The wisdom of extremized crowds: Promoting erroneous divergent opinions increases collective accuracy

Autoría ditelliana: Navajas, Joaquín

Otras autorías: Barrera Lermarchand, Federico; Balenzuela, Pablo; Bahrami,Bahador; Deroy, Ophelia

Fecha de publicación: 2023

The wisdom of extremized crowds: Promoting erroneous divergent opinions increases collective accuracy

Federico Barrera Lermarchand[1,2,3], Pablo Balenzuela[2,3], Bahador Bahrami[4], Ophelia Deroy[5,6,7], & Joaquin Navajas[1,2,8]

[1] Laboratorio de Neurociencia, Universidad Torcuato Di Tella, Buenos Aires, Argentina

[2] National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina

[3] Physics Department, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina

[4] Crowd Cognition Group, Department of General Psychology and Education, Ludwig Maximilian University, Munich, Germany

[5] Faculty of Philosophy, Ludwig Maximilian University, Munich, Germany

[6] Munich Centre for Neuroscience, Ludwig Maximilian University, Munich, Germany

[7] Institute of Philosophy, School of Advanced Study, University of London, London, UK

[8] Escuela de Negocios, Universidad Torcuato Di Tella, Buenos Aires, Argentina

Corresponding author: Joaquin Navajas (joaquin.navajas@utdt.edu)

Abstract

The aggregation of many lay judgements can generate surprisingly accurate estimates. This effect, known as the "wisdom of the crowd", has been demonstrated in domains such as medical decision-making, fact-checking news, and financial forecasting. Therefore, understanding the conditions that enhance the wisdom of the crowd has become a crucial issue in the social and behavioral sciences. Previous theoretical research identified two key factors driving this effect: the accuracy of individuals and the diversity of their opinions. Most available strategies to enhance the wisdom of the crowd have exclusively focused on improving individual accuracy while neglecting the potential of increasing opinion diversity. Here, we study a complementary approach to reduce collective error by promoting divergent and extreme opinions, using a cognitive bias called the "anchoring effect". This method proposes to anchor half of the crowd to an extremely small value and the other half to an extremely large value before eliciting and averaging their estimates. As predicted by our mathematical modeling, three behavioral experiments demonstrate that this strategy concurrently increases individual error, opinion diversity, and collectively accuracy. Most remarkably, we show that this approach works even in a forecasting task where the experimenters did not know the correct answer at the time of testing. Overall, these results not only provide practitioners with a new strategy to forecast and estimate variables but also have strong theoretical implications on the epistemic value of collective decision-making.

Introduction

The aggregation of many lay estimates often outperforms the expert individual judgement (De Condorcet, 1785; Galton, 1907). This phenomenon, popularly known as the "wisdom of the crowd" (Surowiecki, 2005), has been applied to a wide range of problems such as improving medical diagnoses (Kurvers et al., 2016), forecasting geopolitical events (Mellers et al., 2014), predicting financial markets (Ray, 2010), reverse-engineering the smell of molecules (Keller et al., 2017), and fact-checking news (Allen, 2021), among many others. Given its practical relevance, understanding the conditions under which crowds produce accurate estimates has become a relevant issue in the behavioral sciences (Kameda, Toyokawa, & Tindale, 2022; Karachiwalla & Pinkow, 2021; Navajas et al., 2018; Kao & Couzin, 2014).

One important driver of collective accuracy is the diversity of opinions in the crowd (Hong & Page, 2004; Page, 2008; Becker, Porter & Centola, 2019; Shi et al., 2019; Jönsson, Hahn & Olsson, 2015). A simple intuition underlies this claim: when crowds produce diverse estimates, it is likely that some individuals will underestimate the correct answer, while others will overestimate it. Therefore, the more diverse the crowd, the higher the chance that individual errors will cancel out in the aggregation process. More formally, the "Diversity Prediction Theorem" (Page, 2007) states that the crowd's error (E) can be expressed as the mean individual error (ε) minus the crowd's predictive diversity (δ, also known as the population variance):

$$E = \varepsilon - \delta \quad [1]$$

One implication of this theorem is that, in principle, the crowd's accuracy could be increased by reducing the individual error (ε) or, equivalently, by increasing the predictive diversity (δ). Indeed, an extensive literature has demonstrated that it is possible to increase the wisdom of the crowd by reducing the individual error. For example, previous studies have proposed to aggregate information from "select" crowds composed by individuals who are more accurate across estimation problems (Mannes, Soll & Larrick, 2014). Other studies have shown that individual error can be reduced by counteracting individual biases

(Kao et al., 2018) or by exposing individuals to social information (Jayles et al., 2017; Frey & Van de Rijt, 2021; Madirolas & de Polavieja, 2015; Lorenz et al., 2011). All these different approaches share a common feature: they decrease collective error by increasing accuracy at the individual level (i.e. they reduce mean individual error), while preserving, to varying degrees, some of the initial diversity of opinions. However, although Eq. 1 suggests that it should be equally possible to reduce the crowd error (E) by increasing its predictive diversity (δ), there is hardly any empirical evidence examining this implication of the theorem.

Showing that it is possible to decrease collective error by increasing diversity is non-trivial for several reasons. First, from a theoretical point of view, it would provide empirical evidence for a hitherto untested consequence of the Diversity Prediction Theorem (Eq. [1]). Second, from a practical standpoint, it would provide practitioners with a novel approach to increase collective estimation accuracy. Third, given that the wisdom of crowds has been previously interpreted as empirical evidence for the epistemic value of democratic judgements, this putative dissociation between individual and collective accuracy should mitigate concerns about the increase of misinformed voters in recent elections.

Put together, one converges to a counterintuitive, albeit somewhat uncomfortable possibility: if collective error is reduced by increasing diversity, this would imply that the wisdom of the crowd may be enhanced by persuading individuals to adopt erroneous divergent opinions. In this paper, we present theoretical simulations and empirical evidence for this claim. We introduce a new approach to increase collective accuracy by boosting the crowd's predictive diversity at the expense of reducing individual accuracy, even when the truth is completely unknown and unavailable, including to the experimenters.

Results

We propose to promote the adoption of extreme estimates, and therefore to increase diversity, by means of a cognitive bias known as the anchoring effect (Tversky & Kahneman, 1974). The proposed method consists in anchoring one half of the crowd to a small value ("low anchor") and the other half to a large value ("high anchor"), and then averaging all estimates. We hypothesized that this technique should lead to an increase in the predictive diversity that surpasses the increase in mean individual error, thus leading to lower collective error. Using a simple mathematical model, we first demonstrated that this method is expected to enhance collective accuracy across a wide range of parameters. We then empirically tested the procedure across three different experiments and showed that it indeed leads to a substantial reduction of collective error.

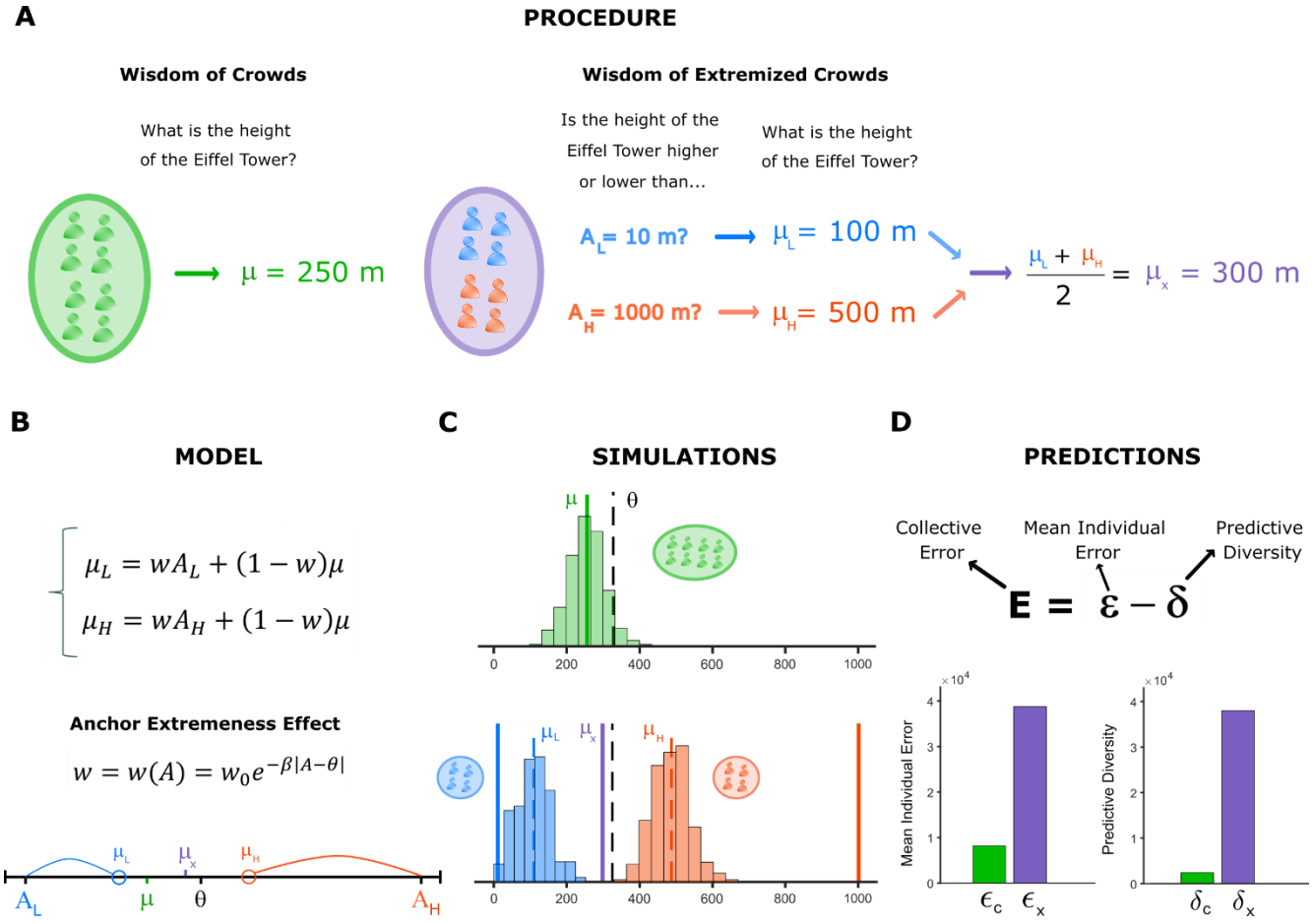

Fig. 1. Wisdom of extremized crowds. (A) The method for estimating a variable through the wisdom of crowds consists in asking a crowd of individuals to estimate a given quantity, and averaging all answers. The proposed alternative method consists in promoting the adoption of divergent opinions wihin a crowd by dividing it in two halves, asking an anchoring question with either a low or a high value to each half, and then averaging all answers. **(B)** Mathematical model of the anchoring effect. The anchored mean is a weighted average of the anchor and the wisdom-of-crowds value. The weight *w* depends on the difference between the anchor and the correct answer, reflecting an internal sensitivity to the correct answer. **(C)** Simulations performed using the proposed model show that the method is expected to outperform the wisdom of crowds. **(D)** Mean individual error and predictive diversity on simulated data show an increase in both individual error and predictive diversity. However, the increase in predictive diversity (right panel) is larger than the increase in individual error (left panel), resulting in an overall reduction in collective error.

Procedure

Let us consider the scenario where someone needs to estimate a numerical variable that is unknown to them; for example, the height of the Eiffel Tower. Based on the standard wisdom-of-crowds effect, one could obtain an approximate value by asking a large number of individuals to provide an estimate. Then, to estimate the height of the Eiffel Tower, the person would aggregate these values, for example, by averaging them (**Fig. 1A**). In this work, we propose an alternative approach that consists in dividing the crowd into two halves and extremizing opinions in opposite directions (**Fig. 1B**). We suggest doing so by using the anchoring effect: before estimating the relevant variable, individuals are first asked to consider either an extremely low or high value. In the previous example, half of the individuals would be asked to consider if the height of the Eiffel Tower is greater or less than 10 meters (low anchor, A_L) and the other half, if it is greater or less than 1000 meters (high anchor, A_H). After providing a categorical answer to this initial question, all individuals would then be asked to provide their best-guess estimate. An extensive literature has shown that these estimates should then be consistently biased towards the initially considered values (Furnham & Boo, 2011; Röseler et al., 2022). Because these anchors are extreme in opposite directions, this procedure should radicalize the estimates produced by the crowd as a whole, leading to an increase in predictive diversity. We therefore propose to average all numbers, across both halves of the crowd.

While this procedure requires pre-defining two extreme values that will be used as anchors, the strategy does not require knowing the correct answer. However, reasonably, its accuracy will depend on the specific choice of anchors. Therefore, to better understand the conditions under which the proposed approach is expected to increase collective accuracy, we developed a simple mathematical model of the anchoring effect.

Model

We consider a set of individuals who, being asked to estimate the variable θ, produce a distribution of values with mean μ. The model assumes that, when those individuals are anchored to a low value A_L, they produce a set of estimates with a different mean

$$\mu_L = w_L A_L + (1 - w_L)\,\mu, \qquad [2]$$

where $0 \leq w_L \leq 1$ is an "anchoring index" reflecting the strength of the anchoring procedure given by the anchor A_L. Similarly, a population anchored to a high value A_H would produce a distribution of estimates given by

$$\mu_H = w_H A_H + (1 - w_H)\,\mu, \qquad [3]$$

where $0 \leq w_H \leq 1$ is the corresponding "anchoring index" given by anchor A_H. Eq. [2] and [3] imply that the anchored mean is a weighted average of the anchor and the mean of the original distribution of values μ.

Following a variety of empirical findings linking the anchoring effect to the plausibility of the anchor (Mussweiler & Strack, 2001; Wegener et al., 2001), we propose that the weights w are given by

$$w_j = w_0 e^{-\beta|A_j - \theta|} \qquad [4]$$

where j indicates whether the weight corresponding to the low or high anchor (w_L or w_H, respectively) and w_0 is a parameter reflecting the anchoring index when individuals are anchored to the correct value θ. The parameter β is an "inverse temperature" encoding the sensitivity of the individuals to the distance between the anchor and θ. Thus, the value of β modulates the strength of the "anchoring extremeness effect" (Röseler et al., 2022).

In this work, we propose averaging estimates from two populations of individuals, each of which is anchored to either a low or a high value (A_L or A_H). Assuming both populations are equally sized, we can estimate the mean estimates from both populations as

$$\mu_r = \frac{\mu_L+\mu_H}{2} \quad [5]$$

Simulations (for a fixed set of parameters, with θ=324, μ=250, A_L=10, A_H=1000, w_0=1, and β=0.0017) show that this model is expected to produce a reduction in collective error (**Fig. 1C**). This increase in collective accuracy is accompanied by a large increasing diversity as well as a reduction in individual accuracy (**Fig. 1D**).

Using this simple model of the anchoring effect, we analytically derived the general conditions under which the proposed method leads to an increase in collective accuracy. We found that the key variable determining the success of the approach is the mid-point of the anchors, defined as $\bar{A} = \frac{A_L+A_H}{2}$. Following a simple mathematical procedure (for details, see Methods), we found that the range of $\bar{A}$ values where the method outperforms the wisdom of crowds (Δ) is

$$\Delta = \frac{4|\mu-\theta|}{w_L+w_H} \quad [6]$$

The expression derived in Eq. [6] implies that the range of values where the method outperforms the wisdom of crowds is always equal to or larger than two times the collective error. This can be shown by examining two opposite extreme scenarios. If the sensitivity β is small (i.e., when the anchoring procedure does not depend on the distance between the anchor and the truth), the range of values where the method outperforms the wisdom of crowds converges to, at very least, twice the collective error (i.e., if $\beta \to 0$, then $w_j \to w_0$, and therefore $\Delta \to \frac{2|\mu-\theta|}{w_0}$). In the opposite case, when the sensitivity β is large (i.e., when the anchoring effect is stronger as anchors become closer to the correct answer), then this method is always better than the wisdom of crowds (i.e., if $\beta \to \infty$, then $w_j \to 0$, and thus $\Delta \to \infty$).

Experiment 1: Estimation of unbounded quantities

To empirically evaluate the effectiveness of the proposed method, we performed three experiments. In Experiment 1, N=120 participants (48 female, aged 37.2 ±11.6 yr, from the USA, recruited online on Amazon Mechanical Turk) provided estimates about 14 general-knowledge quantities (**Table S3**). All variables were positive and unbounded, like the example of used in **Figure 1** (e.g., "What is the height of the Eiffel Tower?"). Participants had monetary incentives to estimate these variables as accurately as possible (for details, see **Methods**). One third of the sample was randomly assigned to a control condition where they simply estimated the variable. The other two thirds of the participants were randomly assigned to the experimental condition where, before estimating the quantity, they were asked to consider either an extremely low or extremely high value. Half of the anchored participants considered a low value, and the other half considered a high value. Crucially, these extreme values were not manually chosen by the experimenters, but set automatically as the 5 and 95 percentiles of the empirical values observed in the control condition.

By employing a simple bootstrapping resampling method, we estimated the collective error of differently-sized groups for both the wisdom of crowds and the wisdom of radicalized crowds (**Fig. 2A**). We observed that the average collective error of the latter was always smaller than the former. For example, the collective error of 34 individuals randomly taken from the control condition was substantially larger than the collective error of 34 radicalized individuals (unpaired t-test: $t(998)=29.7$, $p=2\times10^{-139}$; effect size: Cohen's $d = 1.88$). This collective error reduction was due to an increase in predictive diversity (**Fig. 2B**, unpaired t-test: $t(998)=54.8$, $p<10^{-200}$; effect size: Cohen's $d = 3.46$) that was higher than the increase in mean individual error (**Fig. 2B**, unpaired t-test: $t(998)=24.2$, $p=5\times10^{-102}$; effect size: Cohen's $d = 1.53$).

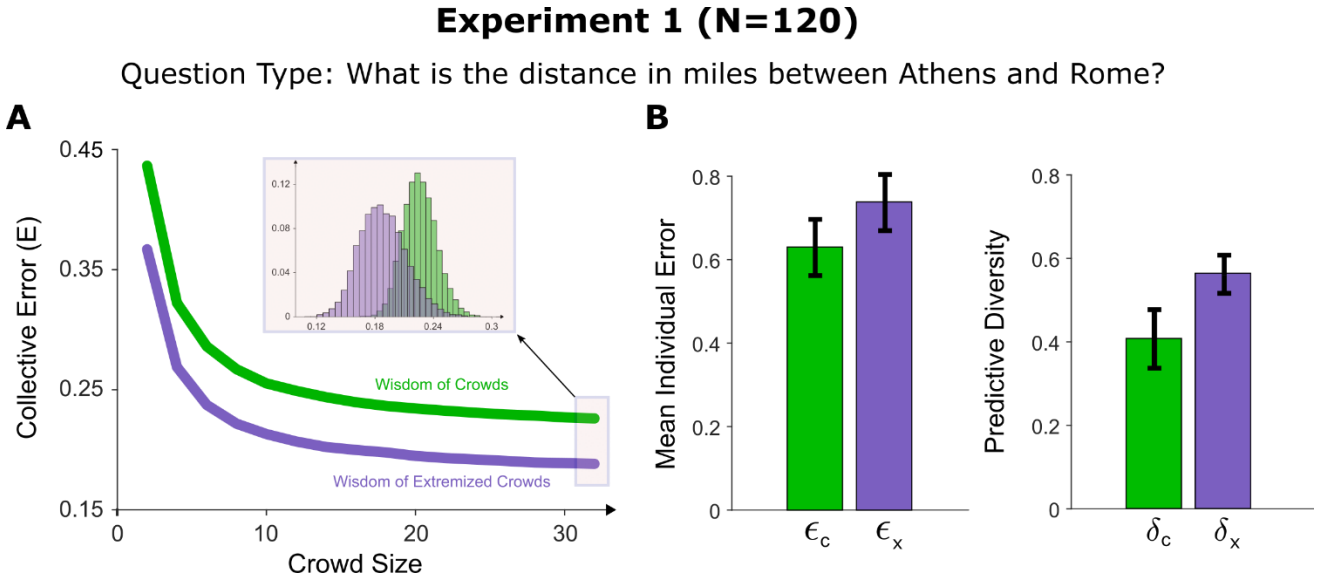

Fig. 2. Empirical results for Experiment 1. (A) Collective error as a function of crowd size, for the standard wisdom of crowds (green) and for extremized crowds (purple). The standard error of each curve is within the line width. The inset show the distribution of values from the resampling method for the largest crowd size (N=34). **(B)** Mean individual error and predictive diversity for both the standard wisdom of crowds (green) and the wisdom of extremized crowds (purple). The error bars are the standard deviation of the means.

Experiment 2: Estimation of bounded quantities with fixed anchors

One limitation of Experiment 1 is that the anchors were defined after collecting the data of the non-anchored population. Because this procedure may be inconvenient from a practical point of view, we performed a second pre-registered experiment where anchors were pre-defined and fixed across all questions (https://aspredicted.org/RYC_4Y5). In Experiment 2, we recruited N=396 participants online (235 female, aged 27.9±8.8 yr), and asked them 30 general-knowledge questions that involved the estimation of a percentage. Therefore, all answers were bounded in the range [0,100] (e.g. what percentage of the population of Argentina is under 15 years old? See Methods for details on the procedure and Table S4 for the full list of questions used in Experiment 2). Unlike the previous study, here we used the same anchors for all questions, always set at either 5% (low anchor A_L) or 95% (high anchor A_H).

We observed very similar results to Experiment 1 (**Fig. 3A and 3B**). Collective error was lower for the radicalized crowd (crowd size N=50, unpaired t-test: $t(998)=19.1$, $p=2x10^{-69}$; effect size: Cohen's $d = 1.21$). This was accompanied by an increase in mean individual error (unpaired t-test: $t(198)=121.1$, $p<10^{-200}$; effect size: Cohen's $d = 7.66$) as well as an increase in predictive diversity (unpaired t-test: $t(198)=162.5$, $p<10^{-200}$; effect size: Cohen's $d = 10.3$).

Given that this experiment used fixed anchors across all questions, it allowed us to test one key element of the model, i.e., the anchoring extremeness effect (Eq. [4]). We did so by performing two separate analyses. First, we examined the biases associated with each experimental condition. We reasoned that, if anchoring was sensitive to the distance to the correct answer, then the effectiveness of the procedure should be higher when the correct answer is close to the anchor. For example, we should see that participants considering a low value (5%) should be more attracted to the anchor when the correct answer is low (below 50%) compared to when the correct answer is high (above 50%). Consistent with this idea, when the correct answer was below 50% (17 questions), we observed that the population considering the "low anchor" provided a distribution of estimates that was similar to the correct answer (paired t-test, $t(16)=0.95$, $p=.36$, effect size: Cohen's $d = 0.23$). In turn, both the non-anchored (paired t-test, $t(16)=4.16$, $p=3x10^{-4}$, effect size: Cohen's $d = 1.11$) and the population anchored to a high value (paired t-test, $t(16)=5.99$, $p=2x10^{-5}$, effect size: Cohen's $d = 1.45$) provided a distribution of estimates that significantly overestimated the correct answer (**Fig. 3C**). Conversely, when the correct answer was above 50% (13 questions), we observed the opposite pattern: the population considering the "high anchor" provided a distribution of estimates that was similar to the correct answer (paired t-test, $t(12)=2.01$, $p=0.07$, effect size: Cohen's $d = 0.56$) and both the non-anchored population (paired t-test, $t(16)=3.90$, $p=0.002$, effect size: Cohen's $d = 1.08$) and the population anchored to a low value (paired t-test, $t(16)=4.90$, $p=4x10^{-4}$, effect size: Cohen's $d = 1.36$) provided a distribution of estimates that significantly underestimated the correct answer.

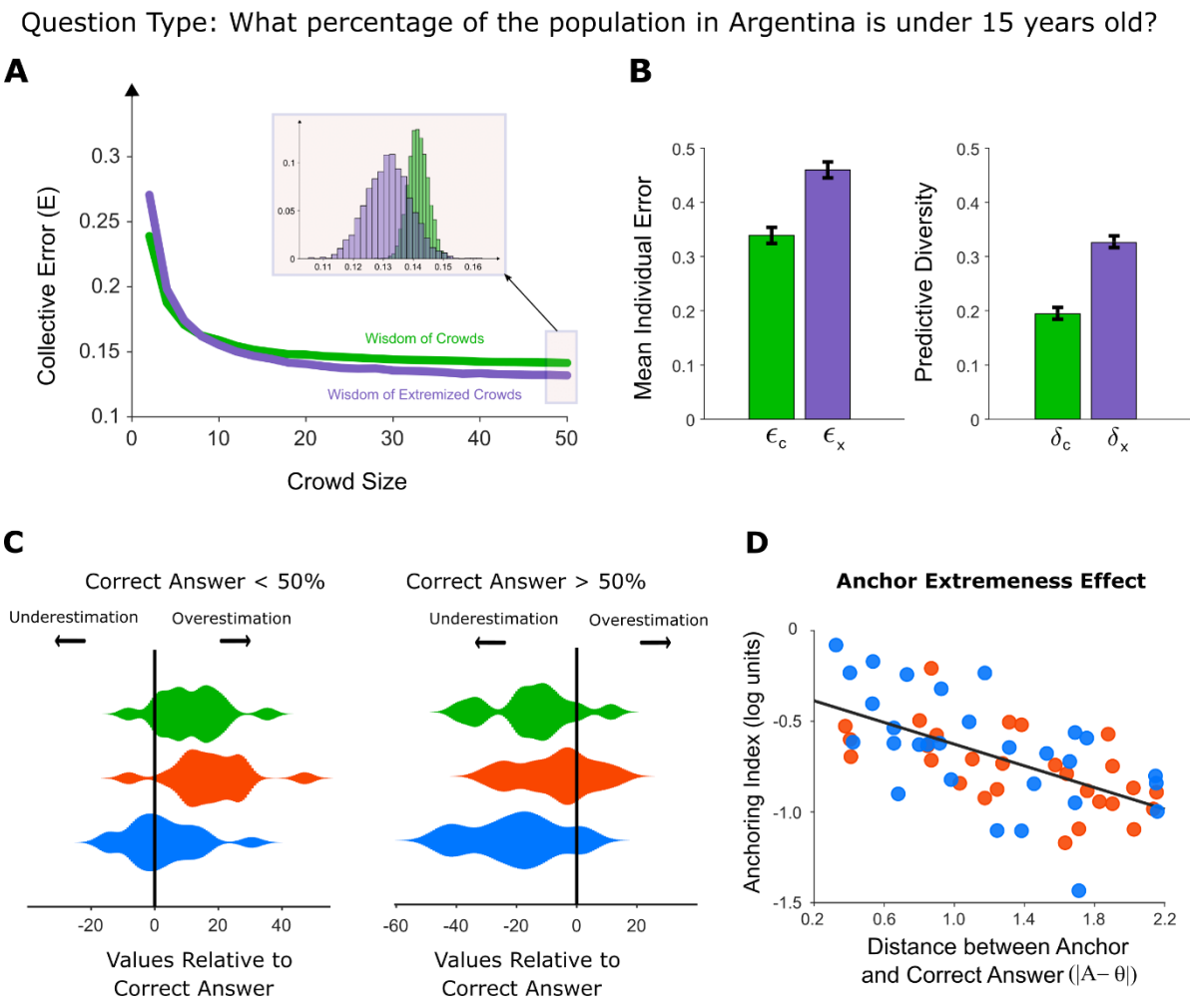

Fig. 3. Empirical results for Experiment 2. (A) Collective error as a function of crowd size, for the standard wisdom of crowds (green) and for extremized crowds (purple). The standard error of the curves is within the line width. The inset shows the distribution of values from the resampling method for the largest crowd size (N=50). **(B)** Mean individual error and predictive diversity for both the standard wisdom of crowds (green) and for extremized crowds (purple). The error bars show the standard deviation of the means. **(C)** Distributions of values corresponding to the difference between the mean answers and the correct answer for each question, for the standard wisdom of crowds (green), the crowd extremized using a high anchor (red), and the crowd extremized using a low anchor (blue). We separate the cases where the correct answer is above 50% (left panel) and where it is above 50% (right panel). The black line depicts the case where the mean value is equal to the correct answer. **(D)** Empirical anchoring index (w) for each question. Blue dots show estimates using low anchors and red dots show the same with high anchors. The horizontal axis represents the distance between the corresponding anchor and the correct answer (logarithmic units on anchoring index), and the black line shows the best linear fit of the data.

Second, we directly examined the anchoring extremeness effect by studying the association between the strength of the anchoring procedure and the distance between the anchor and the correct answer (Eq. 4). We estimated the strength of the anchoring effect by calculating the "anchoring index" as in previous literature (Jacowitz & Kahneman, 1995). The index (see Equation 7 in Methods for details) takes a value of 0 when the anchoring procedure does not produce any effect on the estimates, and a value of 1 when the estimates are on average equal to the anchor. Consistent with the existence of the anchoring extremeness effect (Roseler et al., 2022), we observed a significantly negative correlation between the anchoring index and the distance between the anchor and the correct answer (Pearson correlation coefficient, r=-0.56, $p=3x10^{-6}$). This empirical observation provides support to the proposed model of the anchoring effect used in Eqs 2-4.

Experiment 3: Forecasting

Finally, we asked whether the proposed method can be useful to outperform the wisdom of crowds for forecasting tasks, i.e., for domains where the truth is unknown and unavailable at the time of the experiment. To answer this question, we performed a third pre-registered experiment that took place in the midst of the COVID-19 crisis (https://aspredicted.org/HZC_PTH). We recruited N=620 participants (312 female, aged 46.1 ± 15.7 yr.) from the USA, and asked them to estimate the total number of COVID-19 cases and deaths that would occur in the United States in the following week (from 27 July to 2 August, 2020). Participants had monetary incentives to estimate these variables as accurately as possible. Anchors were selected as extreme values based on historical data, namely, two orders of magnitude less or more the number of COVID-19 cases and deaths reported on the two weeks before the beginning of the experiment (see **Methods** for details on the procedure and Table S4 for the questions used in Experiment 3).

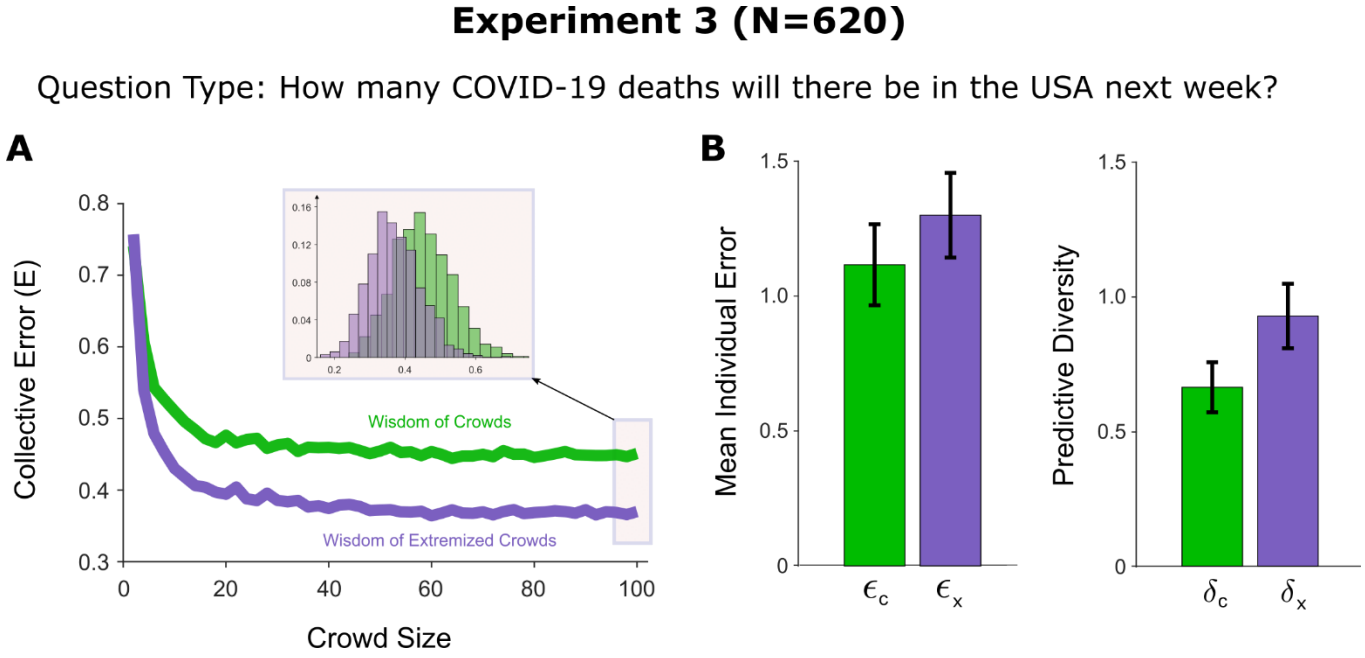

Fig. 4. Empirical results for Experiment 3. (A) Collective error as a function of crowd size, for the standard wisdom of crowds (green) and the radicalized crowds (purple). The standard error of the curves is within the line width. The inset shows the distribution of values from the resampling method for the largest crowd size (N=100). **(B)**. Mean individual error and predictive diversity for both the standard wisdom of crowds (green) and the radicalized crowd (purple). The error bars are the standard deviation of the means.

Again, as in Experiments 1 and 2, we observed that the collective error was lower for the radicalized crowd for all group sizes compared to the standard wisdom of crowds (**Fig. 4A**). We found that the decrease in collective error (specifically, for the largest crowd size N=100, unpaired t-test: $t(998)=16.4$, $p=7\text{x}10^{-54}$; effect size: Cohen's d = 1.04) was due to an increase in predictive diversity (crowd size N=100, unpaired t-test: $t(998)=37.6$, $p=7\text{x}10^{-193}$; effect size: Cohen's d = 2.38) that was greater than the increase in mean individual error (**Fig. 4B,** crowd size N=100, unpaired t-test: $t(998)=18.1$, $p=1\text{x}10^{-64}$; effect size: Cohen's d = 1.15).

Discussion

In this work, we introduce a novel strategy to outperform the wisdom of the crowds (Surowiecki, 2005) that has both practical and theoretical implications. By means of the anchoring effect (Tversky, & Kahneman, 1974), we show we can reduce collective error by increasing predictive diversity (Page, 2007). In previous literature, methods for increasing the wisdom of crowds often involved strategies for reducing individual error (Madirolas & de Polavieja, 2015; Mannes, Soll & Larrick, 2014). However, while theoretical analysis suggested this goal could also be achieved by increasing the predictive diversity (as seen in Eq. [1]), that possibility remained unexplored. Here, we thoroughly studied this approach both theoretically and empirically: first, by developing a mathematical model of the anchoring effect and, second, by performing three behavioral experiments.

In all of the experiments, regardless of differences in sample size, country of implementation, use of bounded or unbounded quantities, and whether the task involved estimation or forecasting, we observed very similar results. Extremized crowds always produced estimates with lower collective error, and this was always accompanied by an increase in both mean individual error and predictive diversity. This demonstrates that it is possible to reduce collective intelligence while concurrently reducing individual accuracy, an approach that remained heretofore empirically untested. Therefore, this should inspire future research aimed at increasing the wisdom of crowds using a similar strategy.

One limitation of the proposed method is that the selection of appropriate anchors could potentially prove hard (for example, in forecasting problems). However, our model-based analyses as well as our empirical findings suggest that the range of values where this method improves the wisdom of crowds is large. We also showed that the a priori selection of anchors is empirically feasible across three very different setups. The third experiment is especially relevant, since it shows a direct application of the proposed method in a real-world forecasting problem, where the answer was unknown at the time. Thus, appropriate anchors can

be chosen for problems with unknown answers. This, of course, requires that the order of magnitude of those answers is not completely indiscernible.

The first study on the wisdom of crowds (Galton, 1907) was regarded as an empirical demonstration that democratic aggregation principles can be reliable and efficient. This was counter-intuitive at the time, since it showed that erroneous individuals could make good collective choices. Nowadays, when political opinions tend to become extremized, these results seem to suggest that democratic decisions can still be surprisingly accurate, even if collective choices proceed from misinformed voters, as long as they are sufficiently diverse. Therefore, one interpretation of these findings is that opinion polarization, which may stem from the attraction towards political extremes (Goldenberg et al., 2023; Zimmerman et al., 2022), can potentially improve democratic judgement.

However, this uncomfortable conclusion may also be interpreted in a different way. Most problems in democratic decision-making involve conditions which are very different to the estimation or forecasting of unknown quantities, as they involve moral values and political identities (Van Bavel & Pereira, 2018; Cohen, 2023) which are absent in these minimalistic setups. Therefore, the costs of individual extremization in terms of increasing outgroup hate (Ivengar et al., 2019) and promoting political violence (Zmigrod & Goldenberg, 2021) could very well exceed the beneficial side effects of polarization, including the increase in diversity of opinion (Abramowitz, 2010). Taking these ideas into consideration, perhaps there is a more comfortable and parsimonious conclusion of this work which is that studying the epistemic value of democracy only through the lens of understanding the drivers of collective accuracy can provide an incomplete picture. Future research should refine this centenarian practice in social science (Galton, 1907), especially in times where democracy seems to be threatened by political polarization.

Methods

Participants

Participants were informed that their participation was completely voluntary, and that they could withdraw their participation at any time. All data were completely anonymous. The experimental protocol was approved by the ethics committee of Centro de Educación Médica e Investigaciones Clínicas Norberto Quirno (Buenos Aires, Argentina), protocol 435, version 5

A total of 120 participants (48 female, mean age 37.2 yr, s.d. 11.6 yr) performed Experiment 1. Participants were recruited online by using Amazon Mechanical Turk, and resided in the United States at the time of the experiment. A total of 396 participants (235 female, mean age 27.9 yr, s.d. 8.8 yr) performed Experiment 2. Participants were recruited online, and resided in Argentina at the time of the experiment. A total of 620 participants (312 female, mean age 46.1 yr, s.d. 15.7 yr) performed Experiment 3. Participants were recruited online by using Prolific (https://www.prolific.co/), and resided in the United States at the time of the experiment.

Questions

The full list of questions used across all experiments is available in **Tables S2-S4**. In Experiment 1, we selected 14 general-knowledge questions that involved the estimation of a positive unbounded quantity (e.g. how many bridges are there in Paris?). In Experiment 2, we selected 30 general-knowledge questions that involved the estimation of a percentage – and therefore, their answers were bounded in the range [0,100]. All of them came from a variety of representative surveys carried out by official entities (a complete list, with the corresponding sources, can be found in the Supplementary Information). The contents of the questions were explicitly related to Argentina, and ranged from demographic (e.g. what percentage of the population over 20 years old is either overweight or obese?) to personal perceptions of the country (e.g., what percentage of the population believes abortion is not morally acceptable?). When

necessary, we referenced the year in which the corresponding survey had taken place. In Experiment 3, we asked participants to forecast the number of new cases and number of new deaths in the USA for the week following the experiment (i.e. on the week from 27 July to 2 August, 2020). Thus, the answers were positive (unbounded), resembling a typical wisdom-of-crowds experiment, but related to quantities that were unknown at the time of the experiment.

Procedure of Experiment 1

Experiment 1 was developed using Psytoolkit (Stoet 2010; Stoet 2017). We tested fourteen general-knowledge questions in which participants had to estimate a given quantity. The population was divided into two groups. Group 1 were directly asked to estimate the answer to the 14 questions (e.g., how many bridges are there in Paris?). We collected these data before those of Group 2, so we could define anchors according to the percentiles of the distribution of answers for each of the unanchored questions. Group 2 were asked the same questions, but after an "anchoring" question (e.g., is the number of bridges in Paris higher or lower than 349?). The high anchors corresponded to the 95-percentile of the distribution of answers for Group 1, and the low anchors corresponded to the 5-percentile. The assignment to the low anchor or high anchor condition was random across questions. Overall, we collected data from 41 participants in Group 1 and 79 participants in Group 2. In all cases, the questions were randomly ordered. All participants had a maximum of 15 seconds to answer. Participants were paid a flat fee of 1.5 USD for their participation. Estimation accuracy was incentivized by rewarding a bonus payment of 0.5 USD to the top 10% most accurate respondents.

Procedure of Experiment 2

In Experiment 2, also developed using Psytoolkit (Stoet 2010; Stoet 2017), we asked 30 general-knowledge questions, which involved the estimation of a percentage. In order to reduce the length of the survey, we divided the questions into two subsets of 15 questions. Each participant was randomly assigned to one of

these set of questions. In addition, following a very similar procedure to the previous experiment, we randomly divided the sample in two groups. Participants in Group 1 were asked to directly estimate the answer to the 15 questions (e.g., What percentage of the population of the USA is under 18 years old?). Participants in Group 2 answered the same questions but after an "anchoring" question (Do you think the percentage of the population of the USA under 25 years old is above or below 95 %?). The assignment to the low anchor or high anchor condition was random across questions. Overall, we collected data from 119 participants in Group 1, and 277 participants in Group 2. In all cases, the questions were randomly ordered (within each subset of 15 questions). All participants had 20 seconds to answer the questions. Estimation was not incentivized for accuracy.

Procedure of Experiment 3

In Experiment 3, developed using Survey Monkey (https://www.surveymonkey.com/), we asked two questions related to the COVID-19 pandemic: we asked participants to forecast the number of COVID-19 deaths and cases in the week following the experiment. Anchors were selected as extreme values based on historical data, namely, two orders of magnitude less or more the number of COVID-19 cases and deaths reported on the two weeks before the beginning of the experiment. The correct answers to those questions were unknown at the time of the experiment. Since all participants answered both questions, there were a total of six conditions in this experiment. In Condition 1, we used a "low anchor" on COVID-19 deaths. We asked participants whether they thought there would be more or less than 40 new deaths in the week following the experiment, and then asked them to forecast the number of new deaths. In the next screen, participants also forecasted the number of new COVID-19 cases in the week following the experiment). Condition 2 was the same as Condition 1, but with a "high anchor" (400,000) on COVID-19 deaths. In Condition 3, we did not anchor people's expectations (which served as a control for Conditions 1 and 2). We first asked people to forecast the number of new deaths in the following week. In the following screen,

participants forecasted the number of new COVID-19 cases in the same week. Condition 4 was analogous to Condition 1, but changing the order of questions and setting an anchor on cases, instead of deaths. First, we asked participants whether they thought there would be more or less than 8,000 new cases in the following week, and then asked them to estimate the number of new cases. In the next screen, participants forecasted the number of new COVID-19 deaths in the same week. Condition 5 was the same as Condition 4, but with a "high anchor" (8,000,000) on COVID-19 cases. In Condition 6, we did not ask any anchoring questions and participants were directly asked to forecast the number of new cases and then the number of new deaths in the week following the experiment. This condition was analogous to Condition 3, but changing the order of questions, and serves as a control for Conditions 4 and 5.

We collected data from 117 participants in Condition 1, 105 participants in Condition 2, 97 participants in Condition 3, 92 participants in Condition 4, 97 participants in Condition 5, and 112 participants in Condition 6. The total sample of 600 participants, obtained through Prolific (https://www.prolific.co/), was representative of the US population in terms of age, gender and ethnicity. All participants received a flat participation fee of 1.0 USD and, to incentivize forecasting accuracy, we paid a bonus of 2.0 USD to the top 10% performers. There was no time limit to answer these questions.

Data Analysis

We discarded data from participants that completed the survey in less than three minutes, or those who failed to complete the survey. We also excluded participants with two or more exactly correct answers (which is likely to reflect cheating). These criteria were preregistered for Experiments 2. For the third experiment, given that it consisted in forecasting questions, and since every participant completed the survey, there was no need to exclude any of them. All of them answered both forecasting questions.

To compare different conditions in a properly balanced way, we developed a resampling bootstrapping strategy. For each crowd size, we randomly selected with replacement a fixed number of individuals and estimated the collective error, predictive diversity, and mean individual error for that crowd size and that iteration. We repeated that procedure 1,000 times for crowd sizes that varied from 2 to different maximum values in different experiments (i.e., n=32 in Experiment 1, n=50 in Experiment 2, and n=100 in Experiment 3). In all cases those values allow perceiving the asymptotic behavior of the collective error as a function of crowd size.

Conditions to Outperform the Wisdom of Crowds

To find the range of parameters under which the collective error of radicalized crowds is lower than the classic wisdom of crowds, the following expression should hold:

$$|\mu_r - \theta| < |\mu - \theta| \quad [7]$$

where μ_r is the mean of the radicalized crowds, μ is the mean of the wisdom of crowds, and θ is the correct answer. Assuming that $\mu \neq \theta$ (i.e., the wisdom of crowds does not reach the correct answer, meaning that there is room for improvement), we have

$$\frac{|\mu_r - \theta|}{|\mu - \theta|} < 1 \quad [8]$$

Furthermore, without loss of generality, we assume that the crowd underestimated the correct answer, and so $\mu - \theta > 0$. (The procedure is analogous in the case of overestimation). We then expand the absolute value $\left|\frac{\mu_r - \theta}{\mu - \theta}\right|$, and multiply by $\mu - \theta$ on each side, reaching

$$-(\mu - \theta) < \mu_r - \theta < \mu - \theta \quad [9]$$

We then substract $\mu - \theta$ on all sides, and reach

$$-2(\mu - \theta) < \mu_r - \mu < 0 \quad [10]$$

For the sake of clarity, we now simplify the problem and assume a constant anchoring index w_0. This particular case is informative, as a constant anchoring index would produce conditions which are less favorable for the model, given that it would lack access to the correct value θ through Eq. [4] (the proof for the general case can be found below). If we have a constant anchoring index, using Eq. [2] and [3], and replacing them in Eq. [5], we have

$$\mu_r = w_0 \, \frac{A_L + A_H}{2} + (1 - w) \, \mu \qquad [11]$$

If we define $\bar{A} = \frac{A_L + A_H}{2}$ (mean value of the anchors), and replace μ_r in Eq. [10], we get

$$-2(\mu - \theta) < w_0 \, \bar{A} + (1 - w) \, \mu - \mu < 0 \qquad [12]$$

Now we divide by w_0 and sum μ on all sides, reaching

$$-\frac{2}{w_0}(\mu - \theta) + \mu < \bar{A} < \mu \qquad [13]$$

The inequality in Eq. [13] implies that the range of values of $\bar{A}$ to outperform the wisdom of crowds is of the form:

$$\Delta_0 = \frac{2|\mu - \theta|}{w_0} \qquad [14]$$

This result indicates that the conditions under which the model improves the wisdom of crowds depend on the mid-point of the anchors (its average $\bar{A}$). It also implies that the range of values for $\bar{A}$ that fulfills those conditions is at least twice the size of its error.

If we drop the assumption of constant anchoring index, we can still follow the same procedure and reach, instead of Eq. [13], the general expression

$$-\frac{4(\mu - \theta)}{(w_L + w_H)} + \mu \; + \; \frac{(w_L - w_H)}{(w_L + w_H)} \Delta A \; < \bar{A} < \; \mu \; + \; \frac{(w_L - w_H)}{(w_L + w_H)} \Delta A \qquad [15]$$

This expression implies that the range of values where radicalized crowds outperform the wisdom of crowds is given by Eq. [6] in the main text.

Definition of "Anchoring Index"

From Eq. [2] and [3], it follows that the anchoring index (Jacowitz & Kahneman, 1995) is defined as

$$w_j = \frac{\mu_j - \mu}{A_j - \mu} \quad [16]$$

where j represents either L for the low anchor or H for the high anchor.

Data and code availability

All data and codes supporting all analyses in this paper are available at the Open Science Framework (https://osf.io/ch5qw/).

Experiment 2 was pre-registered at https://aspredicted.org/RYC_4Y5 and Experiment 3 was pre-registered at https://aspredicted.org/HZC_PTH.

References

Abramowitz, A. (2010). *The disappearing center: Engaged citizens, polarization, and American democracy*. Yale University Press.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science advances*, *7*(36), eabf4393.

Becker, J., Porter, E., & Centola, D. (2019). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22), 10717-10722.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of personality and social psychology*, 85(5), 808.

De Condorcet, N. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'impremerie royale.

Frey, V., & Van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management science*, 67(7), 4273-4286.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The journal of socio-economics*, *40*(1), 35-42.

Galton, F. (1907). Vox populi. *Nature 7*, 450–451.

Goldenberg, A., Abruzzo, J. M., Huang, Z., Schöne, J., Bailey, D., Willer, R., ... & Gross, J. J. (2023). Homophily and acrophily as drivers of political segregation. *Nature Human Behaviour*, 7(2), 219-230.

Hong, L., & Page, S. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385-16389.

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, 129-146.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161-1166.

Jayles, B., Kim, H. R., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., ... & Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, 114(47), 12620-12625.

Jönsson, M. L., Hahn, U., & Olsson, E. J. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*, 142, 191-204.

Kameda, T., Toyokawa, W., & Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*, *1*(6), 345-357.

Kao, A. B., & Couzin, I. D. (2014). Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings of the Royal Society B: Biological Sciences*, 281(1784), 20133305.

Kao, A. B., Berdahl, A. M., Hartnett, A. T., Lutz, M. J., Bak-Coleman, J. B., Ioannou, C. C., ... & Couzin, I. D. (2018). Counteracting estimation bias and social influence to improve the wisdom of crowds. *Journal of The Royal Society Interface*, 15(141), 20180130.

Karachiwalla, R., & Pinkow, F. (2021). Understanding crowdsourcing projects: A review on the key design elements of a crowdsourcing initiative. *Creativity and innovation management*, *30*(3), 563-584.

Keller, A., Gerkin, R. C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., ... & Meyer, P. (2017). Predicting human olfactory perception from chemical features of odor molecules. *Science*, *355*(6327), 820-826.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, 113(31), 8777-8782.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, 108(22), 9020-9025.

Madirolas, G., & de Polavieja, G. G. (2015). Improving collective estimations using resistance to social influence. *PLoS computational biology*, *11*(11), e1004594.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of personality and social psychology*, *107*(2), 276.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, *25*(5), 1106-1115.

Mussweiler, T., & Strack, F. (2001). Considering the impossible: Explaining the effects of implausible anchors. *Social Cognition*, *19*(2), 145-160.

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, 2(2), 126-132.

Page, S. (2007). Making the difference: Applying a logic of diversity. *Academy of Management Perspectives*, *21*(4), 6-20.

Page, S. (2008). The difference. In *The Difference*. Princeton University Press.

Ray, R. (2006). Prediction markets and the financial" wisdom of crowds". *The Journal of Behavioral Finance*, *7*(1), 2-4.

Röseler, L., Weber, L., Helgerth, K., Stich, E., Günther, M., Tegethoff, P., ... & Schütz, A. (2022). The Open Anchoring Quest Dataset: Anchored Estimates from 96 Studies on Anchoring Effects. *Journal of Open Psychology Data*, *10*(1), 16.

Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature human behaviour*, 3(4), 329-336.

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior research methods*, 42, 1096-1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24-31.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. S*cience*, *185*(4157), 1124-1131.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, 22(3), 213-224.

Wegener, D. T., Petty, R. E., Detweiler-Bedell, B. T., & Jarvis, W. B. G. (2001). Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology*, *37*(1), 62-69.

Zimmerman, F., Garbulsky, G., Ariely, D., Sigman, M., & Navajas, J. (2022). Political coherence and certainty as drivers of interpersonal liking over and above similarity. *Science Advances*, 8(6), eabk1909.

Zmigrod, L., & Goldenberg, A. (2021). Cognition and emotion in extreme political action: Individual differences and dynamic interactions. *Current Directions in Psychological Science*, 30(3), 218-227.

Supplementary Information

Experiment	N	Number of Questions	Type of Questions	Range of Answers	Country	Representative Sample	Time Limit	Incentive for Accuracy	Pre-reg
1	120	14	Estimation	[0,+inf)	USA	No	Yes	Yes	No
2	396	30	Estimation	[0,100]	Argentina	No	Yes	No	Yes
3	620	2	Forecasting	[0,+inf)	USA	Yes	No	Yes	Yes

Table S1: main design variables of Experiments 1-3.

Question	Correct Answer
What is the distance (miles) between Memphis (Tennessee, USA) and Oklahoma City (Oklahoma, USA)?	421.51
What is the distance (miles) between Milwaukee (Wisconsin, USA) and Faith (South Dakota, USA)?	714.18
What is the distance (miles) between Davenport (Iowa, USA) and Indianapolis (Indiana, USA)?	261.47
What is the distance (miles) between Paris (France) and Florence (Italy)?	550.81
What is the distance (miles) between Athens (Greece) and Rome (Italy)?	652.97
What is the height (to the tip, in yards) of the Eiffel Tower?	354.331
What is the height (to the tip, in yards) of Mount Vesuvius?	1400.92
How many floors does the Franklin Center have?	60
How many times does the word Jesus appear in the New Testament (New International Version, case insensitive)?	1273
How many times does the word Allah appear in the Qur`an (case insensitive)?	2699
How many times does the word Wisdom appear in the Qur`an (case insensitive)?	50
How many emperors did the Roman Empire have (Unified Empire only)?	71
How many heart transplantations were made in 2016 in the USA?	3191
How many bridges are there in Paris, France?	37

Table S2: Questions tested in Experiment 1 and correct answers

Question	Correct Answer
What percentage of the population of Argentina is not affiliated with any religion?	12.2
What percentage of the population of Argentina aged 20 or older is either overweight or obese?	52.0
What percentage of the members of the House of Representatives of Argentina are women?	39.0
What percentage of the population of Argentina is 14 years old or younger?	24.9
According to a representative survey conducted in 2013, what percentage of the population of Argentina believes that homosexuality is not morally acceptable?	27.0
According to a representative survey conducted in 2013, what percentage of the population of Argentina believes that abortion is not morally acceptable?	56.0
According to a representative survey conducted in 2014, what percentage of the population of Argentina believes they have good or very good health?	75.0
What percentage of male deaths between 15 and 24 years of age in Argentina were due to suicide?	15.3
What percentage of the population of Argentina aged 13 or older has a Facebook account?	60.0
According to a representative survey conducted in 2017, what percentage of the population of Argentina believes they are able to distinguish between real news and “fake news” (completely invented stories or facts)?	72.0
According with the 2010 census in Argentina, what percentage of the population between 3 and 18 years old attends an educational institution?	88.9
According to the 2010 census in Argentina, what percentage of the population lives in a housing unit without water discharge or without a toilet?	15.2
According to the 2010 census in Argentina, what percentage of women over 14 years old have never had a child?	30.3
According to the 2010 census in Argentina, what percentage of the population over 20 years old either has or is looking for a job?	70.6
According to the 2010 census in Argentina, what percentage of the population living in private homes has health coverage?	63.9

What percentage of women aged between 18 and 60 years old have jobs in Argentina?	50.2
What percentage of the Argentine population has access to the internet at home, either through a computer or mobile device?	73.0
According to a representative survey conducted in 2013, what percentage of the Argentine population believes that it is not morally acceptable for unmarried adults to have sexual relationships?	22.0
What percentage of the Argentine population lives in a housing unit that they own?	72.0
What percentage of the Argentine population owns a smartphone?	48.0
According to a representative survey from 2013, what percentage of the Argentine population believes that most people are trustworthy?	19.2
What percentage of the Argentine population believes that some vaccines can cause autism in healthy children?	10.0
In Argentina, what percentage of the total real estate wealth belongs to the richest 1% of the population?	44.0
According to the 2010 census in Argentina, what percentage of people over 20 years old have completed high school?	20.1
According to the 2010 census in Argentina, what percentage of the population over 80 years old was born in another country?	11.8
According to a representative survey conducted in 2013, what percentage of the population of Argentina is not at all interested in politics?	30.7
According to the 2010 census in Argentina, what percentage of the population living in the Santa Fe province is under 65 years old?	88.2
According to the 2010 census in Argentina, what percentage of the population is single?	33.5
According to the 2010 census in Argentina, what percentage of the population does not live in an apartment?	88.1
According to the 2010 census in Argentina, what percentage of people aged 20 to 29 both work and study?	19.2

Table S3: Questions tested in Experiment 2 and correct answers

Question	Correct Answer
How many new deaths by COVID-19 do you think there will be in the United States in the upcoming week (27 July – 2 August)?	7987
How many new cases by COVID-19 do you think there will be in the United States in the upcoming week (27 July – 2 August)?	442,417

Table S4: Questions tested in Experiment 3 and correct answers