Interactive crowdsourcing to fact-check politicians

Santos Espina Mairal[1,*] Florencia Bustos[1] Guillermo Solovey[2,3] Joaquin 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

[*] 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 empirical gap and also studies the effect of social influence on the accuracy of collective judgements about the veracity of phrases pronounced by politicians. Participants (N=180) first read 20 politically-balanced phrases and made individual judgements. Then, they were randomly assigned to discuss the same phrases with another politically homogeneous or heterogeneous person, or to a control condition with no social influence. Finally, we asked them to provide revised individual judgements. We found that only heterogeneous dyads largely increased their accuracy after social influence and that crowdsourced revised judgements can yield above 75% performance which is comparable with the accuracy observed in previous studies about crowdsourcing and fake news. Overall, our results extend past efforts on developing crowdsourcing tools for aiding fact-checking agencies across a formerly unexplored information format, and shows how setups with minimalistic social influence can further increase the ability of the wisdom of crowds to fact check politicians.

1. Introduction

The spread of false information has substantially sped up in recent years, posing tangible risks to public health[1,2], democratic life[3], and the fight against climate change[4]. 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 largely overloaded and they cannot keep pace with the amount of false information that they need to process[1]. 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[5].

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"[6], recent research tested the reliability of crowdsourcing as a potential tool to assist fact-checkers[7–9]. 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[9]. 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 largely 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[10]. The first goal of this 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[11–15]. 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[16], and that support to political figures can remain unchanged even after their claims have been fact-checked as false[17]. 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 this work is to investigate whether partisan biases can be reduced in social interactive settings and to test whether social influence 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[18–20], several studies reported remarkably positive effects[21–25]. 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[26]. This study suggests that interacting with politically diverse individuals may reduce partisan biases by breaking filter bubbles and promoting informational flow[27–29]. This is consistent with findings relating homogeneous social settings with polarization of political

stances, and heterogeneous interaction resulting in individuals becoming more accuracy-based[30]. However, other studies have shown that the benefit of heterogeneous social influence is absent in small groups[25], and yet another study has found that interacting in politically homogeneous networks largely increases the accuracy of the wisdom of crowds[21]. Therefore, on the basis of these disparate findings, here we considered the two possibilities and examined the effect of social interaction 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 interaction with politically heterogeneous or homogeneous individuals. We present results from a behavioural experiment where participants interacted in dyads with someone supporting the same or opposite political party and compared their accuracy against a control condition with no social influence. To anticipate our results, we found that social influence improved crowdsourced judgements to fact-check politicians but only if participants interacted with politically heterogenous, but not homogenous, individuals.

2. Materials and Methods

2.1. 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) 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 prior to the experiment.

2.2. Participants

This experiment was performed in Argentina, a politically polarized country[31–33] 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. Participants first completed a form consisting of a series of demographic items, a CRT test[34,35] and 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.

2.3. 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 assortment throughout the experiment.

2.4. 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 to 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.

2.5. Procedure

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[24], the probability of reaching consensus in polarized moral issues[32], and the effect of diversity on herding behavior[36]. 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 pronouncer as well as a brief description of his/her public position, and the approximate date when the phrase was pronounced. 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 between them whatsoever. 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.

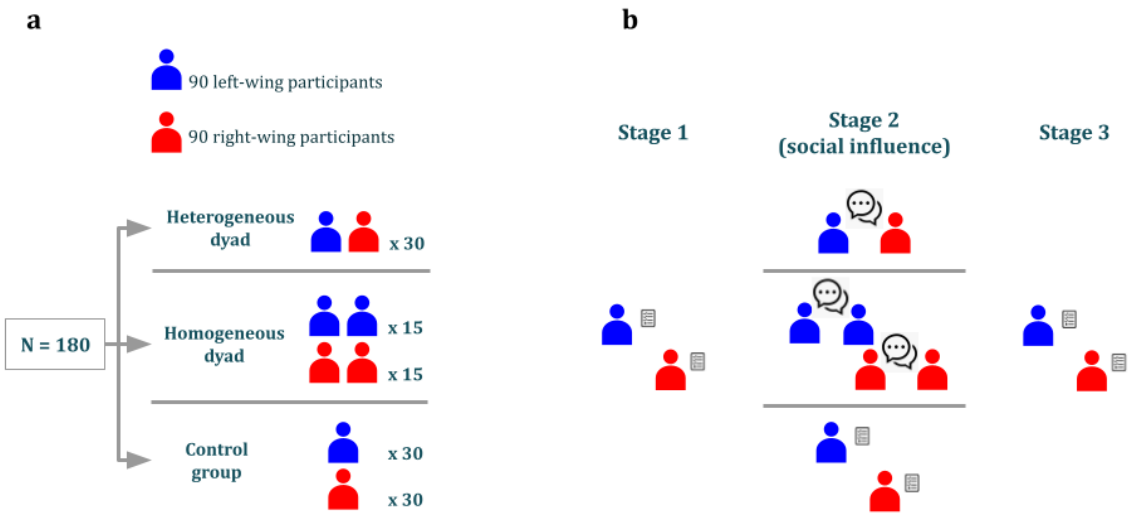

Fig. 1 | Experimental procedure. a. Participants were paired in dyads of similar or opposed partisan affinity, depending on the experimental condition. **b.** The experiment had a 3-stage structure: after individually classifying 20 statements as true or false, participants were exposed to social influence,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

2.6. 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 in paper and then type them into a pre-assigned, individual Google Sheet (https://docs.google.com/spreadsheets) file sent to them (**Figure S2**). 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 blindly assigned to each dyad and so they were unaware about whether they were testing participants in the "homogeneous" or "heterogeneous" condition.

2.7. Data analysis

Reported statistical analyses were performed in Matlab R2018b. Full specification and details on the mixed-effects logit model reported in Section 3.2 can be found in **Table S3**. Figures were generated in Python Jupyter Notebooks and Matlab.

For aggregating individual judgements into collective estimates (Section 3.3), 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*.

2.8. Data and code availability

All data and codes to reproduce our findings are available at https://osf.io/ngw8p.

3. Results

3.1. 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 (t_{179} = 11.6, p < 0.001; two-tailed t-tests are employed from now on if not specified). There were no differences in initial accuracy between experimental groups (F_{179} = 0.59, p = 0.55), and performance was approximately equal across politically concordant and discordant statements (t_{358} = -0.92, p = 0.35). 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[13].

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 phrases pronounced by politicians of the same political party they support compared to phrases pronounced 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 ($t_{179} = 16.4$, $p < 0.001$), suggesting that partisanship predicts beliefs about the veracity of statements pronounced 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 ($F_{179} = 2.05$, $p = 0.13$).

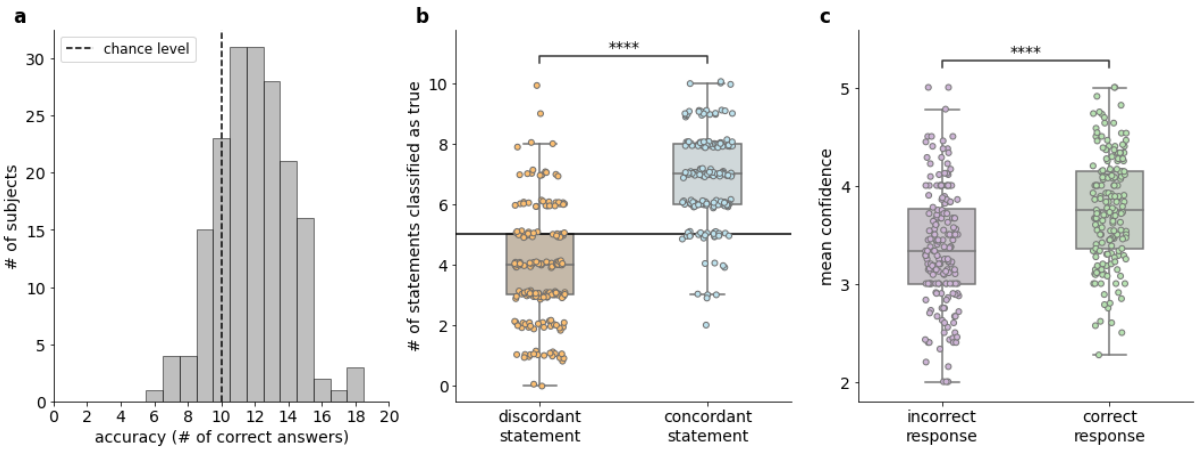

Fig. 2 | Stage 1 responses. a. Accuracy of initial individual responses. **b.** Number of statements classified as true by partisan concordance of pronouncer. Theoretically unbiased participants would result in no significant differences between both conditions. Each point marks the count of answers 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.

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, $t_{179} = 11.3$, $p < 0.001$), revealing they had -to some extent- insight into their performance, without significant pre-treatment differences displayed by experimental conditions ($F_{179} = 0.22$, $p = 0.81$).

3.2. 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**). As a benchmark, individuals in the Control condition showed a small, albeit significant, increase in performance ($t_{59} = 3.01$, $p = 0.004$). We found that participants in the Homogeneous condition did not significantly alter their performance compared to Stage 1 ($t_{59} = 0.71$, $p = 0.48$) and this change was statistically indistinguishable to the one observed by the individuals lacking social influence ($t_{118} = 1.25$, $p = 0.21$). Instead, Heterogeneous dyad members showed an increase in performance compared to Stage 1 ($t_{59} = 4.5$, $p < 0.001$) which was significantly larger in magnitude to the one observed by participants in the Control condition ($t_{118} = 2.40$, $p = 0.018$).

To test the robustness of this result, 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, 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**, 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**, 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 (t_{59} = 1.16, p = 0.25) and individuals in the Control condition (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 (t_{118} = 4.84, p < 0.001 and 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.

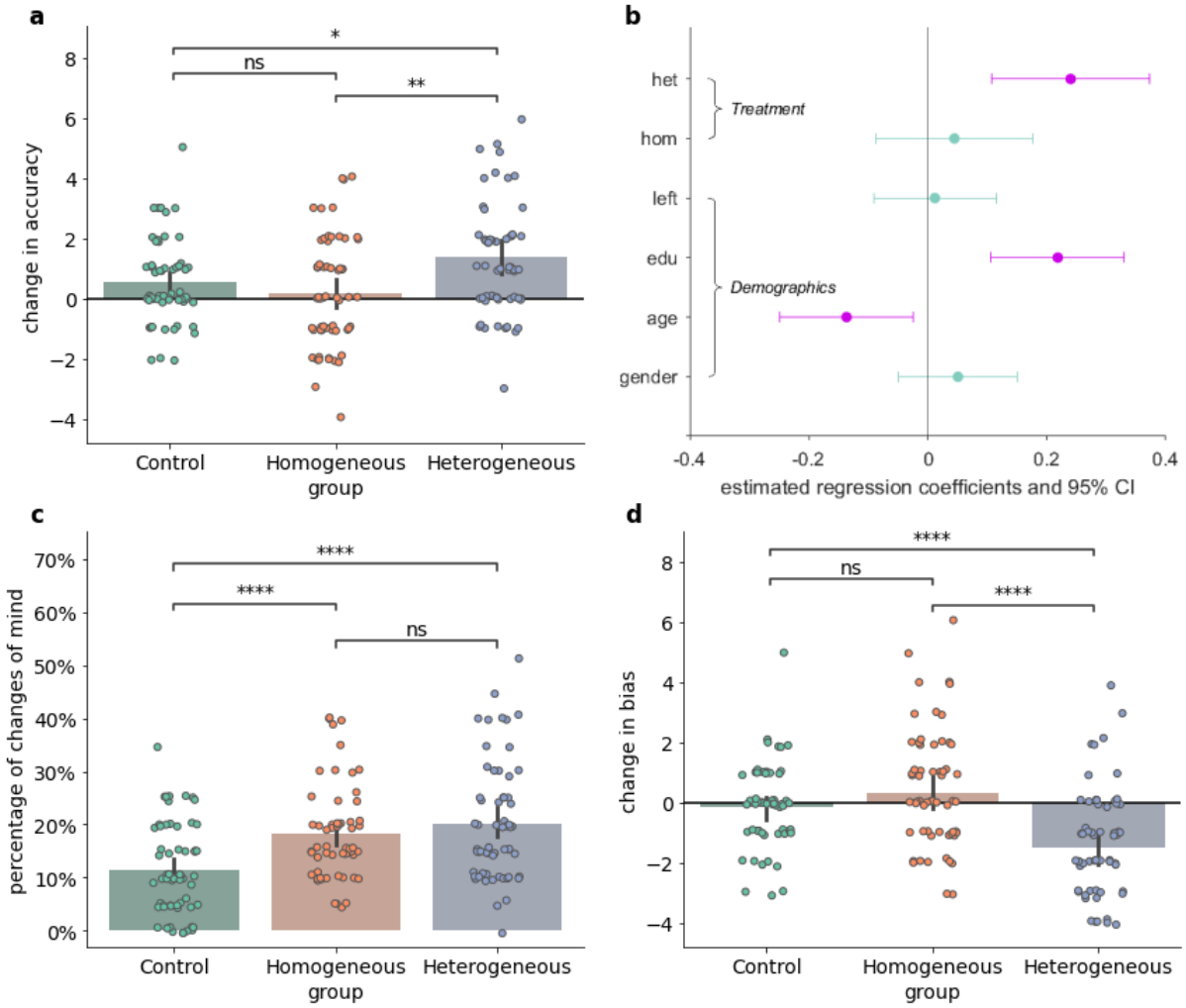

Fig. 3 | Treatment effect. a. Mean and s.e.m. of change in the number of individual correct answers between Stage 1 and Stage 3, by group. Outliers are omitted in the plots, but included in the analysis. **b.** Coefficients and CIs of predictors resulting from running a GLME Logistic regression on the probability of giving a correct answer in Stage 3. **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.

3.3. Crowd performance: aggregating individual knowledge

Our main goal involved studying crowdsourcing as a potential tool for fact-checking politicians. By aggregating individual revised responses, 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 (β = 0.313±0.037, t = 8.36, p < 0.001), Heterogeneous crowds performed better than Control crowds (β = 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[7].

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, 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 (t_{29} = 35.1, p < 0.001) and the Heterogeneous conditions (t_{29} = 30.2, p < 0.001) and significantly smaller than zero for the Homogeneous condition (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 (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 (t_{58} = 31.8, p < 0.001).

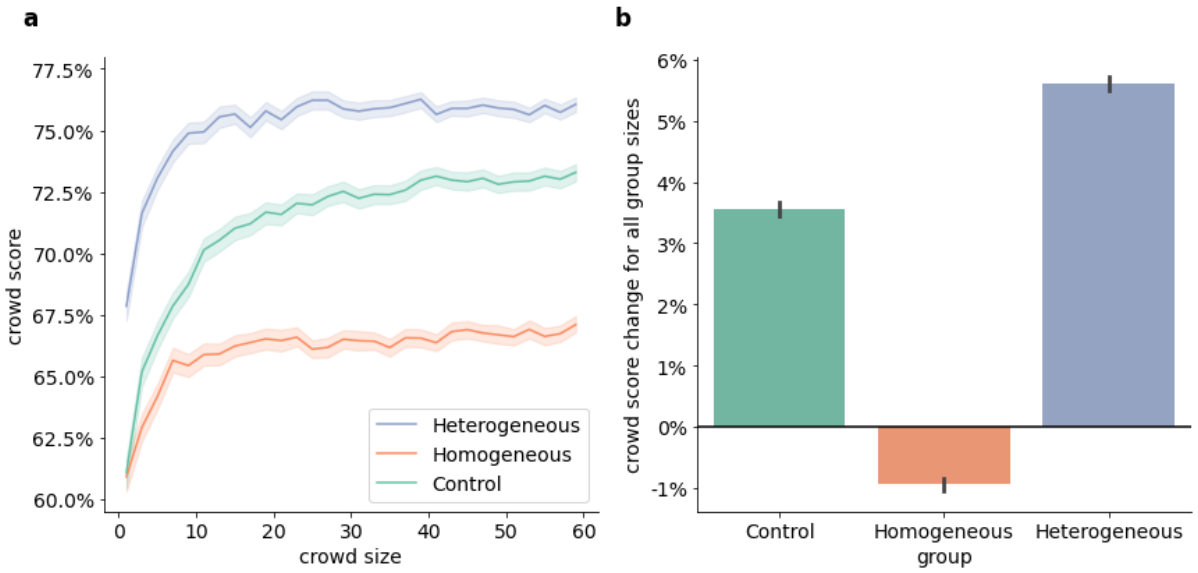

Fig. 4 | Wisdom of crowds. a. Accuracy of the crowd by group and by crowd size, using Stage 3 responses. Score is defined as the number of correct responses as a % of total statements (20). Mean (solid line) and s.e.m. (shaded area) across 1,000 iterations. **b.** Change in crowd accuracy between Stage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

4. 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. 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[22,37] and fake news[11,13–15] to a previously unexplored format. Studying the effect of dyadic interaction on pre-existing partisan bias, we found social influence to act as a partial antidote to polarization only when happening in heterogeneous settings. While social interaction increased the overall likelihood of participants changing their minds over initial answers, only inter-partisan communication resulted in revised answers that were opposed to partisanship. Regarding homogeneous dyad members, who had their partisan biases aligned, social interaction failed to increase their accuracy. Contrarily, interaction in heterogeneous dyads (characterized by opposing partisan biases) seemed to nudge their members to prioritize accuracy over partisanship, depolarizing pre-existing biases. As a result, performance only improved amongst participants who had interacted with peers of opposing political ideology, a result consistent with Klar[30] but differing from Becker et al.[21] 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.[21] used a factual estimation task, whereas we asked participants to directly report whether or not political leaders were telling the truth. Further research is needed to

understand to what extent these differences in task characteristics influence the effect of homogeneous conversations on individual and collective accuracy.

Given the experimental design of this study, these results can causally link the boost in detection accuracy to the structure of social influence. However, one main limitation of this work is that these results cannot directly pinpoint the mechanism underlying the reduction in partisan biases and the subsequent error reduction. We have evidence to partially rule out the quantity of social interaction as a potential candidate, as the number of phrases discussed in dyads was not a predictor of changes in bias or accuracy. A possible mechanism involves the straightforward act of realizing that someone else thinks differently than ourselves. Future research could explore in detail the underlying contribution of social influence in determining which aspects of interaction account for partisan depolarization when happening in heterogeneous settings.

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, our findings stem from a sample of participants with a defined, disclosed partisan affinity. However, extending this setup to moderate participants is challenging given the strength of political polarization in Argentina[16,31]. Finally, individual accuracy was possibly limited by time and information constraints determined by experimental design, especially taking into account the fact that participants only had access to the spoken phrase and the name of the pronouncer. Enabling lay raters with additional resources could increase overall accuracy, although future research is required in order to test this idea.

Overall, these results suggest that lay people can rate the veracity of political speech with above-chance performance and reduce their partisan biases through social interaction 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. However, the fact that crowds' performance was below 80% indicates this instrument cannot be envisioned to replace the work of fact-checkers, but rather to aid it. Further research is needed to extend this setup to diverse contexts, countries and formats.

5. References

1. Burel, G., Farrell, T., Mensio, M., Khare, P. & Alani, H. Co-Spread of Misinformation and Fact-Checking Content during the Covid-19 Pandemic. in (2020).
2. Burki, T. Vaccine misinformation and social media. *Lancet Digit. Health* **1**, e258–e259 (2019).
3. Frau-Meigs, D. Societal costs of fakenews in the digital single market. *Eur. Parliam.* 40 (2018).
4. van der Linden, S., Leiserowitz, A., Rosenthal, S. & Maibach, E. Inoculating the Public against Misinformation about Climate Change. *Glob. Chall.* **1**, 1600008 (2017).
5. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
6. Surowiecki, J. *The Wisdom of Crowds*. *Nature* (2005).
7. Allen, J., Arechar, A. A., Pennycook, G. & Rand, D. G. Scaling up fact-checking using the wisdom of crowds. *Sci. Adv.* (2021).
8. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci.* **116**, 2521 (2019).
9. Resnick, P., Alfayez, A., Im, J. & Gilbert, E. Informed Crowds Can Effectively Identify Misinformation. *ArXiv210807898 Cs* (2021).
10. Ajzenman, N., Cavalcanti, T. & Da Mata, D. *More Than Words: Leaders' Speech and Risky Behavior during a Pandemic*. https://papers.ssrn.com/abstract=3582908 (2020) doi:10.2139/ssrn.3582908.
11. Faragó, L., Kende, A. & Kreko, P. We Only Believe in News That We Doctored Ourselves: The Connection Between Partisanship and Political Fake News. *Soc. Psychol.* **51**, 1–14 (2019).
12. Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A. & Petersen, M. B. Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *Am. Polit. Sci. Rev.* **115**, 999–1015 (2021).

13. Pennycook, G. & Rand, D. G. The Psychology of Fake News. *Trends Cogn. Sci.* **25**, 388–402 (2021).
14. Pereira, A., Harris, E. A. & Bavel, J. J. V. Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. Preprint at https://doi.org/10.31234/osf.io/7vc5d (2018).
15. Vegetti, F. & Mancosu, M. The Impact of Political Sophistication and Motivated Reasoning on Misinformation. *Polit. Commun.* **37**, 678–695 (2020).
16. Levy Yeyati, E., Moscovich, L. & Abuin, C. Leader over Policy? The Scope of Elite Influence on Policy Preferences. *Polit. Commun.* **37**, 398–422 (2020).
17. Swire-Thompson, B., Ecker, U. K. H., Lewandowsky, S. & Berinsky, A. J. They Might Be a Liar But They're My Liar: Source Evaluation and the Prevalence of Misinformation. *Polit. Psychol.* **41**, 21–34 (2020).
18. Frey, V. & van de Rijt, A. Social Influence Undermines the Wisdom of the Crowd in Sequential Decision Making. *Manag. Sci.* **67**, 4273–4286 (2021).
19. Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci.* **108**, 9020–9025 (2011).
20. Mavrodiev, P. & Schweitzer, F. The ambigous role of social influence on the wisdom of crowds: An analytic approach. *Phys. Stat. Mech. Its Appl.* **567**, 125624 (2021).
21. Becker, J., Porter, E. & Centola, D. The wisdom of partisan crowds. *Proc. Natl. Acad. Sci.* **116**, 10717–10722 (2019).
22. Guilbeault, D., Becker, J. & Centola, D. Social learning and partisan bias in the interpretation of climate trends. *Proc. Natl. Acad. Sci.* **115**, 9714–9719 (2018).
23. Jayles, B. *et al.* How social information can improve estimation accuracy in human groups. *Proc. Natl. Acad. Sci.* **114**, 12620–12625 (2017).
24. Navajas, J., Niella, T., Garbulsky, G., Bahrami, B. & Sigman, M. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nat. Hum. Behav.* **2**, 126–132 (2018).
25. Pescetelli, N., Rutherford, A. & Rahwan, I. Modularity and composite diversity affect the collective gathering of information online. *Nat. Commun.* **12**, 3195 (2021).
26. Shi, F., Teplitskiy, M., Duede, E. & Evans, J. A. The wisdom of polarized crowds. *Nat. Hum. Behav.* **3**, 329–336 (2019).
27. Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proc. Natl. Acad. Sci.* **118**, (2021).
28. Rhodes, S. C. Filter Bubbles, Echo Chambers, and Fake News: How Social Media Conditions Individuals to Be Less Critical of Political Misinformation. *Polit. Commun.* **0**, 1–22 (2021).
29. Sunstein, C. R. *Going to Extremes: How Like Minds Unite and Divide*. (Oxford University Press, 2009).

30. Klar, S. Partisanship in a Social Setting. *Am. J. Polit. Sci.* **58**, 687–704 (2014).
31. Freira, L., Sartorio, M., Boruchowicz, C., Lopez Boo, F. & Navajas, J. The interplay between partisanship, forecasted COVID-19 deaths, and support for preventive policies. *Humanit. Soc. Sci. Commun.* **8**, 1–10 (2021).
32. Navajas, J. *et al.* Reaching Consensus in Polarized Moral Debates. *Curr. Biol.* **29**, (2019).
33. Zimmerman, F., Garbulsky, G., Ariely, D., Sigman, M. & Navajas, J. Political coherence and certainty as drivers of interpersonal liking over and above similarity. *Sci. Adv.* **8**, eabk1909 (2022).
34. Frederick, S. Cognitive Reflection and Decision Making. *J. Econ. Perspect.* **19**, 25–42 (2005).
35. Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A. & Hamilton, J. The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). *J. Behav. Decis. Mak.* **29**, 453–469 (2016).
36. Navajas, J., Armand, O., Moran, R., Bahrami, B. & Deroy, O. Diversity of opinions promotes herding in uncertain crowds. *R. Soc. Open Sci.* **9**, 191497 (2022).
37. Bullock, J. G., Gerber, A. S., Hill, S. J. & Huber, G. A. *Partisan Bias in Factual Beliefs about Politics*. https://www.nber.org/papers/w19080 (2013) doi:10.3386/w19080.

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**). English translation and links to the original information check are provided.

#	***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.

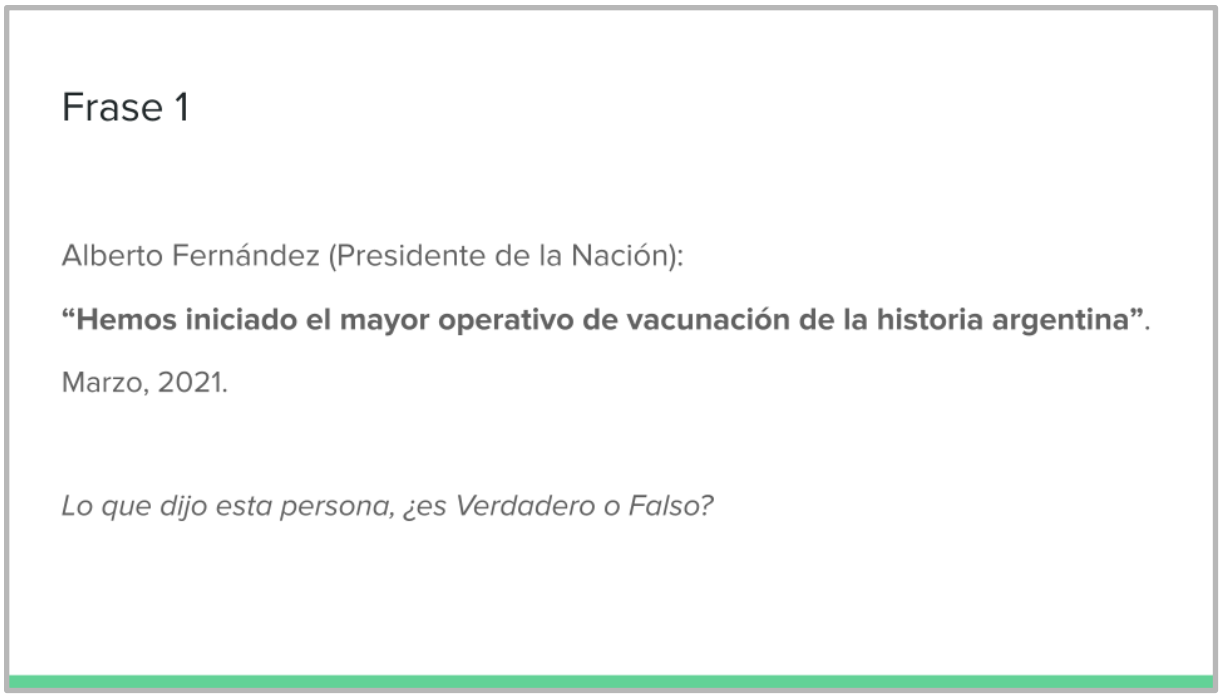

Table S2 | Corpus of true or false statements used for Control group subjects in Stage 2. Translation from Spanish to English is depicted in italics.

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, 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.

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

Fixed effects coefficients		
variable	**coefficient**	**t-statistic**
intercept	0.873***	5.43
correct_1	1.49***	30.9
heterogeneous	0.24***	3.56
homogeneous	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

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

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.