP4** (Auer et al. 2002) for example, maintains weights over the experts, which are updated to favor participants with higher performance. For each decision, it selects one of the participants according to its current probability distribution. This single participant's opinion is then acted upon. In this sense, the decisions made by EXP4 are not collective, as a single participant is decisive. However, among algorithms that select a single participant, EXP4 has optimal theoretical guarantees and is straightforward to implement, leading to its inclusion in this study.

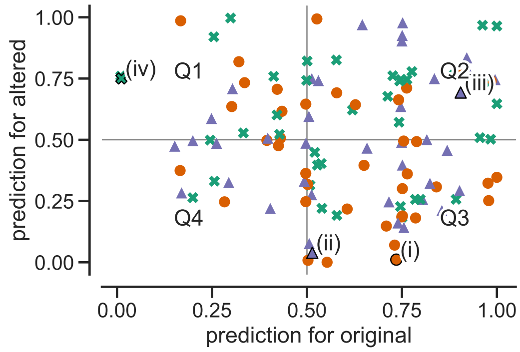


Figure 8. Effects of framing in participants' responses for EXP4. This plot is analogous to those in Figure 3; the quadrants represent false stereotypes (Q1), positive framing effects (Q2), common knowledge (Q3), and negative framing effects (Q4). Each point represents a distinct headline; the x-coordinate displays the expected response to its original form, and the y-coordinate shows the expected response to its altered form.

The reliance on a single expert by EXP4 underscores a significant limitation: the lack of collective decision-making. Recent advancements have introduced algorithms like **MetaCMAB** (Abels et al. 2023a), which, by employing a regression across all participant opinions, enables a more nuanced combination of expert advice. In doing so, MetaCMAB can optimally combine the opinions of diverse groups in order to enable better decisions. This approach allows MetaCMAB to outperform even the best individual participant, showcasing the potential of collective intelligence.

Nonetheless, MetaCMAB and similar algorithms assume uniform performance across different contexts (here, headline categories), an assumption rarely met in practice. Performance can vary significantly with the context, influenced by factors like prior experience and cognitive biases. Such heterogeneous performance permeates into collective decisions, meaning the optimal weighting of opinions also is a function of the headline category. To tackle such heterogeneity, the **ExpertiseTree** (Abels et al. 2023b) learns to split the context space (consisting here of the headline categories) and then fits distinct models to each region. Note that regions are only split if it induces a better estimated performance. In theory, this allows ExpertiseTree to adapt the granularity of its models to the available expertise.

Framing Effect for EXP4

We provide in Figure 8 the framing effects for EXP4 (analogously to those in Figure 3). EXP4 tends to concentrate its weights on a single group member, resulting in predictions which are often aligned with that group member's advice (and are thus close to $0, 0.25, 0.5, 0.75$, or 1). In terms of shifts from the simple average, we find the following: Q1: $19 \rightarrow 16$, Q2: $34 \rightarrow 41$, Q3: $52 \rightarrow 49$, and Q4: $15 \rightarrow 14$. While there are slight shifts from Q3 to Q2, overall these framing effects are comparable to the average.

In terms of group biases, captured by the GEE summary in Table 2, none of the features seem to have a significant impact, however the intercept indicates the model's error is generally large.

Transposed Categories

Our results in the main text suggest that headline categories induce different biases. For example, we found that age-related headlines gave rise to low skepticism, whereas headlines related to ethnicity gave rise to high skepticism. To eliminate the hypothesis that this is not due to the categories but instead is a result of the specific headlines presented in each category, we ran a secondary experiment wherein the sensitive groups were swapped. For example, we swapped a headline's age group by a gender group. If our hypothesis that the biased responses are due to the groups themselves, rather than to the specific headlines, is true, we would expect to find the same patterns of (lack of) skepticism in these altered headlines. In contrast, if the (lack of) skepticism was a result of the headlines themselves, we would find that when permuting categories, the categories susceptible to skepticism would also change.

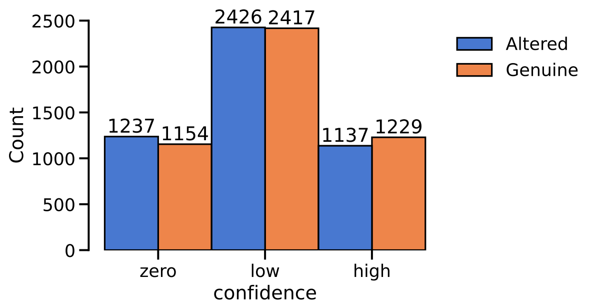
Our results suggest that our hypothesis holds, as we find that the number of age headlines in Q4 is low both when generating age headlines from gender headlines (2/80 headlines in Q4), and when generating age headlines from ethnicity headlines (3/80 headlines in Q4).

In contrast, ethnicity headlines are absent from Q2 when generated from age headlines (0/80 headlines in Q2) or from gender headlines (again 0/80 headlines in Q2).

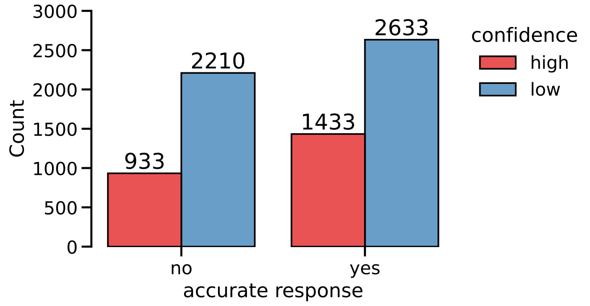
GEE tables

Tables 1 to 4 summarize the GEE analyses in support of the group bias discussion of paragraph ***"Reduced Group Biases"***. These tables analyze the effects of headline class (gender, ethnicity, age) and alteration status (genuine vs. altered) on error. Headline class variables indicate the headline's focus, with the baseline being age. "Genuine" takes as value 1 when the headline is unaltered. Interaction terms (e.g., Gender+Genuine) explore the combined effect of headline class and authenticity.

Reference characteristics are age and altered. Therefore, for example, an estimate of 0.579 for the intercept in Table 1 indicates that for an altered headline conveying an outcome about age groups, the expected error is 0.579. Other estimates are coefficients, such that, for example, a value of -0.093 for the "Headline Class: Gender" parameter indicates that the expected error for altered headlines about gender is $0.579 - 0.093 = 0.486$. In addition, the associated p-value indicates that this difference is significant. We therefore conclude that, when using an average to aggregate, there is a significant difference in error between genuine and altered headlines. Conversely, for MetaCMAB for example (Table 3), the same parameter's estimate is -0.036, but with



(a) Distribution of confidence levels as a function of altered status.



(b) Distribution of accuracy as a function of confidence level.

Figure 9. Confidence histograms

a p-value of 0.660, suggesting this feature no longer has a significant effect.

Secondary Questions

Inclination Towards Conservative or Assertive Responses: Contrast by Error and Demographic Group

People tend to be conservative in their answers, being more likely to answer (un)likely rather than very (un)likely (see Figure 9a).

Note that detailing by demographic groups (Figure 10) shows that in general, all demographic groups tend to more low confidence than high confidence. One exception to this observation is the group of black females which participated in this study, as they tend to express low and high confidence in equal measure. Furthermore, people do tend to answer more extremely when correct than when incorrect, indicating that their confidence estimates are somewhat calibrated. In particular, when maximally confident, participants have an expected accuracy of 60.6%, while, when less confident, they have an expected accuracy of 54.2%.

While the difference is small, it suggests that weighing the participants' contribution by their own confidence is likely to be beneficial. Comparing the accuracy of a majority vote with a confidence-weighted majority vote, we find that the latter obtains an accuracy of 64.0%, while the former's accuracy is 54.8%. A Wilcoxon test suggests that this difference is significant ($p = 0.0008$).

Response Time as an Indicator of Error Rates

We observe two trends in Figure 11. First, the fastest response times induce the highest errors. This suggests that for those response times, users simply provided answers

Parameter	Estimate	Std. Error	t-value	p-value
Intercept	0.579	0.022	26.085	< 0.001
Headline Class: Ethnicity	-0.135	0.035	-3.866	< 0.001
Headline Class: Gender	-0.093	0.047	-1.972	0.049
Genuine	-0.204	0.040	-5.045	< 0.001
Headline Class: Ethnicity:Genuine	0.219	0.058	3.782	< 0.001
Headline Class: Gender:Genuine	0.155	0.063	2.435	0.015

Table 1. GEE summary for the average, predicting the aggregate's error as a function of headline characteristics.

Parameter	Estimate	Std. Error	t-value	p-value
Intercept	0.537	0.046	11.712	< 0.001
Headline Class: Ethnicity	-0.088	0.057	-1.553	0.120
Headline Class: Gender	-0.036	0.081	-0.440	0.660
Genuine	-0.126	0.110	-1.144	0.253
Headline Class: Ethnicity:Genuine	0.034	0.118	0.285	0.776
Headline Class: Gender:Genuine	-0.084	0.170	-0.492	0.622

Table 2. GEE summary for EXP4

Parameter	Estimate	Std. Error	t-value	p-value
Intercept	0.431	0.027	15.819	< 0.001
Headline Class: Ethnicity	-0.087	0.038	-2.304	0.021
Headline Class: Gender	-0.056	0.039	-1.417	0.156
Genuine	-0.079	0.035	-2.274	0.023
Headline Class: Ethnicity:Genuine	0.134	0.023	5.780	< 0.001
Headline Class: Gender:Genuine	0.110	0.068	1.610	0.107

Table 3. GEE summary for MetaCMAB

Parameter	Estimate	Std. Error	t-value	p-value
Intercept	0.319	0.017	19.326	< 0.001
Headline Class: Ethnicity	-0.053	0.033	-1.606	0.108
Headline Class: Gender	-0.043	0.037	-1.154	0.249
Genuine	0.003	0.024	0.136	0.892
Headline Class: Ethnicity:Genuine	-0.015	0.035	-0.426	0.670
Headline Class: Gender:Genuine	0.029	0.026	1.080	0.280

Table 4. GEE summary for ExpertiseTree

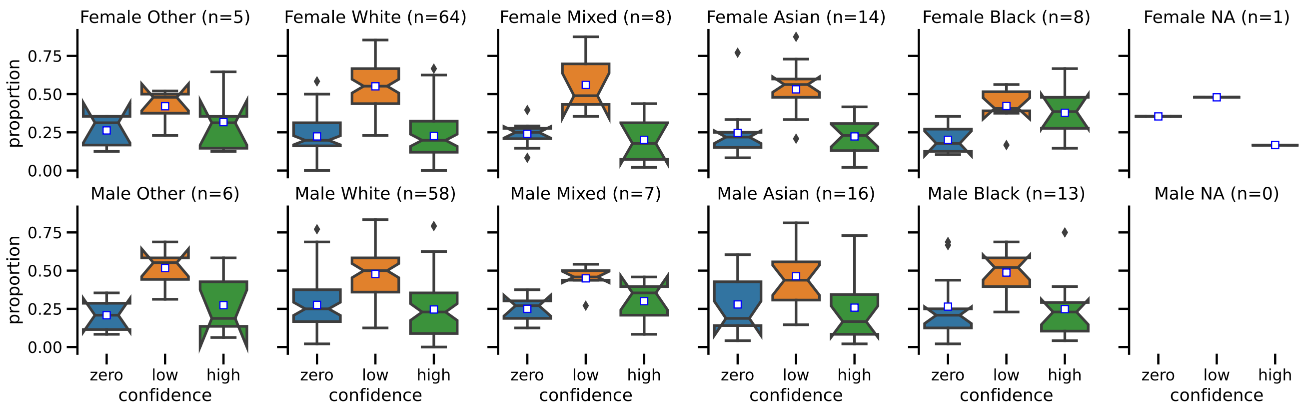


Figure 10. Proportion of confidence categories displayed by participant demographic groups.

without taking the prompts into account. Conversely, as response time increase, the error rate also tends to increase. These slower response time could be a sign of greater uncertainty, which correlates with more mistakes.

Performance of Demographically Diverse Groups

While we found that women perform slightly better than men, this individual improvement is not reflected in an

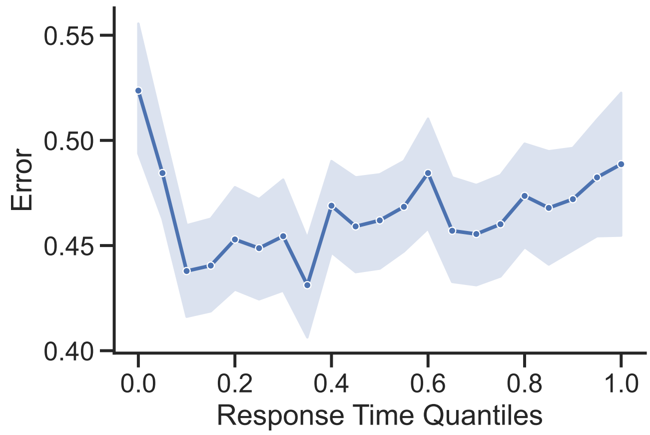


Figure 11. Error by response time.

improved performance of groups; we find that acting on the average of all female responses is not superior to acting on the average of all male responses (based on a Wilcoxon test). This is likely because, while women tend to perform better on average, they are more strongly correlated than men; men have an average Pearson correlation coefficient of 0.178, while women have an average correlation coefficient of 0.220. In practical terms, higher correlation within groups will result in group members making more of the same errors, which are then reinforced by an average aggregate. In contrast, lower correlation means individuals tend to make errors on different instances, which are more easily subdued by an average.

We might therefore wonder whether mixing demographic groups is a better approach to maximizing group performance than selecting the group with the best individuals. Unfortunately we found no significant improvement from mixed groups (again using Wilcoxon tests). While we might hope that mixed groups would be less correlated than the individual groups which compose them, we found that instead the correlation of a mixed group was approximately equal to the average correlation of the individual groups. In particular this means that a mixed group is typically more correlated than the least correlated individual group. For example, the correlation of a group with equal number of women and men is 0.194, approximately halfway the correlation values of the individual groups given above.