Típo de documento: Preprint

Interactive Crowdsourcing to Fact-check Politicians

Autores: Espina Mairal, Santos; Bustos, Florencia; Navajas Joaquín; Solovey, Guillermo

Fecha de publicación: 2023

¿Cómo citar este artículo?
Mairal, S. E., Bustos, F., Solovey, G., & Navajas, J. (septiembre, 2023) *Interactive crowdsourcing to fact-check politicians*. Repositorio Digital Universidad Torcuato Di Tella. https://repositorio.utdt.edu/handle/20.500.13098/12022

Interactive Crowdsourcing to Fact-check Politicians

Santos Espina Mairal[1] Florencia Bustos[1] Guillermo Solovey[2,3] Joaquín Navajas[1,3,4]

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

[2] Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, UBA-CONICET, Buenos Aires, Argentina

[3] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina

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

Author Note

Santos Espina Mairal https://orcid.org/0000-0002-8724-0618

Guillermo Solovey https://orcid.org/0000-0003-4093-2649

Joaquín Navajas https://orcid.org/0000-0001-8765-037X

The data and code that support the findings of this study are openly available at https://osf.io/ufs5c/. Hypotheses and analyses for the second study in this work were pre-registered at: https://aspredicted.org/p6ma7.pdf. We have no known conflict of interest to disclose. This work was supported by the James McDonnell Foundation twenty-first Century Science Initiative in Understanding Human Cognition—Scholar Award (grant no. 220020334) and by a Sponsored Research Agreement between Meta and Fundación Universidad Torcuato Di Tella (grant no. INB2376941).

Correspondence concerning this article should be addressed to Santos Espina Mairal, Laboratorio de Neurociencia, Universidad Torcuato Di Tella, Av. Figueroa Alcorta 7350, C1428. E-mail: santosespinamairal@gmail.com.

Abstract

The discourse of political leaders often contains false information that can misguide the public. Fact-checking agencies around the world try to reduce the negative influence of politicians by verifying their words. However, these agencies face a problem of scalability and require innovative solutions to deal with their growing amount of work. While previous studies have shown that crowdsourcing is a promising approach to fact-check news in a scalable manner, it remains unclear whether crowdsourced judgements are useful to verify the speech of politicians. This paper fills that gap by studying the effect of social influence on the accuracy of collective judgements about the veracity of political speech. In this work, we performed two experiments (Study 1: N=180; Study 2: N=240) where participants judged the veracity of 20 politically balanced phrases. Then, they were exposed to social information from politically homogeneous or heterogeneous participants. Finally, they provided revised individual judgements. We found that only heterogeneous social influence increased the accuracy of participants compared to a control condition. Overall, our results uncover the effect of social influence on the accuracy of collective judgements about the veracity of political speech and show how interactive crowdsourcing strategies can help fact-checking agencies.

Public Significance Statement

This paper studies the effect of social influence on the accuracy of crowdsourcing strategies to fact-check statements made by politicians. We found that only exposing individuals to the judgements of people supporting the opposite political party improves their performance at the individual and collective level.

Keywords

social influence crowdsourcing fact-checking interactive political speech

Introduction

The spread of false information is a major concern in recent years, posing tangible risks to public health (Burel et al., 2020; Burki, 2019), democratic life (Frau-Meigs, 2018), and the fight against climate change (van der Linden et al., 2017). In response to this threat, a vast number of fact-checking organizations aimed at increasing the quality of information in public debates have surged around the World. Despite their enormous efforts, these agencies are very overloaded and they cannot keep pace with the amount of false information that they need to process (Burel et al., 2020). For this reason, finding novel solutions to scale and reduce the time-intensive work of fact-checking organizations has become an urgent issue in social science (Lazer et al., 2018).

On the basis of the observation that combining several layperson estimates about factual issues can outperform expert judgements, a phenomenon popularly known as “the wisdom of crowds” (Larrick et al. 2012; Surowiecki, 2005), recent research tested the reliability of crowdsourcing as a potential tool to assist fact-checkers (Allen et al., 2021; Pennycook & Rand, 2019a; Resnick et al., 2021). For example, one study found that averaging 16 laypeople ratings about the truthfulness of news articles led to more accurate judgements than the ones made by 3 qualified journalists (Resnick et al., 2021). Similarly, Allen et al. (2021) observed that a crowd of 26 lay raters predicted expert judgements with substantial accuracy, suggesting that crowdsourcing may become a promising avenue to boost fact-checking scalability.

However, one limitation of previous research is that it only focused on fake news, while false information can take different forms. For example, a very demanding endeavour of fact-checking agencies is to check politicians for false statements. Far from being a secondary activity, checking the discourse of political leaders is critical, as it is known to directly influence public behaviour. Just to give one clear example, research has found that behavioural metrics of social distancing in Brazil severely reduced right after the president inaccurately minimized

the mortality of COVID-19 (Ajzenman et al., 2020). The first goal of the present work is to test whether crowdsourcing strategies can be useful to fact-check statements made by politicians.

A main challenge in this goal is that laypeople's judgements might be subject to partisan effects. In the context of fake news, partisanship has been previously shown to be a strong predictor of people's beliefs about false information and their willingness to subsequently share it in social media (Faragó et al., 2019; Osmundsen et al., 2021; Pennycook & Rand, 2021; Pereira et al., 2018; Vegetti & Mancosu, 2020). However, given that this research was tested with fake news specifically, it is still unknown whether and how partisan biases influence the perceived accuracy of claims made by politicians. Previous studies have found that leader statements can be interpreted as party cues which may awaken tribal motives (Levy Yeyati et al., 2020), and that support for political figures can remain unchanged even after their claims have been fact-checked as false (Swire-Thompson et al., 2020). Therefore, we hypothesized that statements made by leaders from a supported party (which we here refer to as "concordant" statements) will be more likely to be classified as "true" compared to statements made by leaders from the opposite party (which we call "discordant" statements).

In principle, the above-mentioned hypothesis, if confirmed, would set a clear limitation on the applicability of crowdsourcing to fact-check politicians. For this reason, a second goal of the present work is to investigate whether partisan biases can be reduced in social interactive settings and to test whether social influence that involves diverse opinions can increase the accuracy of aggregated judgements. Previous studies examining the impact of social influence on the wisdom of crowds have also provided mixed evidence. While part of the literature has found that it sometimes reduces the diversity of opinions and degrades the accuracy of crowdsourced judgements (Frey & van de Rijt, 2021; Lorenz et al., 2011; Mavrodiev & Schweitzer, 2021), several studies reported remarkably positive effects (Becker et al., 2019; Guilbeault et al., 2018; Jayles et al., 2017; Navajas et al., 2018; Pescetelli et al., 2021). In the political domain, a large-scale observational study on Wikipedia has found evidence that articles edited by "heterogeneous" collaborators who support opposite political

parties are of higher quality compared to articles edited by “homogenous” collaborators who share the same political affiliation (Shi et al., 2019). This study suggests that interacting with politically diverse individuals may reduce partisan biases by breaking filter bubbles and promoting informational flow (Cinelli et al., 2021; Rhodes, 2021; Sunstein, 2009). This is consistent with findings relating homogeneous social settings with polarization of political stances, and heterogeneous interaction resulting in individuals becoming more accuracy-based (Klar, 2014). Although previous literature highlights how group diversity and dissent can improve decision-making (Schulz-Hardt et al., 2007), other studies have shown that the benefit of heterogeneous social influence is absent in small groups (Pescetelli et al., 2021), and yet another study has found the accuracy of the wisdom of crowds is robust to social interaction within politically homogeneous networks (Becker et al., 2019). Therefore, on the basis of these disparate findings, here we considered the two possibilities and examined the effect of social influence in both politically homogeneous and heterogeneous settings.

To summarize, in this work we empirically tested whether crowdsourced judgements about the veracity of claims made by politicians become more accurate after social influence from politically heterogeneous or homogeneous individuals. Given previous studies suggesting that analytical thinking, quantified by scores obtained in the Cognitive Reflection Test (CRT) (Frederick, 2005), plays an important role in shaping people’s ability to detect misinformation (Pennycook & Rand, 2019b), here we also sought to evaluate whether the effect of social influence persists after controlling for individual differences in CRT scores. We present results from two experiments where participants either interacted in dyads (Study 1) or received feedback from supporters of the same or opposite political party (Study 2). To anticipate our results, we found that social influence improved crowdsourced judgements to fact-check politicians but only if participants interacted with or received feedback from politically heterogenous, but not homogenous, individuals.

Study 1

Method

Transparency and openness

All data and codes necessary to reproduce our findings are publicly available at the Open Science Framework and can be accessed at https://osf.io/ufs5c/. This study's design and its analysis were not pre-registered.

Sample size was determined based on a power analysis assuming a medium-sized effect (Cohen's d=.5) and an unpaired t-test between the experimental and control conditions. The selected sample size yields over 80% power with a two-tailed 5% significance level.

Ethics

The study has been approved by the Ethics Committee of "Centro de Educación Médica e Investigaciones Clínicas" (protocol ID 435 - version 5) and was performed in line with the principles of the Declaration of Helsinki. Participants provided informed consent and were paid a flat fee of 400 Argentine Pesos (roughly 4 USD at the time) for completing the experiment. On top of that, we incentivized accuracy by providing an additional bonus of the same amount to the 10% best-performing participants at the task. These monetary compensations were informed before the experiment.

Participants

This experiment was performed in Argentina, a politically polarized country (Freira et al., 2021; Navajas et al., 2019; Zimmerman et al., 2022) which is largely under-represented in the study of misinformation. We recruited N=180 Argentinian participants (56% female, aged 26.3 ± 8.4 y.o., 22% having completed university education) through student mailing lists at Universidad Torcuato Di Tella and Universidad de Buenos Aires. Potential participants completed an online sign-up form that included several political identity questions. We only recruited participants with a defined political identity, i.e., those who **a)** reported positive affect (categorically, yes/no) for Mauricio Macri (former President of Argentina, and the main political leader of the center-right party *Juntos por el Cambio*) and negative affect for Cristina Fernández de Kirchner

(current Vice-President, former President, and the main political leader if the center-left party *Frente de Todos*), or vice versa; and simultaneously **b)** stated they would vote for the corresponding party in a hypothetical Presidential election taking place the following week. For clarity purposes, within this report we will refer to those participants who identified with Mauricio Macri and *Juntos por el Cambio* as right-wing individuals and those displaying a preference for Cristina Fernandez de Kirchner and the *Frente de Todos* party as left-wing individuals.

Design

Previous to the experiment, we randomly paired participants into dyads and created three experimental conditions. In the "heterogeneous" condition (n=30 dyads), 60 participants (30 left-wing and 30 right-wing individuals) were matched with someone supporting the opposite political party. In the "homogeneous" condition (n=30 dyads), 60 participants (30 left-wing and 30 right-wing individuals) were matched with someone supporting the same political party (15 dyads were composed by two left-wing individuals and 15 dyads by two right-wing individuals). Lastly, the remaining 60 participants (30 left-wing and 30 right-wing individuals) were assigned to a Control condition in which no dyads were formed. Participants were blind to this assignment throughout the experiment.

Statements

We selected a corpus of 20 claims (**Table S1**) made by Argentinian politicians that had already been classified as either true or false by Chequeado (https://chequeado.com/), the only fact-checking agency in Argentina affiliated with the International Fact-Checking Network (https://ifcncodeofprinciples.poynter.org/). Half of these statements were made by left-wing politicians, and the remaining half by right-wing politicians. Moreover, half were classified as true, and half were classified as false by Chequeado. The corpus consisted of 5 true left-wing phrases, 5 false left-wing phrases, 5 true right-wing phrases, and 5 false right-wing phrases. This balanced design implied that each participant, upon presentation of the whole corpus, would be exposed to the same number of concordant as well as discordant phrases, and

simultaneously to an equal number of true and false phrases. Participants were unaware of this balanced design.

Procedure

Before the experiment, participants first completed a form consisting of a series of demographic items and a CRT test (Frederick, 2005; Primi et al., 2016). The experiment consisted of a 3-stage procedure (**Figure 1**). The procedure is identical in structure to the one implemented in two previous studies looking at the effect of deliberation on the wisdom of crowds (Navajas et al., 2018), the probability of reaching consensus in polarized moral issues (Navajas et al., 2019), and the effect of diversity on herding behavior (Navajas et al., 2022). In Stage 1, participants were instructed to individually classify each one of the 20 statements as true or false, as well as to provide a measure of confidence in their reported answer, in a scale ranging from 1 (Low Confidence) to 5 (High Confidence). Each phrase was read out by the experimenter and presented for a duration of 10 seconds. The only information displayed to participants was the statement itself, the name of the politician as well as a brief description of his/her public position, and the approximate date when the statement was made. For example, the information displayed would read: *Alberto Fernández, the president of Argentina, said: "we have started the largest vaccination campaign in Argentinian history". He said this in March, 2021. ¿Is what he said true or false?* **Figure S1** shows how this information was displayed.

After completion of Stage 1, the experimenter showed participants the responses provided by each of them, introducing social influence. Therefore, in Stage 2, participants learned about the responses of another person and were given time to freely discuss their disagreements (one minute per disagreed statement). In the Control condition, we asked participants to privately classify as true or false a set of unrelated general-knowledge statements (i.e., comparing city populations, see **Table S2**) lasting approximately the same time as the one taken to complete Stage 2 in the two treatment conditions. Participants in the Control condition did not interact with each other at all. Stage 3 consisted of a re-run of Stage

1, in which participants were informed that they could revise their initial individual answers and confidence ratings. Timing and format were identical to those of Stage 1.

Data collection

Participants were invited to attend an online meeting hosted in *Zoom* (https://zoom.us/), where they were assigned to separate breakout rooms with their corresponding pair and one research assistant (experimenter). Microphones were active and cameras turned off for the complete duration of the experiment. Participants were instructed to write down their answers on paper and then type them into a pre-assigned, individual Google Sheet (https://docs.google.com/spreadsheets) file sent to them (**Figure S2**).

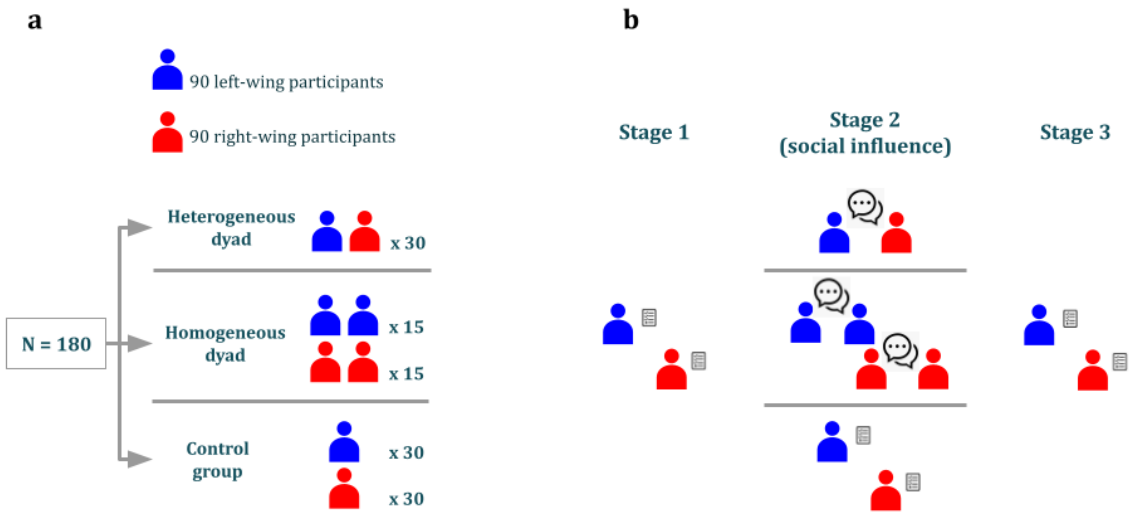

Fig. 1 | Experimental procedure. a. Participants were paired in dyads of similar or opposed partisan affinity, depending on the experimental condition. **b.** The experiment had a 3-stage structure: after individually classifying 20 statements as true or false, participants were exposed to social influence, followed by another individual stage in which they could revise their initial responses.

This manual upload of individual responses happened after Stage 1 was completed and after Stage 3 completion. The experimenter saved a backup copy of each file once participants finished typing their responses to avoid *post hoc* editing. The research assistant had access to a Google Sheet fed by both individual Sheets, enabling them to simultaneously show both members of the dyad their responses during Stage 2. A total of 10 experimental sessions were

carried out between April and May 2020. Research assistants were blind to condition throughout the experiment.

Data analysis

Reported statistical analyses were performed in Matlab R2018b. Full specification and details on the mixed-effects logit model reported in the Results section can be found in **Table S3**. Figures were generated with Python Jupyter Notebooks in Google Colab (https://colab.research.google.com/) and Matlab.

For aggregating individual judgements into collective estimates (see Results section), we randomly sampled *n* individual responses (with replacement) for a given politician statement from each experimental group's pool, where *n* thus indicates *group size*. A group-level response was constructed based on the simple majority of the individual responses: if the majority of the sampled responses for a given group classified the statement as *true* (*false*), the overall group response was *true* (*false*). This collective estimate could in turn be correct or not, upon comparison with the ground truth (assumed to be the fact-checking agency's classification). Running this procedure for each *n* across the total number of politician statements (20) yielded a score (the number of correct answers divided by the number of phrases, 20) for each crowd size. We iterated this sequence 1000 times and calculated the mean score and s.e.m. for each *n*.

Results

Initial individual responses

We initially set out to determine whether participants could discriminate between true and false statements made by politicians. For the purpose of this work, we considered a decision to be "correct" if it matched the classification of the fact-checking agency. Prior to social influence, participants classified correctly, on average, approximately 12 out of 20 statements (**Figure 2a**), leading to a better-than-chance classification performance ($d = 0.87$, $t_{179} = 11.6$, $p < 0.001$; two-tailed t-tests are employed from now, with Cohen's d as measure for effect size). There were no differences in initial accuracy between experimental groups ($\eta 2 = 0.01$, $F_{2, 177} = 0.59$,

p = 0.55; partial eta-squared is employed for effect size in ANOVA tests from now on), and performance was approximately equal across politically concordant and discordant statements (d = -0.1, t_{358} = -0.92, p = 0.36). Thus, in our study we found no evidence of participants being better when classifying concordant information, a pattern that has been documented regarding truth discernment in previous studies focused on fake news (Pennycook & Rand, 2021).

No differences between groups were found regarding CRT scores (η2 = 0.005, $F_{2, 177}$ = 0.48, p = 0.62) and education levels (η2 = 0.004, $F_{2, 177}$ = 0.35, p = 0.71), and no correlation between initial accuracy and CRT performance was found at the individual level (r = 0.03, p = 0.66).

In addition, the scope of this investigation involved studying the effect of partisanship on participants' judgements. If partisanship played a role, we would expect that people would be more likely to believe in statements made by politicians of the same political party they support compared to statements made by politicians of the opposite party. Given that each participant was exposed to the same number of concordant and discordant statements, a participant with no partisan bias should classify as "true" the same number of phrases in both conditions. However, in Stage 1, participants tended to classify politically concordant statements as "true" more frequently than discordant ones (d = 1.22, t_{179} = 16.4, p < 0.001), suggesting that partisanship predicts beliefs about the veracity of statements made by politicians (**Figure 2b**). Defining *bias* as the difference between the number of "true" answers assigned to concordant and discordant statements, we found that, on average, participants classified an excess of approximately 3 statements as "true" (M±SD = 2.99±2.45) comparing both conditions. (Notice that bias could take a maximum value of 10 in this work). We observed no pre-treatment differences in bias between experimental conditions (η2 = 0.023, $F_{2, 177}$ = 2.05, p = 0.13).

To test whether participants had introspective access into their competence as raters (**Figure 2c**), we examined confidence ratings. Individuals assigned higher confidence values in a 1 to 5 scale to correct responses than to incorrect ones (mean confidence in correct trials,

M±SD = 3.77±0.54; mean confidence in incorrect trials, M±SD = 3.36±0.61; d = 0.7, t_{179} = 11.3, p < 0.001), revealing they had -to some extent- insight into their performance, without significant pre-treatment differences displayed by experimental conditions (η2 = 0.002, $F_{2, 177}$ = 0.22, p = 0.81).

Revised individual responses following social influence

The second main goal of this study was to test the effect of social influence on the accuracy of beliefs (**Figure 3a**). Comparing revised individual answers across different treatments we found a medium-sized effect between the heterogenous and control conditions (d = 0.53, t_{118} = 2.93, p = 0.004) and no significant difference between the homogenous and control condition (d = 0.02, t_{118} = 0.09, p = 0.93).

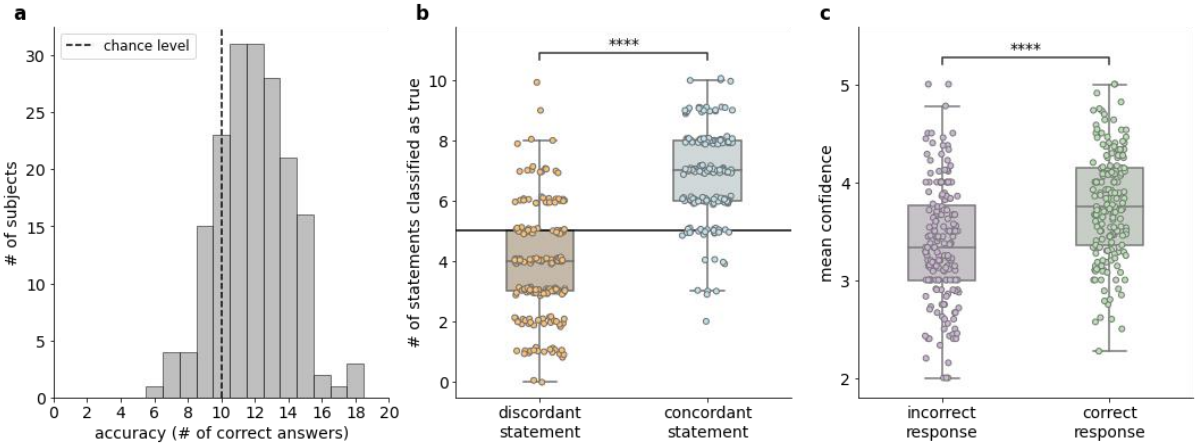

Fig. 2 | Stage 1 responses. a. Accuracy of initial individual responses. **b.** Number of statements classified as true by partisan concordance. Theoretically unbiased participants would result in no significant differences between both conditions. Each point marks the count of answers by participant for each condition. **c.** Mean declared confidence assigned to responses by their accuracy, revealing participants' introspective access. Each point represents mean confidence by participant for each condition.

We then studied the change in accuracy across different conditions. As a benchmark, individuals in the Control condition showed a small, albeit significant, increase in performance (d = 0.39, t_{59} = 3.01, p = 0.004). We found that participants in the Homogeneous condition did not significantly alter their performance compared to Stage 1 (d = 0.09, t_{59} = 0.71, p = 0.48)

and this change was statistically indistinguishable to the one observed by the individuals lacking social influence (d = 0.23, t_{118} = 1.25, p = 0.21). By contrast, Heterogeneous dyad members showed an increase in performance compared to Stage 1 (d = 0.58, t_{59} = 4.5, p < 0.001) which was significantly larger in magnitude to the one observed in participants in the Control condition (d = 0.44, t_{118} = 2.40, p = 0.018).

Although we did not find evidence for a different improvement in accuracy between members of the Homogeneous and Control conditions, we performed an equivalence test to evaluate if the difference fell within a small interval against the null hypothesis of a significant difference between groups (Lakens et al., 2018). We ran a two one-sided tests (TOST) procedure for equivalence assuming that a change in one phrase is the minimum practical difference in accuracy. This test rejected the null hypothesis of a meaningful difference between both conditions (95% CI = [−0.12, 0.85], p = 0.017).

To test the robustness of our results, we ran a multivariate analysis on the probability to provide a correct answer in Stage 3 with dummy variables coding for the two treatments with social influence (**Figure 3b**). In this mixed-effects model, we controlled for a series of variables such as age, gender, education, party affiliation, CRT performance, and whether the statement was debated or not, and added random effects at the individual level and at the phrase level. This analysis provided evidence that the positive effect of heterogeneous social influence is robust under this specification (for full model specification and results, refer to **Table S3**).

To better understand the causes of this increase in accuracy, we then studied whether social influence decreased the partisan bias observed in Stage 1. We found that Heterogeneous and Homogeneous dyad members did not differ in the number of decisions that they revised (**Figure 3c**, d = 0.21, t_{118} = 1.14, p = 0.26, comparing both conditions). However, Heterogeneous dyad members tended to revise their answers in opposition to their partisan stance resulting in a significant reduction in bias (**Figure 3d**, d = -0.74, t_{59} = -5.70, p < 0.001). Hence, by switching ideologically congruent answers for incongruent ones (i.e., revising against their partisanship), they reduced the magnitude of the initial partisan bias by

approximately 50%. Meanwhile, Homogeneous dyad members (d = 0.15, t_{59} = 1.16, p = 0.25) and individuals in the Control condition (d = -0.11, t_{59} = -0.84, p = 0.40) did not significantly change their initial bias across stages. Although social influence increased the overall likelihood to revise answers (d = 0.88, t_{118} = 4.84, p < 0.001 and d = 0.76, t_{118} = 4.20, p < 0.001 comparing Control participants to Heterogeneous and Homogeneous ones, respectively), heterogeneous communication resulted in a reduction of pre-existing partisan bias, while homogeneous interaction failed to do so.

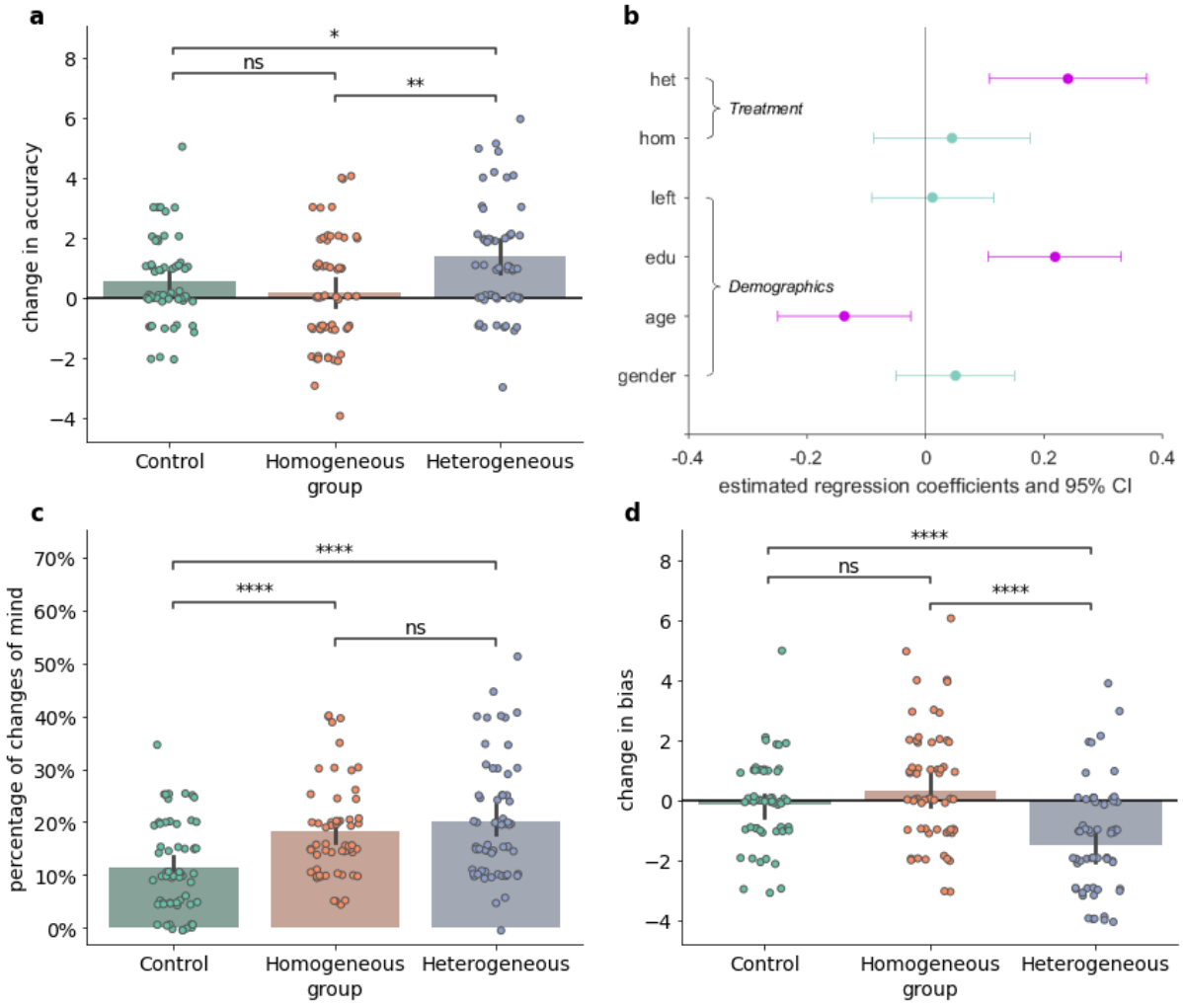

Fig. 3 | Treatment effect. a. Mean and s.e.m. of change in the number of individual correct answers between Stage 1 and Stage 3, by group. Outliers are omitted in the plots but included in the analysis. **b.** Coefficients and CIs of predictors resulting from running a GLME Logistic regression on the probability of giving a correct answer in Stage 3. Abbreviated explanatory variables on display are: *het* and *hom*, dummies that distinguish heterogeneous dyad members from homogeneous dyad members; *left*, a dummy for left-wing participants; *edu*, the number of education years attained. **c.** Mean and s.e.m. of response revisions between Stage 1 and Stage 3, by group. **d.** Mean and s.e.m. of change in individual bias between Stage 1 and Stage 3, by group. Outliers are omitted in the plots but included in the analysis.

In the results presented so far, we used the percentage of correct answers as the main indicator of performance. However, an increase in accuracy could be due to a better classification between true and false statements (truth discernment) or a change in the general tendency to believe that statements are true (overall belief). To better understand the roots of the effect of social influence on performance, we analysed our results under the signal detection theory framework (Batailler et al., 2022; Pennycook & Rand, 2021). We follow the standard calculation of truth discernment as the difference between a participant's belief in true statements (hit rate) and belief in false statements (false alarm rate). Overall belief is defined as the average belief in both true and false statements.

Our main results regarding the effect of social influence hold when examining truth discernment instead of accuracy. We found no initial differences between groups regarding truth discernment ($\eta 2 = 0.006$, $F_{2, 177} = 0.56$, $p = 0.57$) nor overall belief ($\eta 2 = 0.015$, $F_{2, 177} = 1.37$, $p = 0.26$). Post treatment, members of Heterogeneous dyads had a higher truth discernment compared to Homogeneous dyad members ($d = 0.53$, $t_{118} = 2.89$, $p = 0.005$) and Control group members ($d = 0.5$, $t_{118} = 2.71$, $p = 0.008$), whereas no differences in overall belief among groups were detected ($\eta 2 = 0.003$, $F_{2, 177} = 0.22$, $p = 0.80$). This means heterogeneous social influence increased the ability of participants to distinguish between true from false statements without changing their propensity to believe in politicians in general.

Crowd performance: aggregating individual knowledge

Our main goal involved studying crowdsourcing as a potential tool for fact-checking politicians. By aggregating individual revised responses (see Data Analysis section), we found that harnessing the wisdom of crowds after social influence has an amplifying effect on collective accuracy for all groups as crowd size increases (**Figure 4a**). In particular, Heterogeneous crowds (formed with Stage 3 answers of individuals in the Heterogeneous condition) outperformed crowds formed by individuals in the Control and Homogeneous conditions. A multivariate linear regression showed that, controlling for crowd size ($\beta = 0.313 \pm 0.037$, $t = 8.36$, $p < 0.001$), Heterogeneous crowds performed better than Control crowds ($\beta =$

0.442±0.043, t = 10.2, p < 0.001), whereas Homogeneous crowds did significantly worse (β = -0.575±0.043, t = -13.3, p < 0.001). Overall, we observed that crowds formed by individuals who interacted in the Heterogeneous condition correctly classified an average of 15 out of 20 statements, a performance which is comparable with the one obtained by previous studies on crowdsourcing and misinformation (Allen et al., 2021).

To further understand the effect of social influence on crowd accuracy we defined a variable called “crowd score change” as the increase in crowd accuracy from Stage 1 to Stage 3 for each possible group size, and then averaged this variable across all group sizes (**Figure 4b**). We found that the crowd score change was significantly larger than zero for both the Control (d = 6.41, t_{29} = 35.1, p < 0.001) and the Heterogeneous conditions (d = 5.51, t_{29} = 30.2, p < 0.001) and significantly smaller than zero for the Homogeneous condition (d = -1.73, t_{29} = 9.48, p < 0.001). The crowd score change in the Heterogeneous condition was larger than the one observed for the Control condition (d = 2.49, t_{58} = 9.66, p < 0.001), suggesting that social influence had a positive effect on crowd accuracy. Instead, crowds in the Homogeneous condition showed a significantly lower crowd score change compared to the Control condition (d = 8.2, t_{58} = 31.8, p < 0.001).

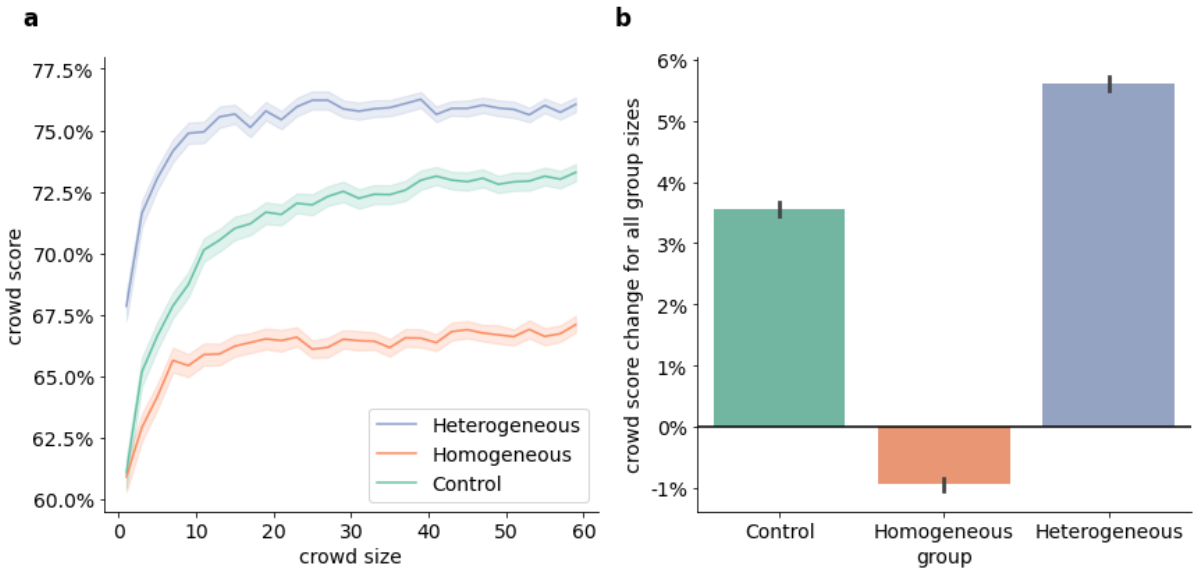

Fig. 4 | Wisdom of crowds. a. Accuracy of the crowd by group and by crowd size, using Stage 3 responses. Score is defined as the number of correct responses as a % of total statements (20). Mean (solid line) and s.e.m. (shaded area) across 1,000 iterations. **b.** Change in crowd accuracy between Stage 1 (initial answers) and Stage 3 (revised answers). Mean (bars) and s.e.m. (error bars) across 1,000 iterations and for all group sizes.

The fact that Heterogeneous crowd accuracy stabilized slightly above 75% deserves further attention. In principle this could reflect that crowds consistently misclassified some of the phrases (i.e., that the crowd correctly classified 75% of the phrases across all iterations) or, alternatively, it could mean that the crowdsourcing approach developed here produced inconsistent classifications across all phrases (i.e., that the crowd correctly classified all phrases across 75% of the iterations), or a combination of both. To study this issue, we computed the crowd score for each phrase individually (**Figure S3**). Our results provide evidence that most crowd classifications were mostly internally consistent and that in only one phrase there was some ambivalence in the crowd recommendation.

Discussion

In this study, we found individuals can distinguish true from false statements made by politicians with above-chance accuracy. We observed their performance was limited by partisan bias, which could be partially corrected by means of social interaction in dyads of opposing political affinity. In line with this observation, the accuracy of participants at fact-

checking politicians increased because of politically heterogeneous communication but failed to do so if social interaction happened in dyads of same partisanship. We also found that crowdsourcing strategies based on aggregating individual fact-checks boosted performance and that the effect of social influence on collective accuracy parallels the one found on individual accuracy.

However, one main limitation of this study points towards the scalability of this tool. Because dyadic interaction may not be cost-effective in crowdsourcing approaches, it remains to be tested if paradigms with minimal social feedback can attain similar results without depending on fact-checkers communicating with one another.

Another limitation in our study concerns the fact that the observed experimental effect of social interaction is conditional on its specific format: dyadic communication between participants was time-constrained and limited to vocal, unstructured interaction in a virtual platform. Further research is required to test whether and how our results are sensitive to the duration and modality of communication, and how they depend on the number of group members.

In addition, the amount of social influence to which each participant was exposed to in this study was unbalanced across conditions, as Heterogeneous dyads naturally tended to disagree more often than Homogeneous dyads and thus discussed relatively more between themselves. Although we found the number of phrases discussed in dyads was not a predictor of changes in bias or accuracy, balancing exposure to social information would be a key extension of this design to further uncover the mechanisms by which the structure of social influence modulates belief in politician speech.

To replicate our main results in a more scalable format involving only minimal social feedback, we ran a second, pre-registered study that also balanced the amount of heterogeneous and homogeneous social influence. This also aimed to extend our findings regarding the effect of heterogeneous social influence on fact-checking ability using a larger sample of participants.

Study 2

Method

Transparency and openness

All data and codes necessary to reproduce our findings are publicly available at the Open Science Framework and can be accessed at https://osf.io/ufs5c/. This study's hypotheses, design and analyses were pre-registered at https://aspredicted.org/p6ma7.pdf.

To determine sample size, we ran a Monte Carlo Power Analysis (Zhang, 2014) using data from Study 1. We generated synthetic datasets by sampling initial and revised responses from Study 1, with replacement. For each possible sample size from N=30 to N=450, we conducted 10,000 simulations where we replicated our main analyses. Specifically, we ran three statistical tests. The first test compared the post-interaction accuracy of the Control condition with the Heterogeneous condition. The second test performed the same comparison for the Homogenous condition against the Heterogeneous condition. The last test compared the increase in accuracy for the two conditions with social influence against the increase in accuracy for the Control condition. In all cases we performed unpaired two-tailed t-tests with a significance level of 5%. We estimated power as the fraction of times where we rejected the null hypotheses on each particular test. Based on this analysis, we decided to collect data from N=240 participants (i.e., 80 individuals per condition), yielding a statistical power of 80% for the most restrictive of our tests (Heterogeneous vs. Control), and over 95% power for the other two mentioned tests (**Figure S4**).

Ethics

This study was approved by the Ethics Committee of "Centro de Educación Médica e Investigaciones Clínicas" (protocol ID 435 - version 5) and was performed in line with the principles of the Declaration of Helsinki. Participants provided informed consent and no monetary compensations were given.

Participants

We recruited N=240 Argentinian participants (44% female, aged 35.9 ± 12.5 y.o., 37% having completed university education) with a defined political identity through student mailing lists at Universidad Torcuato Di Tella and Universidad de Buenos Aires. The criteria for defining partisanship were identical to those involved in recruitment for Study 1. No participant in Study 2 had previously been part of Study 1. As pre-registered, we also asked participants to rate their subjective affect for all politicians mentioned in our corpus of phrases, in a scale ranging from -2 (dislike) to 2 (like).

Design

Participants were randomly assigned one of three possible online forms to complete: a left-wing form, a right-wing form, or a control form, where the form type combined with the partisanship of the participant determined the experimental group he/she belonged to. Left-wing (right-wing) participants assigned to the left-wing (right-wing) form belonged to the Homogeneous condition. Contrarily, participants assigned to a form of opposing political sign, as in left-wing (right-wing) participants assigned to the right-wing (left-wing) form, belonged to the Heterogeneous condition. Control-group participants were assigned the control form. Participants were blind to this arrangement throughout the experiment.

Statements

The corpus of phrases was identical to that in Study 1, and each statement was displayed in the same format.

Procedure

Before the experiment, participants completed a CRT as in Study 1. All experimental forms had a 2-stage structure (**Figure 5a**) where the first stage was identical and involved participants classifying each of the 20 statements as true or false, as well as providing a measure of their confidence in their answer in a 1-5 scale. The information displayed about each statement was identical to the one shown in Study 1. After completion of the first stage, participants answered a simple attentional question before proceeding to the second stage.

In the second stage of the experiment, participants were given the opportunity to revise their original answers. In the two treatment conditions, they were also shown the majority answer for each statement based on the data collected in Study 1. Thus, they had a chance to change their original answers while receiving feedback in the form of "the majority answered True/False". These majority votes were drawn from the left-wing or right-wing participants recruited in Study 1. Participants in the Heterogeneous condition were shown the majority answer given by individuals with opposed political sign, and participants in the Homogeneous condition were exposed to the majority answer participants of the same partisanship. A snapshot example of how this information was displayed can be found in **Figure S5**, and the full list of majority votes for each statement given by left-wing and right-wing participants from Study 1 is listed on **Table S4**. Control-condition participants did not receive any kind of social information and simply had a re-run of the first stage.

Data collection

Data was collected using Google Forms (https://docs.google.com/forms) during the first three weeks of March 2023.

Data analysis

Reported statistical analyses were performed in Matlab R2018b, and figures were generated with Python Jupyter Notebooks in Google Colab (https://colab.research.google.com/).

Regarding aggregation of individual responses into collective ones, the sampling procedure was done exactly as described in the Materials and Method section of Study 1 (see Data analysis subsection).

Results

Initial individual responses

Seeking to replicate results from Study 1, we initially asked whether performance prior to social influence was above chance level. We found participants correctly classified an average of 12.4 statements out of 20, which means that their accuracy was significantly better than chance (d = 1.22, t_{239} = 18.9, p < 0.001). There were no differences in initial performance

between groups (η2 = 0.007, $F_{2, 237}$ = 0.82, p = 0.44), and we found participants were able to better classify discordant statements compared to concordant ones (d = 0.35, t_{478} = 3.86, p < 0.001). No differences between groups were found regarding CRT scores (η2 = 0.001, $F_{2, 237}$ = 0.82, p = 0.94) and education levels (η2 = 0.02, $F_{2, 237}$ = 1.86, p = 0.16), and we found a non-significant positive correlation between initial accuracy and CRT performance at the individual level (r = 0.11, p = 0.09).

We then studied how partisanship modulated participants' beliefs. As in Study 1, we found participants were more likely to classify politically concordant statements as "true" compared to discordant ones (d = 1.24, t_{239} = 19.2, p < 0.001), specifically by an excess of slightly more than 3 statements (M±SD = 3.38±2.73). No differences in bias were observed prior to social influence between experimental conditions (η2 = 0.013, $F_{2, 237}$ = 1.58, p = 0.21).

As pre-registered, in Study 2 we measured the subjective affect that participants had for each political figure involved in the statements. This was done to evaluate if this variable was a driver of bias. We calculated the absolute value of the difference in affect between left and right-wing politicians and normalized this variable to range from 0 to 1, calling it affective bias: values close to 1 indicate a strong difference in how a participant feels for politicians of one party with respect to the other, and values close to 0 imply affective ambivalence or lack of knowledge about the politicians in general. We found no significant differences in affective bias across experimental groups (η2 = 0.006, $F_{2, 237}$ = 0.66, p = 0.52), and that this variable was positively correlated with detection bias at the individual level (r = 0.61, p < 0.001).

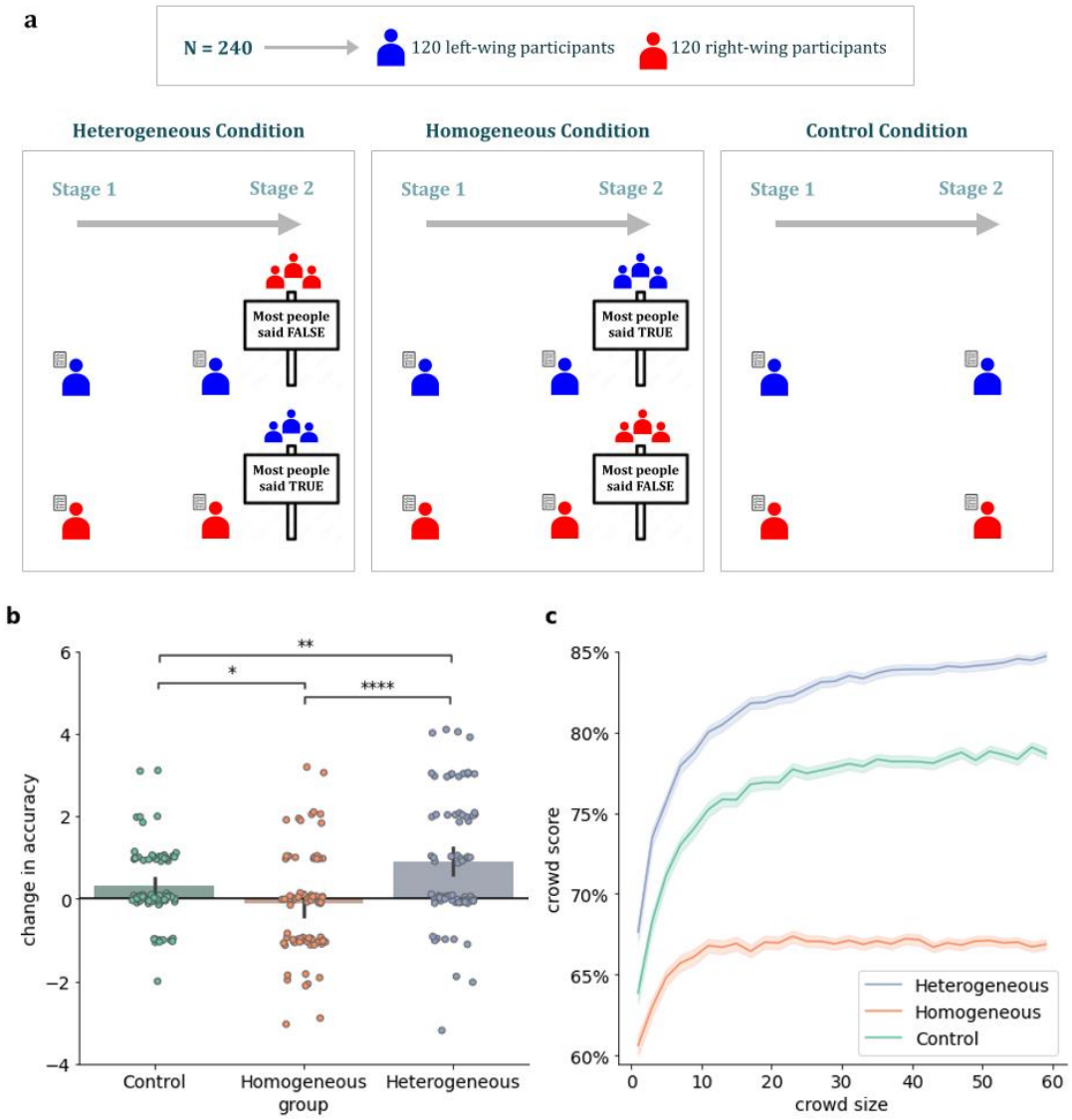

Fig. 5 | Study 2 outline and results. a. Experimental procedure for Study 2. Participants answered individually in both stages of the experiment, but for the Heterogeneous (Homogeneous) condition they were shown the majority answer from a group of participants of opposed (aligned) partisan affinity when given the opportunity to revise their initial responses. **b.** Mean and s.e.m. of change in the number of individual correct answers between Stage 1 and Stage 2, by group. Outliers are omitted in the plots but included in the analysis..**c.** Accuracy of the crowd by group and by crowd size, using Stage 2 responses. Score is defined as the number of correct responses as a % of total statements (20). Mean (solid line) and s.e.m. (shaded area) across 1,000 iterations.

As in Study 1, we found participants had introspective access into their ability as raters, as they tended to assign higher confidence ratings to correct responses than to incorrect ones ones (mean confidence in correct trials, M±SD = 3.64±0.65; mean confidence in incorrect trials, M±SD = 3.32±0.74; d = 0.46, t_{239} = 11.4, p < 0.001), with no significant differences across groups (η2 = 0.011, $F_{2, 237}$ = 1.37, p = 0.26).

Revised individual responses following social influence

The key goal of this pre-registered study was to replicate the main finding of Study 1: the distinctive effect of heterogeneous social influence on the ability of individuals to discriminate between true and false statements made by politicians. When comparing revised answers across experimental groups, we found a small-to-medium effect size in performance between the Heterogeneous and Control conditions (d = 0.35, t_{158} = 2.21, p = 0.029) and between the Control and Homogeneous conditions (d = 0.36, t_{158} = 2.3, p = 0.023).

Focusing on the change in accuracy across experimental groups (**Figure 5b**), a significant improvement was found in participants from the Heterogeneous (d = 0.58, t_{79} = 5.2, p < 0.001) and Control (d = 0.36, t_{79} = 3.18, p = 0.002) conditions, while this was not the case for participants in the Homogeneous group (d = -0.09, t_{79} = -0.81, p = 0.42). On the contrary, participants in the Control condition outperformed those in the Homogenous condition (d = 0.38, t_{158} = 2.4, p = 0.018). As per our pre-registration, we found that participants in the Heterogeneous condition had an improvement in performance that was significantly larger than that one observed in both the Control (d = 0.47, t_{158} = 2.98, p = 0.003) and Homogeneous (d = 0.7, t_{158} = 4.45, p < 0.001) conditions.

Examining the underlying mechanisms behind the improvement in accuracy, we found no differences in the number of responses revised by participants in the Heterogeneous and Homogeneous conditions (d = 0.006, t_{158} = 0.04, p = 0.97). Nonetheless, individuals in the Heterogeneous condition were more likely to revise their answers against their partisan ideology, significantly reducing their initial bias (d = 0.59, t_{79} = 5.3, p < 0.001). On the other hand, participants in the Control condition did not change their bias when given the opportunity to revise their initial answers (d = 0.1, t_{79} = 0.91, p = 0.36), and those in the Homogeneous condition incremented their bias (d = 0.42, t_{79} = 3.74, p < 0.001). Even though social feedback increased the propensity to revise initial answers (d = 0.57, t_{158} = 3.58, p < 0.001 and d = 0.54, t_{158} = 3.44, p < 0.001 comparing Control participants to Heterogeneous and Homogeneous

ones, respectively), heterogeneous social influence reduced pre-existing partisan biases, whereas homogeneous feedback actually incremented it.

As in Study 1, we performed an analysis based on signal detection theory, focusing on truth discernment and overall belief rather than simply on accuracy. We found no initial differences in truth discernment ($\eta 2 = 0.005$, $F_{2, 237} = 0.56$, $p = 0.57$) and overall belief ($\eta 2 = 0.005$, $F_{2, 237} = 0.62$, $p = 0.54$) between experimental groups. After being given the opportunity to revise their answers, participants in the Heterogeneous condition had a higher truth discernment when compared to Homogeneous individuals ($d = 0.69$, $t_{158} = 4.39$, $p < 0.001$) and Control group members ($d = 0.33$, $t_{158} = 2.06$, $p = 0.04$), although post-treatment overall belief remained similar across groups ($\eta 2 = 0.003$, $F_{2, 237} = 0.34$, $p = 0.72$). To summarize, politically heterogeneous social influence distinctively enhanced the capacity of individuals to discriminate true from false statements without affecting their overall proclivity to believe in their veracity.

Crowd performance: aggregating individual knowledge

Our last main pre-registered analyses involved replicating the findings of Study 1 about collective judgements. We found that accuracy is boosted when individual judgements were aggregated into collective estimates for all experimental groups as the size of the crowd increases (**Figure 5c**; see Data Analysis section in Study 1 for details on the aggregation procedure). Crowd judgements based on revised responses from Heterogeneous participants outperformed Homogeneous and Control crowds, as verified by running a multivariate linear regression controlling for crowd size ($\beta = 0.298 \pm 0.033$, $t = 8.99$, $p < 0.001$). We also found that Heterogeneous crowds achieved a higher accuracy than Control crowds ($\beta = 0.355 \pm 0.038$, $t = 9.27$, $p < 0.001$), while Homogeneous crowds underperformed them ($\beta = -0.673 \pm 0.038$, $t = -17.6$, $p < 0.001$). Altogether, crowds formed by individuals belonging to the Heterogeneous condition correctly classified an average of 17 out of 20 statements.

Studying the change in crowd accuracy between Stage 1 and Stage 2 (averaged for all group sizes) to understand how social influence modulates collective performance, we

found the improvement was significantly larger than zero in Control (d = 7.56, t_{29} = 41.4, p < 0.001) and Heterogeneous crowds (d = 2.88, t_{29} = 15.8, p < 0.001), and significantly lower than zero in Homogeneous crowds (d = 3.46, t_{29} = 19.0, p < 0.001). The change in crowd score for the Heterogeneous condition was larger than for the Control condition (d = 0.59, t_{58} = 2.28, p = 0.026), implying heterogeneous social feedback amplified collective accuracy. Contrarily, Homogeneous crowds registered a lower improvement in performance between Stage 1 and Stage 2 when compared to the Control crowd benchmark (d = 7.42, t_{58} = 28.7, p < 0.001).

Finally, to increase the comparability of both studies presented in this work, we replicated the exact the same analyses displayed in **Figures 2-4** and **Table S3**, which are displayed in **Figures S5-S7** and **Table S5** respectively.

Discussion

In this pre-registered study, we replicated all of our main findings of Study 1. Specifically, we verified that politically heterogeneous social influence distinctively improved the ability of individuals to discriminate true from false statements relative to a control benchmark with no social feedback, and that this effect generalizes when participants' judgements are aggregated into collective estimates that outperform individual ones. Unlike Study 1, we achieved this by means of a minimal experimental procedure with no direct social interaction where participants were simply exposed to the majority answer of another group of individuals. This hints towards potential scalable solutions to fact-check politicians with minimal social feedback paradigms.

The most striking difference in results with respect to Study 1 is that Homogeneous groups significantly increased their bias after social influence. We believe this is because, unlike Study 1, participants in Study 2 received feedback for every single phrase. In Study 1, participants discussed only those phrases that they disagreed upon, and so, on average, Homogeneous dyads discussed fewer statements than Heterogeneous dyads. In Study 2, the amount of social feedback was balanced across both experimental conditions and this could explain why we observed greater bias in politically homogeneous groups. This result is consistent with the documented group polarization effect, which predicts people's attitudes or

judgements become more extreme following group discussion with others holding similar opinions (Lord et al., 1979; Moscovici & Zavalloni, 1969; Myers & Lamm, 1976).

In this study, we were also able to measure participants' affective bias towards the political figures being evaluated, and we found this to be closely related to the bias stemming from their direct responses. By reciprocity, this result implies that one should be able to extract the underlying subjective partisanship of lay raters based on their direct responses, with no actual need to ask them about their partisan stance or their subjective affect for political figures.

General Discussion

We found that lay raters are able to detect false information enclosed in claims made by politicians with above-chance performance and that this ability can be exploited by crowdsourcing approaches. While their accuracy was partly bounded by pre-existing partisan biases, here we show that these biases could be significantly reduced through social communication in dyads with opposing political ideology or by means of politically heterogeneous minimal social feedback paradigms. We believe that these results are promising regarding the potential usefulness of interactive crowdsourcing approaches to check politicians and reduce the workload of fact-checking agencies.

We found that belief about the veracity of political discourse was significantly distorted by partisanship, extending previous results focused on factual information (Bullock et al., 2013; Guilbeault et al., 2018) and fake news (Faragó et al., 2019; Pennycook & Rand, 2021; Pereira et al., 2018; Vegetti & Mancosu, 2020) to a previously unexplored format. Studying the effect of social influence on pre-existing partisan bias, we found it to act as a partial antidote to polarization only when happening in heterogeneous settings.

In principle, one explanation for this finding is that individuals in the Heterogeneous condition were exposed to more social information differing from their initial opinions and this could have nudged them to engage in more revisions, therefore increasing their accuracy. However, our data allows ruling out this trivial explanation. While exposure to social

information differing from initial beliefs was indeed higher in the Heterogeneous condition, we found participants in both conditions to be equally likely to change their minds. Instead, the fundamental difference observed between the Heterogeneous and Homogeneous conditions was the nature of those revisions. Participants in the Homogeneous condition tended to change their mind in a direction that increased their partisan congruence, whereas participants in the Heterogeneous condition were more likely to change their mind prioritizing accuracy at the expense of political consistency (**Figure S8**). This suggests that the underlying cause behind the depolarization of pre-existing biases was not the amount of exposure to incongruent social information but rather the exposure to inter-partisan judgements per se. One candidate mechanism to explain changes in accuracy after cross-party social influence involves the straightforward act of realizing that someone else, who in these tasks is not necessarily perceived as an outgroup, may think negatively about ingroup political leaders. Future research should examine the psychological mechanisms underlying partisan depolarization when in heterogeneous settings.

Our findings are consistent with Klar (Klar, 2014) but differing from Becker et al. (Becker et al., 2019) in the sense that task accuracy may be sensitive to the partisan composition of interacting networks. We speculate that this discrepancy may be because Becker et al. (Becker et al., 2019) used a factual estimation task, whereas we asked participants to directly report whether or not political leaders were telling the truth. Relative to numerical tasks, whatever their political baggage, directly evaluating the trustworthiness of politicians might exacerbate the salience of partisanship in individuals, a hypothesis coherent with the observation that people do not seem to change their level of support for political leaders even after their claims have been fact-checked as false (Swire-Thompson et al., 2020). In this sense, fact-checking politicians might be an endeavour vulnerable to homogeneous social influence, rather than resistant to it, because the enhancement of partisanship may trump the benefits of information exchange. Further research is needed to understand to what extent these differences in task characteristics modulate the effect of homogeneous influence on individual and collective accuracy.

An unresolved limitation of this work is that individual accuracy was possibly limited by time and information constraints determined by experimental design, especially taking into account that participants of both studies only had access to the spoken phrase and the name of the politician. Enabling lay raters with additional resources could increase overall accuracy, although future research is required to test this idea. In addition, our findings stem from a sample of participants with a defined, disclosed partisan affinity. Although extending this setup to moderate participants is challenging given the strength of political polarization in Argentina (Freira et al., 2021; Levy Yeyati et al., 2020), further efforts could be made to understand whether and how partisans perform at identifying false politician statements when compared to neutral or politically uninvolved individuals.

Overall, these results suggest that people can rate the veracity of political speech with above-chance performance and reduce their partisan biases through social influence in heterogeneous settings. Moreover, these findings are a proof of concept that crowdsourcing tools are useful to reduce the workload of agencies aiming to fact-check political statements. Further research is needed to extend this setup to diverse contexts, countries and formats.

References

Ajzenman, N., Cavalcanti, T., & Da Mata, D. (2020). *More Than Words: Leaders' Speech and Risky Behavior during a Pandemic* (SSRN Scholarly Paper ID 3582908). Social Science Research Network. https://doi.org/10.2139/ssrn.3582908

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*. https://www.science.org/doi/abs/10.1126/sciadv.abf4393

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A Signal Detection Approach to Understanding the Identification of Fake News. *Perspectives on Psychological Science*, *17*(1), 78–98. https://doi.org/10.1177/1745691620986135

Becker, J., Porter, E., & Centola, D. (2019). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, *116*(22), 10717–10722. https://doi.org/10.1073/pnas.1817195116

Bullock, J. G., Gerber, A. S., Hill, S. J., & Huber, G. A. (2013). *Partisan Bias in Factual Beliefs about Politics* (Working Paper No. 19080; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w19080

Burel, G., Farrell, T., Mensio, M., Khare, P., & Alani, H. (2020). *Co-Spread of Misinformation and Fact-Checking Content during the Covid-19 Pandemic*. Proceedings of the 12th International Social Informatics Conference (SocInfo), Pisa, Italy. http://oro.open.ac.uk/71786/

Burki, T. (2019). Vaccine misinformation and social media. *The Lancet Digital Health*, *1*(6), e258–e259. https://doi.org/10.1016/S2589-7500(19)30136-0

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9). https://doi.org/10.1073/pnas.2023301118

Espina Mairal, S., Navajas, J., & Solovey, G. (2023). Interactive crowdsourcing to fact-check politicians. Retrieved from osf.io/ufs5c

Faragó, L., Kende, A., & Kreko, P. (2019). We Only Believe in News That We Doctored Ourselves: The Connection Between Partisanship and Political Fake News. *Social Psychology*, *51*, 1–14. https://doi.org/10.1027/1864-9335/a000391

Frau-Meigs, D. (2018). Societal costs of fakenews in the digital single market. *European Parliament*, 40.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Freira, L., Sartorio, M., Boruchowicz, C., Lopez Boo, F., & Navajas, J. (2021). The interplay between partisanship, forecasted COVID-19 deaths, and support for preventive policies. *Humanities and Social Sciences Communications*, *8*(1), Article 1. https://doi.org/10.1057/s41599-021-00870-2

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. https://doi.org/10.1287/mnsc.2020.3713

Guilbeault, D., Becker, J., & Centola, D. (2018). Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences*, *115*(39), 9714–9719. https://doi.org/10.1073/pnas.1722664115

Jayles, B., Kim, H., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., Sire, C., & Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, *114*(47), 12620–12625. https://doi.org/10.1073/pnas.1703695114

Klar, S. (2014). Partisanship in a Social Setting. *American Journal of Political Science*, *58*(3), 687–704. https://doi.org/10.1111/ajps.12087

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259–269. https://doi.org/10.1177/2515245918770963

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In Social judgment and decision making (pp. 227–242). Psychology Press.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Levy Yeyati, E., Moscovich, L., & Abuin, C. (2020). Leader over Policy? The Scope of Elite Influence on Policy Preferences. *Political Communication*, *37*(3), 398–422. https://doi.org/10.1080/10584609.2019.1698681

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37, 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

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. https://doi.org/10.1073/pnas.1008636108

Mavrodiev, P., & Schweitzer, F. (2021). The ambigous role of social influence on the wisdom of crowds: An analytic approach. *Physica A: Statistical Mechanics and Its Applications*, *567*, 125624. https://doi.org/10.1016/j.physa.2020.125624

Moscovici, S., & Zavalloni, M. (1969). The Group as a Polarizer of Attitudes. Journal of Personality and Social Psychology, 12, 125–135. https://doi.org/10.1037/h0027568

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. Psychological Bulletin, 83(4), 602–627. https://doi.org/10.1037/0033-2909.83.4.602

Navajas, J., Armand, O., Moran, R., Bahrami, B., & Deroy, O. (2022). Diversity of opinions promotes herding in uncertain crowds. *Royal Society Open Science*, *9*(6), 191497. https://doi.org/10.1098/rsos.191497

Navajas, J., Heduan, F., Garrido, J., González, P., Garbulsky, G., Ariely, D., & Sigman, M. (2019). Reaching Consensus in Polarized Moral Debates. *Current Biology*, *29*. https://doi.org/10.1016/j.cub.2019.10.018

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), Article 2. https://doi.org/10.1038/s41562-017-0273-4

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review*, *115*(3), 999–1015. https://doi.org/10.1017/S0003055421000290

Pennycook, G., & Rand, D. G. (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences, 116(7), 2521. https://doi.org/10.1073/pnas.1806781116

Pennycook, G., & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition, 188, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pereira, A., Harris, E. A., & Bavel, J. J. V. (2018). *Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news*. PsyArXiv. https://doi.org/10.31234/osf.io/7vc5d

Pescetelli, N., Rutherford, A., & Rahwan, I. (2021). Modularity and composite diversity affect the collective gathering of information online. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-23424-1

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453–469. https://doi.org/10.1002/bdm.1883

Resnick, P., Alfayez, A., Im, J., & Gilbert, E. (2021). Informed Crowds Can Effectively Identify Misinformation. *ArXiv:2108.07898 [Cs]*. http://arxiv.org/abs/2108.07898

Rhodes, S. C. (2021). Filter Bubbles, Echo Chambers, and Fake News: How Social Media

Conditions Individuals to Be Less Critical of Political Misinformation. *Political Communication*, *0*(0), 1–22. https://doi.org/10.1080/10584609.2021.1910887

Schulz-Hardt, S., Brodbeck, F., Mojzisch, A., Kerschreiter, R., & Frey, D. (2007). Group Decision Making in Hidden Profile Situations: Dissent as a Facilitator for Decision Quality. Journal of Personality and Social Psychology, 91, 1080–1093. https://doi.org/10.1037/0022-3514.91.6.1080

Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, *3*(4), 329–336. https://doi.org/10.1038/s41562-019-0541-6

Sunstein, C. R. (2009). *Going to Extremes: How Like Minds Unite and Divide*. Oxford University Press.

Surowiecki, J. (2005). The Wisdom of Crowds. In *Nature*.

Swire-Thompson, B., Ecker, U. K. H., Lewandowsky, S., & Berinsky, A. J. (2020). They Might Be a Liar But They're My Liar: Source Evaluation and the Prevalence of Misinformation. *Political Psychology*, *41*(1), 21–34. https://doi.org/10.1111/pops.12586

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the Public against Misinformation about Climate Change. *Global Challenges*, *1*(2), 1600008. https://doi.org/10.1002/gch2.201600008

Vegetti, F., & Mancosu, M. (2020). The Impact of Political Sophistication and Motivated Reasoning on Misinformation. *Political Communication*, *37*(5), 678–695. https://doi.org/10.1080/10584609.2020.1744778

Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: Methods and software. *Behavior Research Methods*, *46*(4), 1184–1198. https://doi.org/10.3758/s13428-013-0424-0

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. https://doi.org/10.1126/sciadv.abk1909

Supplementary Information

Table S1 | Corpus of 20 phrases pronounced by Argentinian politicians, previously checked by Chequeado as True or False. Statement contents are shown as exhibited to subjects (see **Figure S1**). Translation from Spanish to English is provided in italics.

#	***Statement***	***Check by Chequeado***
1	Alberto Fernández (Presidente de la Nación): “Hemos iniciado el mayor operativo de vacunación de la historia argentina”. Marzo, 2021. *Alberto Fernández (President of Argentina): “We have initiated the largest vaccination operative in Argentinian history”. March, 2021.*	True
2	Oscar Parrilli (Senador nacional del Frente de Todos): “Más del 70% de la deuda que hoy tiene la Argentina fue tomada durante la gestión de Macri”. Agosto, 2020. *Oscar Parrilli (national senator for Frente de Todos party): “More than 70% of Argentina’s present debt was taken during Macri’s government”. August, 2020.*	False
3	Soledad Acuña (Ministra de Educación de CABA): “Movilizamos alrededor de 700 mil personas en torno a las escuelas y sólo se contagiaron el 0,17%”. Abril, 2021. *Soledad Acuña (Health Minister in Buenos Aires City): “We mobilised about 700 thousand people around schools and only 0.17% were infected”. April, 2021.*	True
4	Alberto Fernández (Presidente de la Nación): “Cuando nosotros llegamos en diciembre nos encontramos un Banco Central lánguido, sin reservas, vacío” . Octubre, 2020. *Alberto Fernández (President of Argentina): “When we arrived in December we found a languid Central Bank, without reserves, empty” . October, 2020.*	False
5	Alfredo De Angeli (senador nacional por PRO-Entre Ríos): “En el Uruguay se despenalizó [el aborto] pero no resolvió el problema de la mortalidad materna. Siguen con ese mismo problema”. Diciembre, 2020. *Alfredo De Angeli (national senator for Juntos por el Cambio party): “In Uruguay [abortion] was decriminalized but it did not solve the problem of maternal mortality. They still have the same problem." December, 2020.*	False
6	Alberto Fernández (Presidente de la Nación): “Mejoramos la situación fiscal”. Marzo, 2020. *Alberto Fernández (President of Argentina): “We have improved the fiscal situation”. March, 2020.*	False

7	Diego Santilli (Vicejefe de Gobierno de la Ciudad Autónoma de Buenos Aires): “Buenos Aires es la tercera ciudad después de Ottawa y La Paz con menor homicidio en toda América”. Enero, 2020. *Diego Santilli (Deputy Head of Government in Buenos Aires City): "Buenos Aires is the third city after Ottawa and La Paz with the lowest homicide in all of America." January, 2020.*	True
8	Horacio Rodríguez Larreta (Jefe de Gobierno de la Ciudad de Buenos Aires): “El año pasado, la cantidad de chicos que no alcanzó los contenidos mínimos fue el doble que en años anteriores”. Febrero, 2021. *Horacio Rodríguez Larreta (Head of Government of Buenos Aires City): "Last year, the number of children who did not meet the minimum content was double that of previous years." February, 2021.*	True
9	Alfonso Prat Gay (ex Ministro de Hacienda y Finanzas de la Nación por Cambiemos): “En 9 meses los precios de los alimentos ya aumentaron más que durante todo el 2016”. Noviembre, 2020. *Alfonso Prat Gay (former Minister of Treasury and Finance of the Nation for Cambiemos party): “In 9 months, food prices have already increased more than during all of 2016”. November, 2020.*	False
10	Patricia Bullrich (presidenta del PRO): “El préstamo mayor de plata que tuvo Vicentin fue en la época del kirchnerismo, con más de US$ 200 millones”. Junio, 2020. *Patricia Bullrich (president of PRO party): "The largest loan that Vicentin had was at the time of Kirchnerism, with more than US$ 200 million." June, 2020.*	False
11	Mario Negri (Diputado de la Nación, presidente del interbloque de Juntos por el Cambio), sobre la jubilación mínima: “Se decretó una minúscula alza para llegar a $19.035. Por la fórmula de Cambiemos, correspondía $19.995”. Noviembre, 2020. *Mario Negri (national deputy, president of the Juntos por el Cambio interblock), on minimum retirement: “A tiny increase was decreed to reach $19,035. According to the Cambiemos formula, $19,995 would have corresponded”. November, 2020.*	True
12	Alfonso Prat Gay (ex Ministro de Hacienda y Finanzas de la Nación por Cambiemos): “En el primer mes y medio de la cuarentena la actividad económica cayó más que en los 4 años que marcaron (hasta hoy) la peor recesión de la historia, al final de la convertibilidad”. Julio, 2020. *Alfonso Prat Gay (former Minister of Treasury and Finance of the Nation for Cambiemos party): “In the first month and a half of the quarantine, economic activity fell more than in the 4 years that marked (until today) the worst recession in history, at the end of the Convertibility plan. July, 2020.*	True

13	Esteban Bullrich (senador nacional de Juntos por el Cambio): “Claramente hubo en las PASO un fraude muy, muy grande”. Septiembre, 2020. *Esteban Bullrich (national senator for Juntos por el Cambio party): "Clearly there was a very, very big fraud in the PASO elections." September, 2020.*	False
14	Nicolás Trotta (Ministro de Educación de la Nación): “Sufrimos 4 años de fuerte caída de la inversión educativa”. Diciembre, 2020. *Nicolás Trotta (Education Minister of the Nation): "We suffered 4 years of sharp drop in investment on education." December, 2020.*	True
15	Alberto Fernández (Presidente de la Nación) dijo que respeta el distanciamiento con la gente y que “todas” las selfies son “a un metro y medio”. Junio, 2020. *Alberto Fernández (President of Argentina) said that he respects social distancing and that “all” the selfies are “one and a half metres away”. June, 2020.*	False
16	Mauricio Macri (ex Presidente de la Nación): “El 11 de agosto, cuando terminó mi gobierno económico, estábamos en el mismo nivel de pobreza que habíamos heredado”. Octubre, 2020. *Mauricio Macri (former President of Argentina): "On August 11, when my economic government ended, we had the same level of poverty that we had inherited." October, 2020.*	False
17	Alberto Fernández (Presidente de la Nación): “El Estado ha asistido a 21 millones de argentinos de los 45 millones que somos”. Julio, 2020. *Alberto Fernández (President of Argentina): "The State has assisted 21 million Argentines of the 45 million that we are." July, 2020.*	True
18	Cecilia Todesca (Vicejefa de Gabinete de Ministros): “Los salarios de la administración pública cayeron durante el gobierno de Macri un 40%”. Septiembre, 2020. *Cecilia Todesca (Vice-Chief of the Ministers’ Cabinet): "Public administration wages fell by 40% during the Macri government." September, 2020.*	True
19	Axel Kicillof (Gobernador de la Provincia de Buenos Aires): “Hay una superpoblación del 100% en las cárceles”. Marzo, 2020. *Axel Kicillof (Governor of the Province of Buenos Aires): "There is 100% overpopulation in prisons." March, 2020.*	True
20	Santiago Cafiero (Jefe de Gabinete de Ministros): “Hace por lo menos 4 años que los jubilados no le ganaban a la inflación”. Julio, 2020. *Santiago Cafiero (Chief of the Ministers’ Cabinet): "It has been at least 4 years since retirement payments beat inflation." July, 2020.*	False

Figure S1 | Snapshot of a slide from the experiment, showing how statement information was exhibited to subjects.

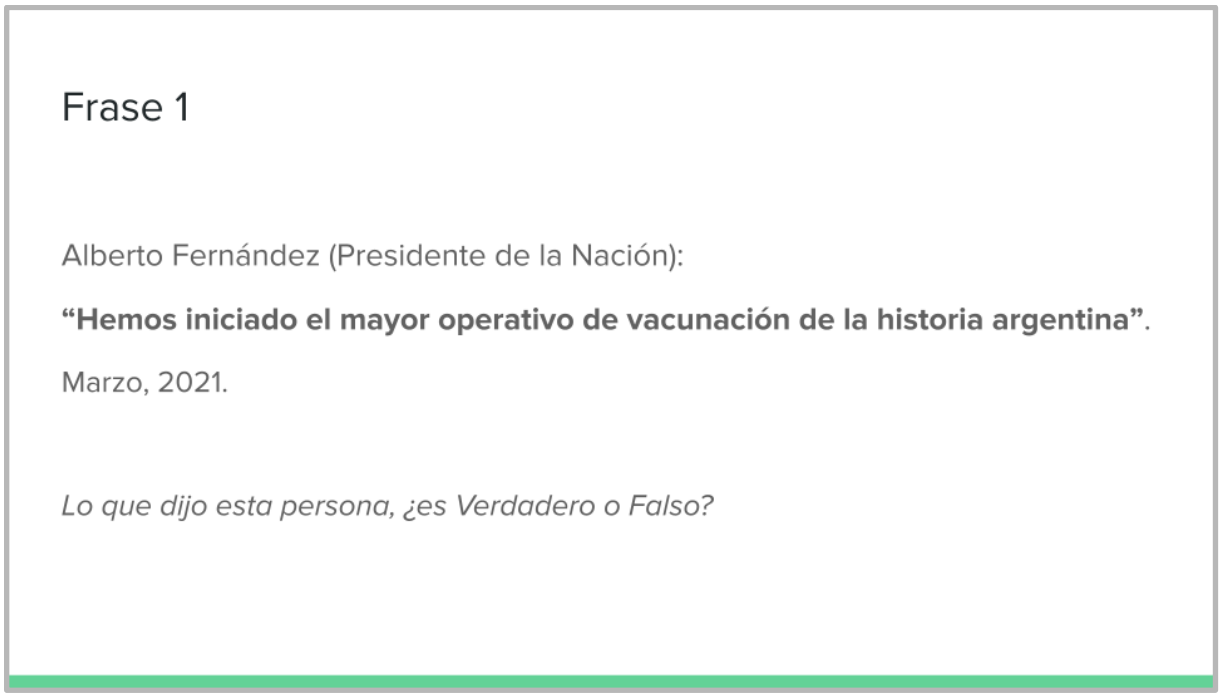

Table S2 | Corpus of true or false statements used for Control group subjects in Stage 2. Translation from Spanish to English is depicted in italics.

1	Estambul tiene más habitantes que Moscú. *(Istanbul has more inhabitants than Moscow).*
2	San Pablo tiene más habitantes que Dehli. *(São Paulo has more inhabitants than Delhi).*
3	México DF tiene más habitantes que Londres. *(Mexico City has more inhabitants than London).*
4	Hong Kong tiene más habitantes que Bogotá. *(Hong Kong has more inhabitants than Bogotá).*
5	Fortaleza tiene más habitantes que Brasilia. *(Fortaleza has more inhabitants than Brasilia).*

Figure S2 | Snapshot of model spreadsheet sent to subjects for completing during the experiment. True-false answers and confidence in each answer had to be provided for each of the 20 statements in Stage 1 and Stage 3.

Nombre y Apellido:					
Parte 1			**Parte 3**		
Frase	Respuesta (indique únicamente V o F)	Del 1 al 5, ¿cuánta seguridad tiene en su respuesta? Donde 1 es "no estoy nada seguro" y 5 es "estoy totalmente seguro".	Frase	Respuesta (indique únicamente V o F)	Del 1 al 5, ¿cuánta seguridad tiene en su respuesta? Donde 1 es "no estoy nada seguro" y 5 es "estoy totalmente seguro".
1			1		
2			2		
3			3		
4			4		
5			5		
6			6		
7			7		
8			8		
9			9		
10			10		
11			11		
12			12		
13			13		
14			14		
15			15		
16			16		
17			17		
18			18		
19			19		
20			20		

Table S3 | Mixed-effects Logit model specification and results at the answer level for Study 1, where the endogenous variable was the probability of giving a correct response in Stage 3 (*p_correct_3*). Explanatory variables introduced as fixed effects were a dummy for being correct on the initial response (*correct_1*), group dummies (*het, hom*), a dummy that indicated if the statement had been subject to dialogue in Stage 2 (*dialogue*), an interaction term between being part of a Heterogeneous dyad and having established dialogue on the statement (*dialhet*), a dummy coding the ideological stance of the responder (*left*), the subject's years of education (*edu*) and age in years (*age*), a dummy coding for gender (*gender*), the declared confidence assigned to each particular revised response (*confidence_3*), the CRT score of the subject (*crtscore*), as well a dummy signaling if the statement was aligned with subject's political stance (*concordant_phrase*). All fixed-effects explanatory variables were z-scored previous to model fit. Random effects were included in the model for controlling for phrase (*phrase*) and subject (*subject*) effects.

$$logit(p\_correct\_3) = \beta_0 + \beta_1\, correct\_1 + \beta_2\, het + \beta_3\, hom + \\ + \beta_4\, dialogue + \beta_5\, dialhet + \beta_6\, confidence\_3 + \beta_7\, concordant\_phrase + \\ + \beta_8\, left + \beta_9\, edu + \beta_{10}\, age + \beta_{11}\, gender + \beta_{12}\, crtscore + \\ + u_1\, phrase + u_2\, subject + \varepsilon$$

Model outline	
Endogenous variable	correct_3
Number of observations	3600
Fixed effects coefficients	13
Random effects coefficients	200
Model fit statistics	
AIC	18125
BIC	18218
Log-likelihood	-9047.7
Deviance	18095

Fixed effects coefficients		
variable	**coefficient**	**t-statistic**
intercept	0.873***	5.43
correct_1	1.49***	30.9
het	0.24***	3.56
hom	0.045	0.659
dialogue	-0.108	-1.87
dialhet	0.054	1.05
left	0.012	0.222
edu	0.218***	3.82
age	-0.137*	-2.38
gender	0.051	0.989
confidence_3	0.122*	2.46
crtscore	-0.080	-1.49
concordant_phrase	0.017	0.354

*** p<0.001, ** p<0.01, * p<0.05

Figure S3 | Taking the score to which crowds converge by phrase reveals that they can be systematically biased regarding some statements. That is, they may converge to a different answer than that given by the fact-checking agency. This was constructed using revised (Stage 3) responses, taking the score by phrase to which crowds converge at a size of 59 subjects, averaging across 1000 iterations.

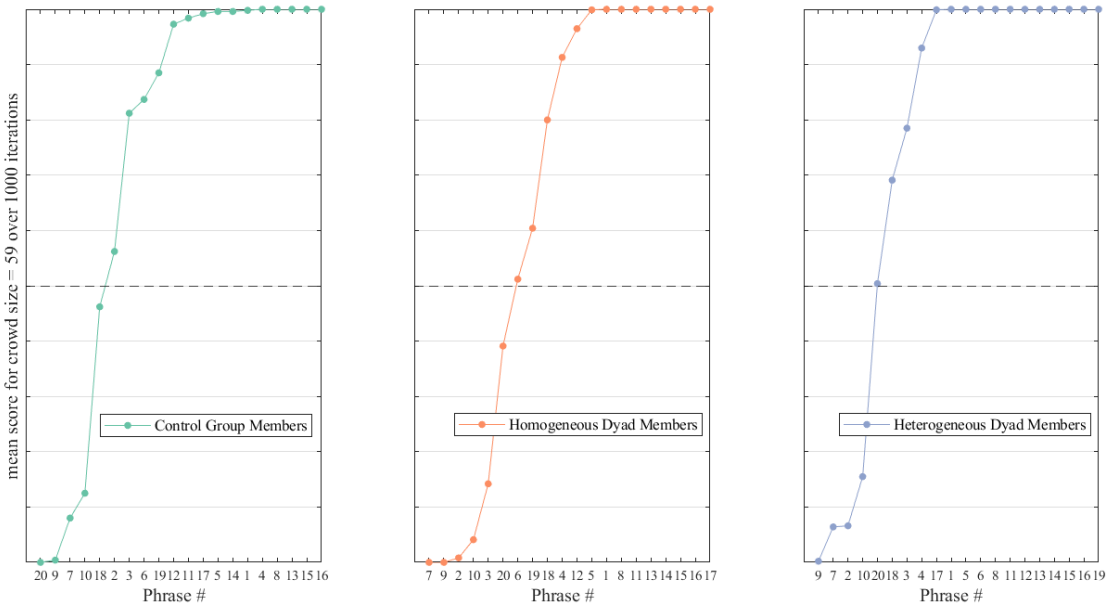

Figure S4 | Results of the Monte Carlo Power Analysis using data from Study 1. Synthetic datasets were generated by sampling (with replacement) initial and revised responses from Study 1. For each possible sample size ranging from N=30 to N=150, we conducted our main analyses on 10,000 different simulations (comparing the change in accuracy between groups) and calculated Power as the fraction of times where the null hypothesis was rejected for each test.

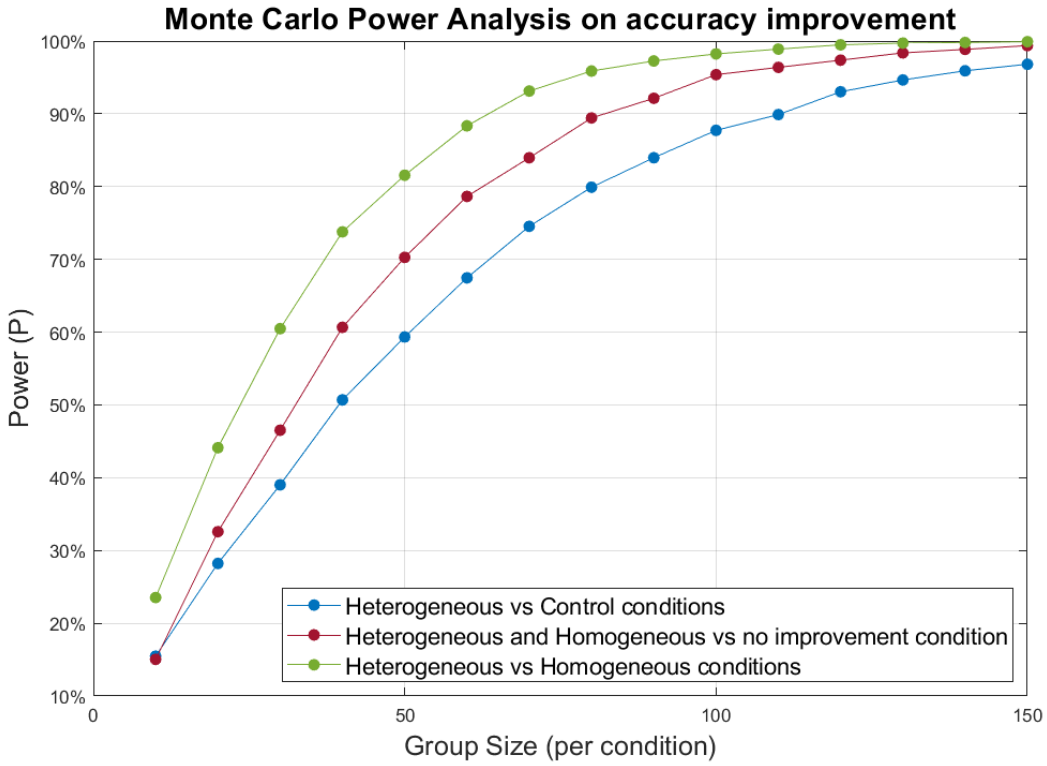

Figure S5 | Snapshot of a slide from the form, showing how statement information was exhibited to subjects at Stage 2 when given social feedback. This example corresponds to the right-wing form.

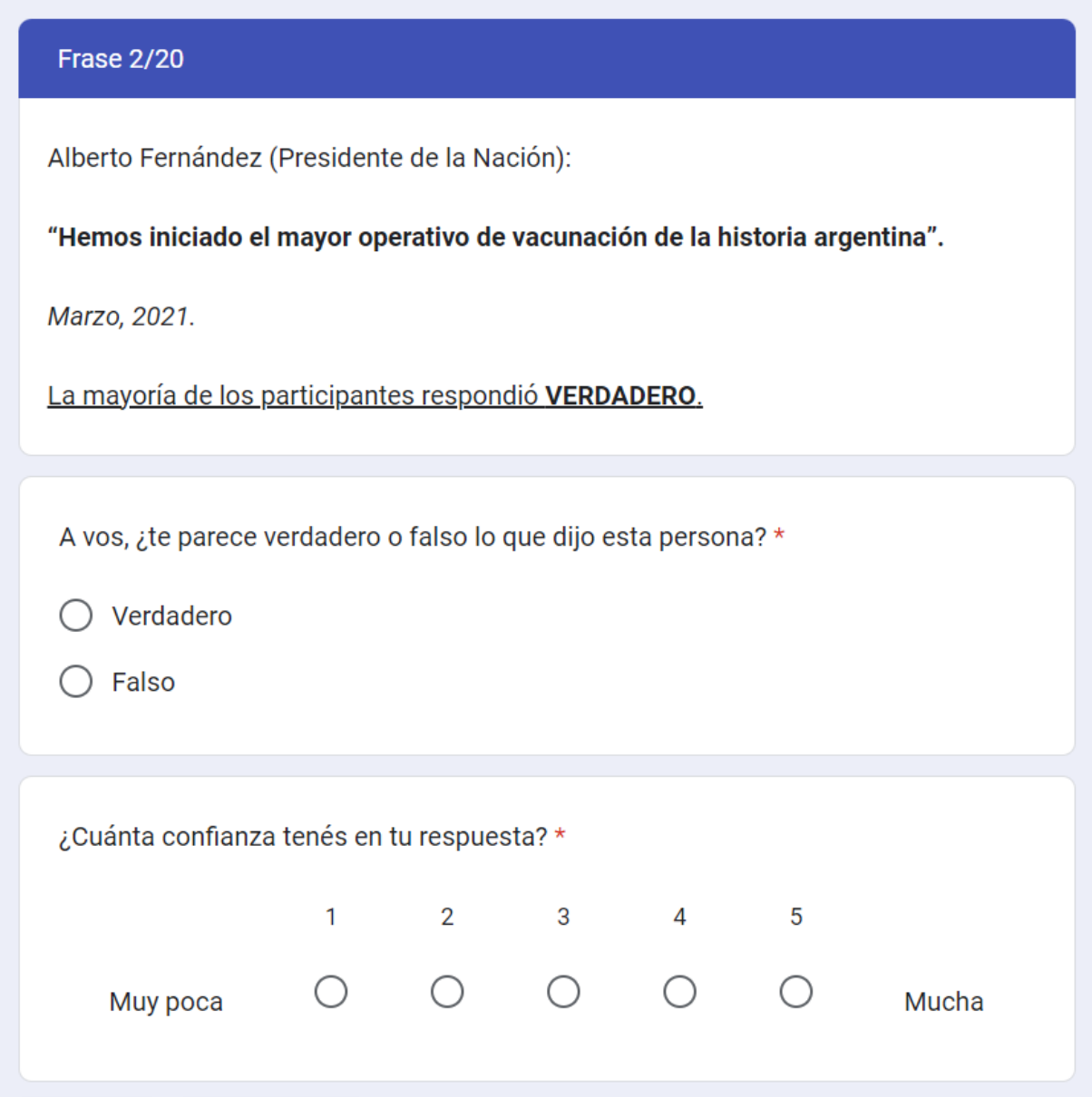

Table S4 | Majority votes from Study 1 left and right-wing participants, by phrase. Statement numbers refer to **Table S1** numbering. This information was displayed in each form type: the right-wing form contained right-wing majority votes as social feedback, and the left-wing form displayed left-wing majority votes as feedback to participants. The Control form did not display any kind of feedback.

Statement #	Left-wing majority vote	Right-wing majority vote
1	TRUE	TRUE
2	TRUE	FALSE
3	FALSE	TRUE
4	TRUE	FALSE
5	FALSE	FALSE
6	TRUE	FALSE
7	FALSE	FALSE
8	TRUE	TRUE
9	TRUE	TRUE
10	FALSE	TRUE
11	TRUE	TRUE
12	FALSE	TRUE
13	FALSE	FALSE
14	TRUE	FALSE
15	FALSE	FALSE
16	FALSE	FALSE
17	TRUE	TRUE
18	TRUE	FALSE
19	TRUE	FALSE
20	TRUE	TRUE

Figure S5 | Replica of Figure 2 using data from Study 2, illustrating pre-treatment performance, bias and introspection. **a.** Accuracy of initial individual responses. **b.** Number of statements classified as true by partisan concordance. Theoretically unbiased participants would result in no significant differences between both conditions. Each point marks the count of answers by participant for each condition. **c.** Mean declared confidence assigned to responses by their accuracy, revealing participants' introspective access. Each point represents mean confidence by participant for each condition.

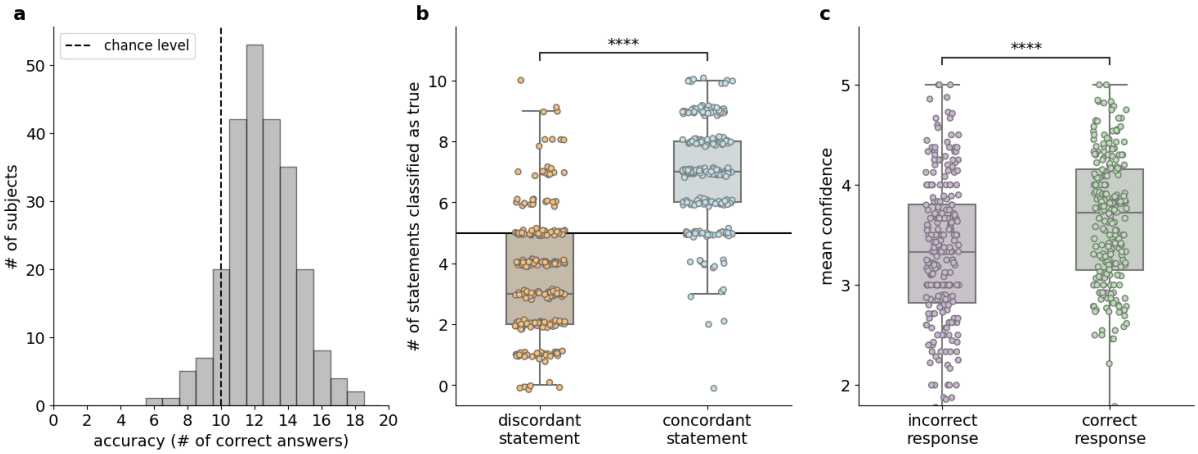

Figure S6 | Replica of Figure 3 using data from Study 2, depicting treatment effect. **a.** Mean and s.e.m. of change in the number of individual correct answers between Stage 1 and Stage 2, by group. **b.** Coefficients and CIs of predictors resulting from running a GLME Logistic regression on the probability of giving a correct answer in Stage 2. Abbreviated explanatory variables on display are: *het* and *hom*, dummies that distinguish heterogeneous dyad members from homogeneous dyad members; *left*, a dummy for left-wing participants; *edu*, the number of education years attained. **c.** Mean and s.e.m. of response revisions between Stage 1 and Stage 2, by group. **d.** Mean and s.e.m. of change in individual bias between Stage 1 and Stage 2, by group.

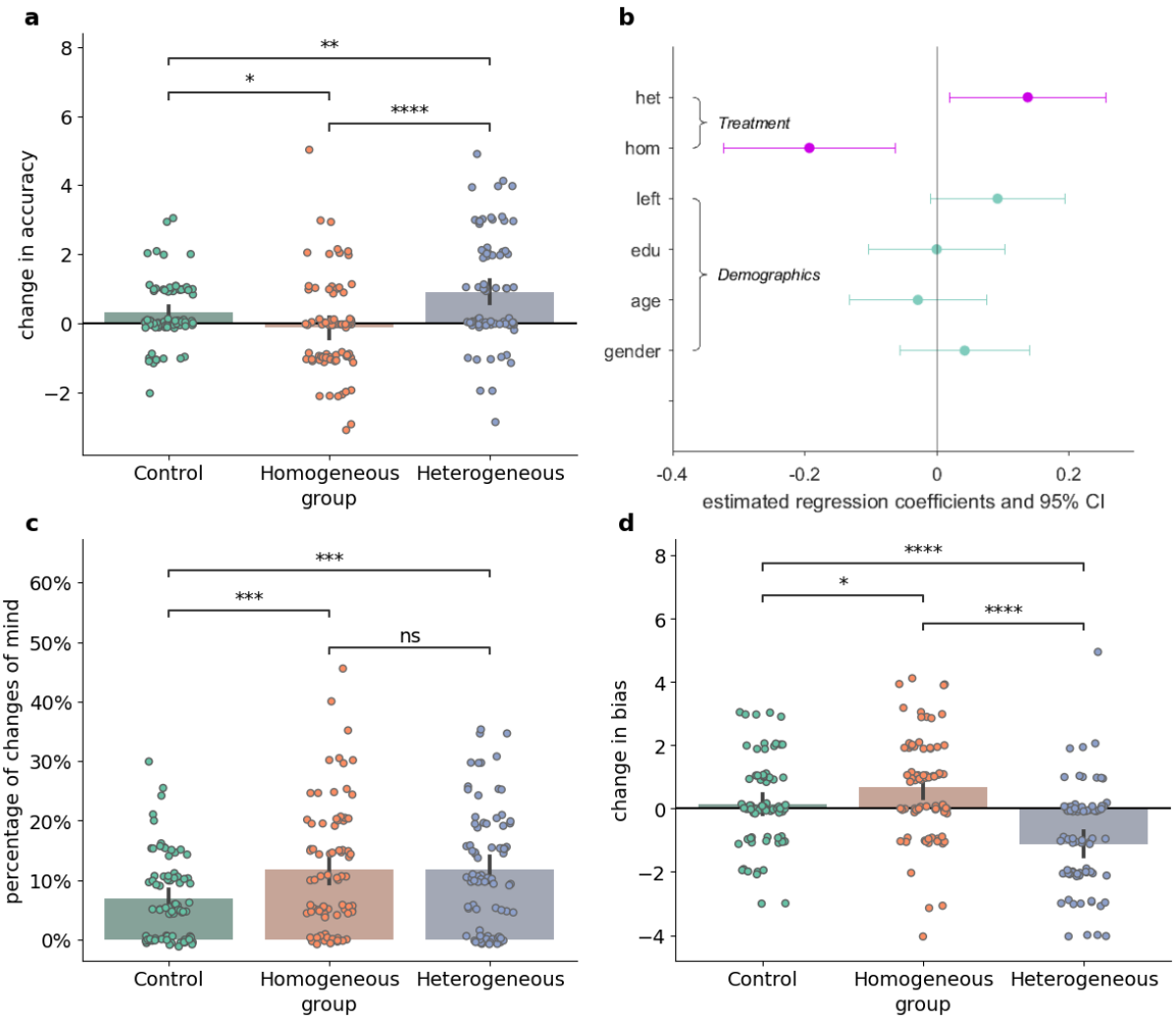

Table S5 | Mixed-effects Logit model specification and results at the answer level for Study 2, where the endogenous variable was the probability of giving a correct response in Stage 2 (*p_correct_2*). Explanatory variables introduced as fixed effects were a dummy for being correct on the initial response (*correct_1*), group dummies (*het, hom*), a dummy that indicated if social feedback had been concordant with the participant's initial answer (*feedcon*), an interaction term between being part of a Heterogeneous dyad and having established dialogue on the statement (*dialhet*), a dummy coding the ideological stance of the responder (*left*), the subject's years of education (*edu*) and age in years (*age*), a dummy coding for gender (*gender*), the declared confidence assigned to each particular revised response (*confidence_2*), the CRT score of the subject (*crtscore*), as well a dummy signaling if the statement was aligned with subject's political stance (*concordant_phrase*). All fixed-effects explanatory variables were z-scored previous to model fit. Random effects were included in the model for controlling for phrase (*phrase*) and subject (*subject*) effects.

$$logit(p\_correct\_2) = \beta_0 + \beta_1\, correct\_1 + \beta_2\, het + \beta_3\, hom + \\ + \beta_4\, feedcon + \beta_5\, confidence\_2 + \beta_6\, concordant\_phrase + \\ + \beta_7\, left + \beta_8\, edu + \beta_9\, age + \beta_{10}\, gender + \beta_{11}\, crtscore + \\ + u_1\, phrase + u_2\, subject + \varepsilon$$

Model outline	
Endogenous variable	correct_2
Number of observations	4800
Fixed effects coefficients	12
Random effects coefficients	260
Model fit statistics	
AIC	26214
BIC	26305
Log-likelihood	-13093
Deviance	26186

Fixed effects coefficients		
variable	**coefficient**	**t-statistic**
intercept	1.08***	6.79
correct_1	2.06***	40.6
het	0.138*	2.27
hom	-0.193**	-2.91
feedcon	0.099	1.52
left	0.092	1.77
edu	-0.0004	-0.007
age	-0.029	-0.54
gender	0.042	0.84
confidence_2	0.019	0.37
crtscore	-0.022	-0.43
concordant_phrase	-0.11*	-2.19

*** p<0.001, ** p<0.01, * p<0.05

Figure S7 | Replica of Figure 4 using data from Study 2. **a.** Accuracy of the crowd by group and by crowd size, using Stage 2 responses. Score is defined as the number of correct responses as a % of total statements (20). Mean (solid line) and s.e.m. (shaded area) across 1,000 iterations. **b.** Change in crowd accuracy between Stage 1 (initial answers) and Stage 2 (revised answers). Mean (bars) and s.e.m. (error bars) across 1,000 iterations and for all group sizes.

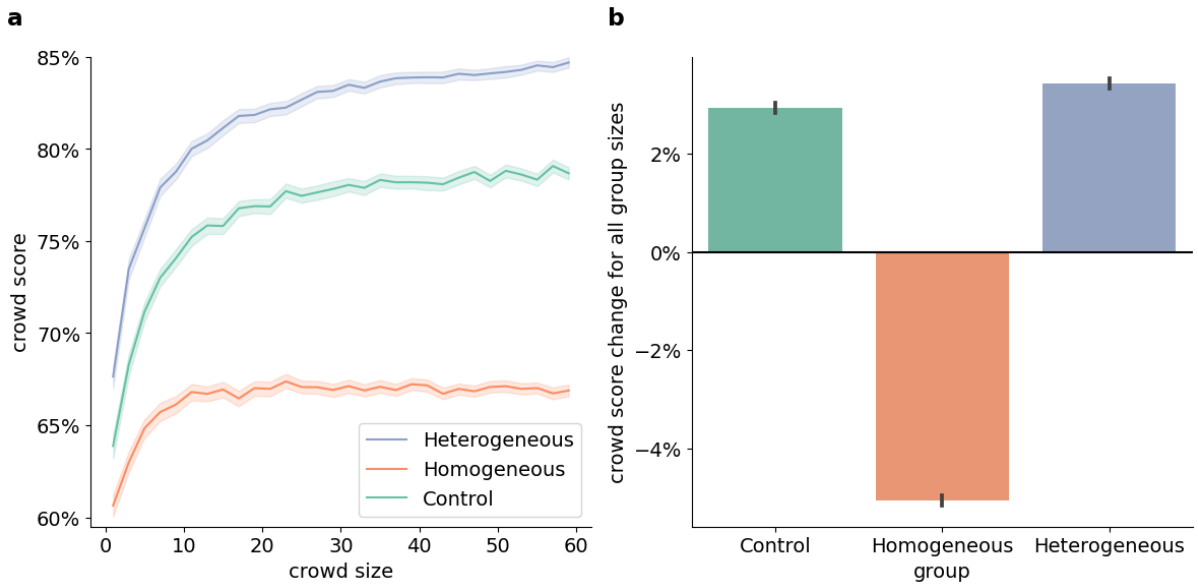

Figure S8 | Heterogeneous and Homogeneous group participants differ in the type of their revisions: Homogeneous-group participants are more likely to change their mind towards politically consistent beliefs (panel **a** for Study 1, panel **c** for Study 2), whereas Heterogeneous-group participants tend to revise their answers more accurately (panel **b** for Study 1, panel **d** for Study 2) at the expense of changing their minds in opposition to their partisanship.