Report

Current Biology

Reaching Consensus in Polarized Moral Debates

Highlights

- We asked two live crowds to deliberate about polarized moral issues (e.g., abortion)
- In small groups, they sought consensus on the acceptability of controversial actions
- Consensus was influenced by participants with a moderate view but high confidence
- Group ratings and changes of mind suggest that people adopted a mediation process

Authors

Joaquin Navajas, Facundo Álvarez Heduan, Juan Manuel Garrido, Pablo A. Gonzalez, Gerry Garbulsky, Dan Ariely, Mariano Sigman

Correspondence

msigman@utdt.edu

In Brief

Navajas et al. study the conditions under which small groups can reach consensus about polarized moral issues. Two large-scale experiments show converging evidence that consensus was triggered by a mediation procedure. Participants with high confidence in moderate opinions ("confident grays") played an important role in resolving moral disagreement.

Navajas et al., 2019, Current Biology 29, 1–6
December 2, 2019
https://doi.org/10.1016/j.cub.2019.10.018

CellPress

Please cite this article in press as: Navajas et al., Reaching Consensus in Polarized Moral Debates, Current Biology (2019), https://doi.org/10.1016/j.cub.2019.10.018

Current Biology
Report

Reaching Consensus in Polarized Moral Debates

Joaquin Navajas,[1,2] Facundo Álvarez Heduan,[3] Juan Manuel Garrido,[3] Pablo A. Gonzalez,[3] Gerry Garbulsky,[4] Dan Ariely,[5] and Mariano Sigman[1,6,7,*]
[1]Universidad Torcuato Di Tella, Av. Figueroa Alcorta 7350, Buenos Aires C1428BCW, Argentina
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Godoy Cruz 2290, Buenos Aires C1425FQB, Argentina
[3]El Gato y la Caja, Teodoro García 2474, Buenos Aires C1426DMR, Argentina
[4]TED, Araoz 727, Buenos Aires C1414DPO, Argentina
[5]The Fuqua School of Business, Duke University, 100 Fuqua Drive, Durham, NC 27708, USA
[6]Facultad de Lenguas y Educación, Universidad Nebrija, Calle de Sta. Cruz de Marcenado 27, Madrid 28015, Spain
[7]Lead Contact
*Correspondence: msigman@utdt.edu
https://doi.org/10.1016/j.cub.2019.10.018

SUMMARY

The group polarization phenomenon is a widespread human bias with no apparent geographical or cultural boundaries [1]. Although the conditions that breed extremism have been extensively studied [2–5], comparably little research has examined how to depolarize attitudes in people who already embrace extreme beliefs. Previous studies have shown that deliberating groups may shift toward more moderate opinions [6], but why deliberation is sometimes effective although other times it fails at eliciting consensus remains largely unknown. To investigate this, we performed a large-scale behavioral experiment with live crowds from two countries. Participants (N = 3,288 in study 1 and N = 582 in study 2) were presented with a set of moral scenarios and asked to judge the acceptability of a controversial action. Then they organized in groups of three and discussed their opinions to see whether they agreed on common values of acceptability. We found that groups succeeding at reaching consensus frequently had extreme participants with low confidence and a participant with a moderate view but high confidence. Quantitative analyses showed that these "confident grays" exerted the greatest weight on group judgements and suggest that consensus was driven by a mediation process [7, 8]. Overall, these findings shed light on the elements that allow human groups to resolve moral disagreement.

RESULTS AND DISCUSSION

The adoption of more extreme views after social interaction, an effect known as group polarization [9], is a cognitive bias that has been observed in diverse contexts, including attitudes toward gender equality [2], race [3], punishment [4], and religious matters [5]. Although a vast literature has studied the conditions that propagate extremism after social influence, comparably little research has examined the opposite phenomenon, namely, the factors that enable humans to shift toward moderate views. Early experiments in social psychology showed that groups can sometimes depolarize and adopt moderate attitudes through deliberation [6, 10, 11]. However, the procedures that humans use to succeed in resolving large disagreements have yet to be uncovered. In this study, we aimed at understanding whether and how people can reach consensus about moral judgments, a domain with vast relevance to policy making (e.g., legislation on abortion) [12]. To examine the mechanisms underlying consensus, we study whether and how collective decisions, when they are reached, can be inferred from the initial beliefs of group members.

We performed two large-scale behavioral studies capitalizing on a program to perform experiments with crowds attending popular events [13–15]. The experimental design was almost identical in structure to a previous experiment where we studied the effect of deliberation on the wisdom of crowds [15] (see STAR Methods for details). In a first stage, participants were presented with four moral scenarios that described a controversial action (see STAR Methods). After listening to each scenario, participants were asked to rate how acceptable they found the action from 0 ("completely unacceptable") to 10 ("completely acceptable"; Figure 1A). Participants also reported how confident they felt about their previous judgement in a scale from 0 ("completely uncertain") to 10 ("completely certain"; Figure 1B). In the second stage, participants organized in groups of three and discussed each scenario for a maximum of 2 min. If, after deliberation, all group members agreed to summarize their views in a shared value of acceptability, then they wrote down this number. If, instead, they could not find a value that was acceptable for all, they simply wrote down an "X," indicating that they had not reached consensus. Finally, in a third stage, participants expressed a revised degree of acceptability, having the possibility to change their minds and revise their opinions after deliberation.

The shape of the distribution of acceptability ratings varied across the scenarios (Figures 1A and S1), but in all of them, a substantial fraction of participants selected the most extreme ratings (34%–56% selected either 0 or 10). As shown previously [16, 17], confidence followed a quadratic relationship with acceptability (Figures 1B and S1). Participants with intermediate ratings were typically less confident than the ones giving extreme ratings. However, and in agreement with previous findings [18, 19],

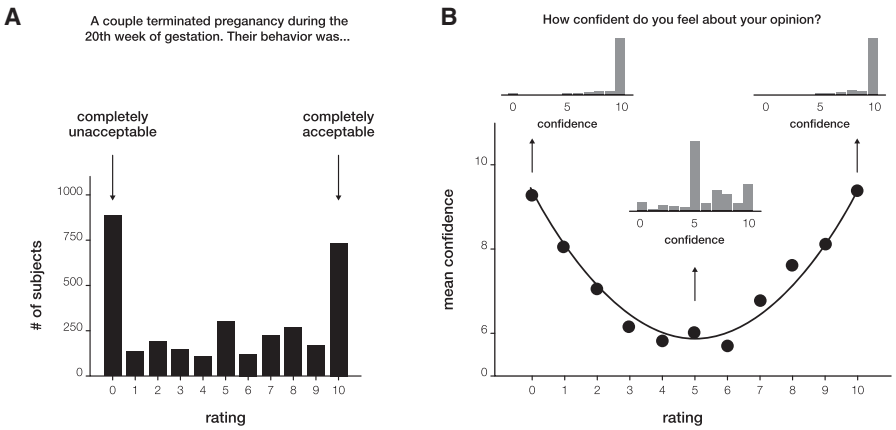

Figure 1. Acceptability and Confidence Ratings in a Polarized Moral Scenario

Participants were asked to judge, in a scale from 0 to 10, the acceptability of an action described in a moral scenario.

(A) Histogram of acceptability ratings for one of these scenarios (of a total of 4), which was about the voluntary interruption of pregnancy after 20 weeks of gestation (see STAR Methods for the full version of all scenarios and Figure S1 for the histograms of all scenarios).

(B) Mean confidence about their previous judgement (in a scale from 0 to 10) as a function of the initial acceptability rating. Circles show mean confidence, SEM is within the height of the circles, and the line shows the best-fitting quadratic function. See also Figure S1 for the distribution of confidence on all scenarios. Insets depict the normalized histogram of confidence for extreme and intermediate acceptability ratings.

confidence varied widely between subjects. More specifically, participants reporting moderate ratings of acceptability showed a multimodal distribution of confidence across participants (middle inset of Figure 1B), with almost one-third of them providing high values of confidence (29.4% of the participants with an acceptability rating of 5 provided a confidence rating of 8, 9, or 10).

Once participants provided initial opinions and confidence, they proceeded to deliberate in groups of three. We estimated the likelihood of reaching consensus by measuring the fraction of groups that agreed to provide a collective rating. To organize the data of each group, we sorted acceptability ratings in ascending order. This way, we assigned three different roles; participants with lowest, middle, and highest ratings were respectively defined as p_{min}, p_{med}, and p_{max}. Accordingly, we refer to their initial ratings as r_{min}, r_{med}, and r_{max}. To quantify the diversity of opinions, we measured the range of ratings within groups (i.e., $\delta = r_{max} - r_{min}$; Figure 2A). As expected, the probability of reaching consensus decreased for groups with increasing range of opinions (Figures 2A and S2; logistic regression, study 1: $\beta_\delta = -1.03 \pm 0.04$, $p = 4 \times 10^{-116}$; study 2: $\beta_\delta = -1.98 \pm 0.25$, $p = 3 \times 10^{-14}$; see STAR Methods for details).

We then focused on answering a crucial question in our study: what distinguishes those groups that reached consensus from those that failed at doing so? Based on theoretical arguments [20, 21], we started by comparing two mechanisms that can trigger consensus (Figure 2B). One simple strategy is a majority rule, by which the minority herds toward the most popular answer (upper panel of Figure 2B). Majority rules are ubiquitous in collective systems, from animal swarms [22] to democratic societies [20]. An alternative way of reaching consensus is through mediation (lower-left panel of Figure 2B). In this procedure, opposing views are brought together by an interlocutor with an intermediate opinion that can express with conviction arguments in favor of the opposing views [7].

Although both accounts make the same reasonable prediction that similar groups should reach consensus more often than diverse groups (Figure 2A), they make opposite predictions about how the symmetry of initial ratings relates to the likelihood of reaching consensus. The majority rule predicts that groups with asymmetric ratings, in which a majority of the individuals start the discussion with highly similar opinions (upper panel, Figure 2B), have a higher probability of reaching consensus. Instead, mediation processes rely on an intermediate agent that can bridge both opinions (lower panel of Figure 2B) and hence predict that symmetric distributions are the most likely configurations to result in consensus.

To quantify the degree of symmetry in the distributions of group ratings, we measured their absolute non-parametric skewness (see Equations 1 and 2 in STAR Methods for the definition of symmetry). Figure 2C shows how p(consensus) changes relative to baseline for groups with different symmetry, i.e., Δp(consensus) (see Equation 4 in STAR Methods). We observed that the more symmetric groups were more likely to reach consensus (median split of symmetry, study 1: $\chi^2(1) = 10.5$, $p = 0.001$; study 2: $\chi^2(1) = 5.9$, $p = 0.01$). We then fitted a logistic regression with range (δ) and symmetry (S) as predictor variables (see Equation 3 in STAR Methods) and looked at whether S modulated the probability of reaching consensus with positive sign ($\beta_S > 0$, as predicted by a mediation process) or negative sign ($\beta_S < 0$, as predicted by a majority model). We found that symmetry had a positive main effect on the probability of reaching consensus (logistic regression, study 1: $\beta_S = 0.09 \pm 0.04$, $p = 0.02$; study 2: $\beta_S = 1.0 \pm 0.45$, $p = 0.03$).

We then examined an additional prediction of the mediation hypothesis that relates to the strength and conviction of the mediator. Specifically, mediators are agents with intermediate opinions but high confidence [7, 23], who are able to persuade extreme participants to adopt a moderate position. To test this prediction, we measured the change in p(consensus) associated

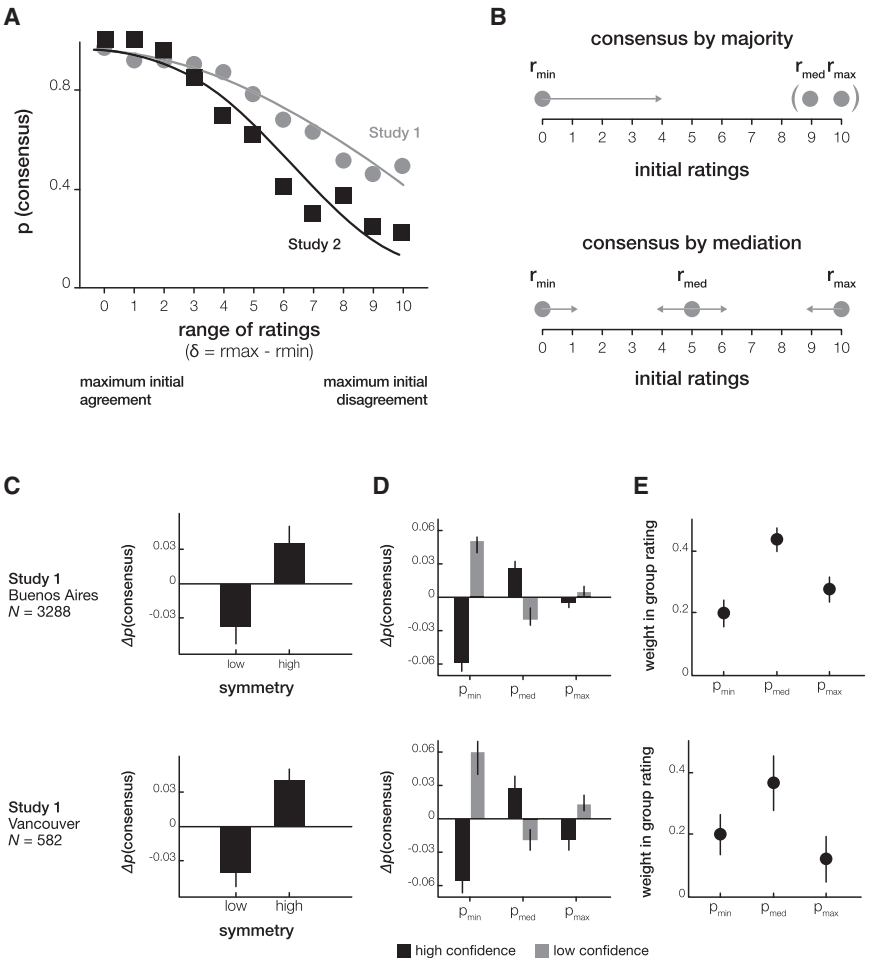

Figure 2. Reaching Consensus in Polarized Moral Issues

(A) The probability of reaching consensus as a function of initial disagreement in the group, quantified by the range of ratings $\delta = r_{max} - r_{min}$. Dots (study 1) and squares (study 2) show the fraction of groups reaching consensus as a function of δ across all scenarios, SEM is within the heights of the markers, and the line shows the best-fitting logistic regression.

(B) Groups may reach consensus through different strategies. In this panel, we consider a majority rule (MR) or a mediation process (MP). MR (upper panel) predicts that more asymmetric groups should be more likely to reach consensus. MP (lower panel) predicts the opposite.

(C–E) Upper panels refer to study 1; lower panels refer to study 2.

(C) Change in the probability of reaching consensus relative to baseline, i.e., Δp(consensus), for groups with low or high symmetry in their distribution of ratings.

(D) Change in the probability of reaching consensus relative to baseline, i.e., Δp(consensus), associated with having a participant with low or high confidence at each of the three roles.

(E) Circles show the weight that each participant exerted on group ratings, as estimated by a linear regression. Vertical lines depict SEM. The participant with intermediate acceptability had the largest weight in the group judgement.

See also Figure S2.

with having a participant with "low" or "high" confidence at each of the three roles (Figure 2D). We found that groups where extreme participants reported low confidence had a higher probability of reaching consensus than groups where extreme participants expressed high confidence (median split of confidence, study 1: $\chi^2(1) = 73.17$, $p = 4 \times 10^{-10}$ for p_{min} and $\chi^2(1) = 5.01$, $p = 0.03$ for p_{max}; study 2: $\chi^2(1) = 29.07$, $p = 2 \times 10^{-8}$ for p_{min} and $\chi^2(1) = 0.16$, $p = 0.68$ for p_{max}). Conversely, groups where moderate participants reported high confidence had a larger probability of reaching consensus compared to groups where these intermediate participants had low confidence (study 1: $\chi^2(1) = 10.28$, $p = 0.001$; study 2: $\chi^2(1) = 5.17$, $p = 0.02$).

These two findings jointly provide support for the mediation model for two reasons. First, mediation relies on extreme individuals being open to change their minds, and this is more likely to happen when extreme participants display low confidence [24, 25]. This effect was more pronounced for participants with low acceptability ratings (p_{min}), which suggests that a key factor in resolving moral disagreements consists in persuading individuals that initially oppose to the acceptability of controversial actions. Second, mediation is more likely to be successful if intermediate participants are confident enough to bridge the two extremes. Our data are consistent with this observation, given the positive correlation between confidence and consensus for moderate participants.

The mediation hypothesis not only makes predictions about the configurations that are likely to trigger consensus but also about the kind of consensus that groups should reach. Mediation processes imply the existence of an interlocutor that dominates the discussion and may exert a large influence on the group judgment. This can be examined by measuring the weight that the opinion of each of the three group members (p_{min}, p_{med}, and p_{max}) had on the collective rating (see Equation 5 in STAR Methods). As predicted by the mediation hypothesis, we found that participants with intermediate opinions exerted the largest weight on group judgments (Figures 2E and S2). We also observed that this effect became larger as their confidence increased (study 1: $\beta = 0.011 \pm 0.005$, $p = 0.006$; study 2: $\beta = 0.02 \pm 0.01$, $p = 0.05$).

We performed a series of control analyses to discard alternative accounts that could potentially explain these findings. First, we found that two models previously proposed to explain group decisions, such as the simple average [26] and confidence-weighted average [27] rules, were inconsistent with the large weight associated to p_{med} in Figure 2E. This suggests that groups did not implement these procedures (Figure 3). We also found no significant effects of gender or age (Figure S3) on p(consensus).

We also evaluated whether the convergence to intermediate ratings was explained by regression to the mean. To test this possibility, we performed a control experiment (N = 98) where individuals provided initial and revised ratings without any kind of

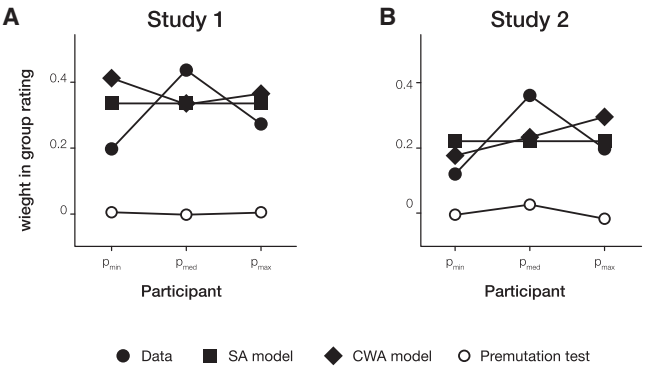

Figure 3. Simple Models of Consensus Do Not Explain the Large Weight Associated to Intermediate Participants

Weights obtained by regressing collective ratings against a linear combination of the initial ratings of participant p_{min}, p_{med}, and p_{max} (see Equation 5 in STAR Methods for details). Left (right) panel shows data from study 1 (study 2). Black circles: weights extracted by fitting the empirical data are shown. Squares: weights expected by a simple-average (SA) model are shown. Diamonds: weights expected by a confidence-weighted average (CWA) model are shown. Blank circles: permutation test where we shuffled the labels of the initial and group ratings, expecting no correlation between them, is shown. See also Figure S3.

social interaction between them. With this dataset, we constructed a set of "virtual triads" (i.e., groups of three participants that did not interact with each other) and performed a series of analyses to see which of our findings could be replicated in this setting that lacks social influence. Our data suggest that resampling alone could not explain several of our previous observations (see STAR Methods for details).

We then examined another prediction of the mediation hypothesis regarding the revision of private opinions following deliberation. Specifically, mediation should increase exchange of views between extreme individuals, and thus, there should be an increased influence of the initial opinion of one extreme person on the final rating of the other extreme person after consensus (i.e., the initial rating r_{max} should have an influence of the final rating r_{min} and vice versa). Moreover, if this process relies on mediation, this influence should be modulated by the intermediate person's confidence. To test these predictions, we performed a bivariate linear regression with "revised rating at one extreme" (i.e., RE, revised r_{min} or r_{max}) as dependent variable (Equation 6 in STAR Methods). As predictor variables, we included "initial rating at that extreme" (i.e., IE, initial r_{min} or r_{max}, respectively) and "initial rating at the other extreme" (IO, initial r_{max} or r_{min}, respectively).

In accordance with the mediation model, the influence of IO on RE was not significantly different from zero in the absence of interaction (Figure 4A; circle in condition NI; $\beta = .001 \pm .01$, $p = .94$) and for groups where interaction did not lead to consensus (Figure 4A; circle in condition NC; $\beta = .01 \pm .03$, $p = .71$). In turn, we observed a significantly positive effect of IO on RE for groups reaching consensus (Figure 4A; circle in condition C;$\beta = .33 \pm .03$, $p = 10^{-12}$). We also observed that this effect was modulated by the intermediate person's confidence (Figure 4B). The partial correlation between IO and RE after controlling by IE was significantly larger for groups with a high-confident intermediate participant (Figure 4B; $r_{high} = .28$; $r_{low} = .17$; $p = .005$). For a direct comparison between interacting groups and the control condition, see Equation 7 in STAR Methods and Table S1. Overall, these analyses provide more evidence in favor of mediation and against the idea that the resampling procedure fully explains our results.

Finally, we sought to understand the relationship between our findings and the phenomena of group polarization [2, 28] and depolarization [6]. To study these effects, we first quantified the attitude shift for all debates (Figure 4C) by taking the average individual rating of each group before versus after the debate (see Equation 8 in STAR Methods). We then looked whether this average became more extreme (polarization) or less extreme (depolarization). Considering all debates proceeding from all groups, we observed polarization in 41.1% of them and depolarization in a similar proportion (42.5%; chi-square test; $\chi^2(1) = 1.5$; p = 0.23). The remaining interactions (16.4%) showed no change in attitude extremity. This result suggests that both phenomena—group polarization and depolarization—are present in our data. However, and consistent with previous experiments [6], we observed that the size of depolarization (mean ± SEM = 1.49 ± 0.03) was larger than the size of polarization (mean ± SEM = 1.11 ± 0.02; unpaired t test; $t(3,182) = 10.8$; $p = 10^{-27}$).

We then asked what distinguishes those debates showing depolarization from those showing polarization. We found that the likelihood to observe depolarization was higher in groups reaching consensus ($\beta = 1.39 \pm .22$, $p = 10^{-10}$) and also for groups with high symmetry ($\beta = .67 \pm .30$, $p = .02$). We also found an interaction between the effects of consensus and symmetry ($\beta = -1.11 \pm .36$, $p = .002$), by which low-symmetry groups reaching consensus have a lower probability to show depolarization (Figure 4D). We believe that this result is also consistent with the idea that intermediate participants played an important role in reaching consensus. Because the middle participant of an asymmetric group is more extreme than the average opinion (e.g., in a group formed by 0-9-10, the average is 6.3, but the intermediate participant has a rating of 9), a large influence of intermediate participants in consensus implies that asymmetric groups should polarize. This explains why we observed a lower probability of depolarization in asymmetric groups reaching consensus (black dots in Figure 4D).

Our findings provide new insights about how people achieve consensus on controversial moral issues. We show that participants with intermediate opinions seem to play an important role in resolving disagreement and that their confidence acts as a moderator of their influence (Figure 4B; Table S1). We refer to these individuals who might promote consensus as "confident grays" and show evidence that these participants are key in the process of reaching consensus (Figures 2D, 2E, and 4B).

Previous theoretical research has defined mediation as a "conflict resolution process where groups of people negotiate mutually acceptable solutions that resolve their differences" [25]. Based on the observation that mediators should engage with all parties involved in the resolution of their differences, we used group symmetry as a proxy for the potential success of mediation. Our results suggest that symmetry not only matters for reaching consensus (as mediation suggests; Figure 2C) but also for the likelihood to observe depolarization (Figure 4D).

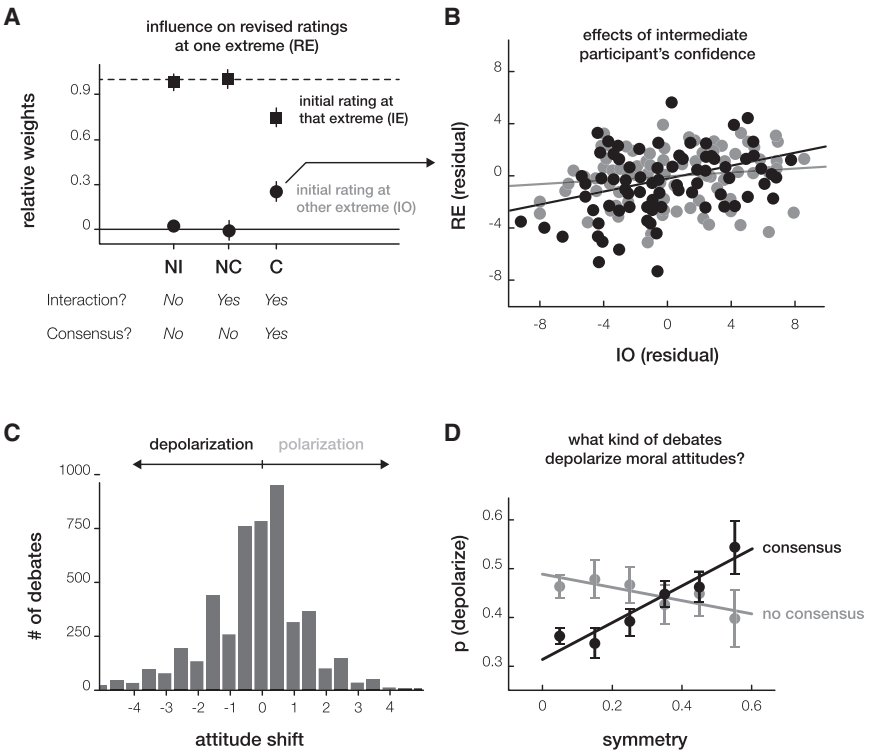

Figure 4. Effect of Deliberation on Revised Ratings

(A) Testing for a potential influence between extremes. We performed a bivariate regression by which the revised ratings at one extreme (RE) were modeled as a linear combination of the initial rating at that extreme (IE) and the initial rating at the other extreme (IO). The y axis shows the relative weights of that regression for three different conditions (x axis). No interaction (NI): virtual triads constructed using data from the control experiment where individuals revised their initial ratings in the absence of social influence are shown. No consensus (NC): triads that failed at reaching consensus after deliberation are shown. Consensus (C): triads that succeeded at reaching consensus after deliberation are shown.

(B) Only for groups reaching consensus, we looked at the partial correlation between RE and IO for groups where the intermediate participant displayed low versus high confidence. Dots show the residuals of RE (y axis) versus the residuals of IO (x axis) after controlling both variables by IE. This is shown separately, with their best-fitting lines, for groups where intermediate participants that displayed low (gray dots) and high (black dots) confidence.

(C) Histogram of attitude shifts (see Equation 8 in STAR Methods). Negative values depict depolarization, and positive values show polarization.

(D) Fraction of debates showing depolarization (y axis) as a function of group symmetry (x axis) for groups reaching consensus (black dots) and groups that did not reach consensus (gray dots). Error bars depict SE of proportion, and lines show the best-fitting linear functions. All the analyses reported in this figure were based on revised ratings that could only be collected in study 1, and hence, they were not replicated within this work.
See also Table S1.

Our study shows the conditions that maximize the probability of reaching consensus about moral judgements and that people might do so by implementing a mediation procedure. We believe that these results could be interesting to policy makers working on the development of deliberative polls [29, 30]. We hope that this work will inspire future research into the design of democratic innovations seeking to elicit consensus in controversial issues through deliberation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Context
  - Materials
  - Experimental procedure
  - First stage: Individual ratings
  - Second stage: Deliberation and group ratings
  - Third stage: Revised decisions
  - Control Experiment
  - Selection criteria for moral scenarios
  - Study 1: Moral scenarios
  - Study 2: Moral scenarios
  - Exclusion criterion
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Measuring the variability of opinions
  - Descriptive statistics of ratings
  - Modeling the probability of consensus
  - Modeling group ratings
  - Analysis of the control experiment
  - Analysis of revised private ratings
  - Attitude shifts
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2019.10.018.

ACKNOWLEDGMENTS

We thank Paloma Urtizberea and Florencia Gonzalez for assistance in data collection. M.S. was supported by the James McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition—Scholar Award (grant no. 220020334) and by Agencia Nacional de Promoción Científica y Tecnológica (Argentina)—Préstamo BID PICT (grant no. 2013-1653).

AUTHOR CONTRIBUTIONS

Conceptualization, J.N. and M.S.; Investigation, J.N. (control experiment), F.Á.H., J.M.G., P.A.G. (study 1), G.G., D.A., and M.S. (study 2); Writing – Original

Draft, J.N., D.A., and M.S.; Writing – Review & Editing, J.N. and M.S.; Visualization, J.N., J.M.G., and M.S.; Formal Analysis, J.N.; Resources, G.G.; Funding Acquisition, M.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 23, 2019
Revised: September 9, 2019
Accepted: October 10, 2019
Published: November 21, 2019

REFERENCES

1. Sunstein, C.R. (2018). Republic: Divided Democracy in the Age of Social Media (Princeton University Press).
2. Myers, D.G. (1975). Discussion-induced attitude polarization. Hum. Relat. *8*, 699–714.
3. Myers, D.G., and Bishop, G.D. (1970). Discussion effects on racial attitudes. Science *169*, 778–779.
4. Schkade, D., Sunstein, C.R., and Kahneman, D. (2000). Deliberating about dollars: the severity shift. Columbia Law Rev. 1139.
5. Myers, D.G., Wojcicki, S.B., and Aardema, B.S. (1977). Attitude comparison: is there ever a bandwagon effect? J. Appl. Soc. Psychol. *4*, 341–347.
6. Vinokur, A., and Burnstein, E. (1978). Depolarization of attitudes in groups. J. Pers. Soc. Psychol. *36*, 872–885.
7. Rahwan, I., Ramchurn, S.D., Jennings, N.R., Mcburney, P., Parsons, S., and Sonenberg, L. (2004). Argumentation-based negotiation. Knowl. Eng. Rev. *18*, 343–375.
8. McCubbin, P.R. (1988). Consensus through mediation: a case study of the Chesapeake Bay Land Use Roundtable and the Chesapeake Bay Preservation Act. J. Law Polit. *5*, 827–863.
9. Moscovici, S., and Zavalloni, M. (1969). The group as a polarizer of attitudes. J. Pers. Soc. Psychol. *12*, 125–135.
10. Fraser, C., Gouge, C., and Billig, M. (1971). Risky shifts, cautious shifts, and group polarization. Eur. J. Soc. Psychol. *1*, 7–30.
11. Hong, L.K. (1978). Risky shift and cautious shift: some direct evidence on the culture-value theory. Soc. Psychol. *41*, 342–346.
12. Mooney, C.Z., and Lee, M.-H. (1995). Legislative morality in the American states: the case of pre-Roe abortion regulation reform. Am. J. Polit. Sci. *39*, 599–627.
13. Lopez-Rosenfeld, M., Calero, C.I., Fernandez Slezak, D., Garbulsky, G., Bergman, M., Trevisan, M., and Sigman, M. (2015). Neglect in human communication: quantifying the cost of cell-phone interruptions in face to face dialogs. PLoS ONE *10*, e0125772.
14. Niella, T., Stier-Moses, N., and Sigman, M. (2016). Nudging cooperation in a crowd experiment. PLoS ONE *11*, e0147125.
15. Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., and Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. Nat. Hum. Behav. *2*, 126–132.
16. Skitka, L.J., Bauman, C.W., and Sargis, E.G. (2005). Moral conviction: another contributor to attitude strength or something more? J. Pers. Soc. Psychol. *88*, 895–917.
17. Lebreton, M., Abitbol, R., Daunizeau, J., and Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. Nat. Neurosci. *18*, 1159–1167.
18. Ais, J., Zylberberg, A., Barttfeld, P., and Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. Cognition *146*, 377–386.
19. Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P.E., and Bahrami, B. (2017). The idiosyncratic nature of confidence. Nat. Hum. Behav. *1*, 810–818.
20. Condorcet, M. (1785). Essai sur L'application de L'analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix (L'impremerie royale).
21. Susskind, L., and Cruikshank, J. (1987). Breaking the Impasse: Consensual Approaches to Resolving Public Disputes (Basic Books).
22. Couzin, I.D., Krause, J., Franks, N.R., and Levin, S.A. (2005). Effective leadership and decision-making in animal groups on the move. Nature *433*, 513–516.
23. Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., and Makse, H.A. (2010). Identification of influential spreaders in complex networks. Nat. Phys. *6*, 888–893.
24. van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M.N., and Wolpert, D.M. (2016). A common mechanism underlies changes of mind about decisions and confidence. eLife *5*, e12192.
25. Folke, T., Jacobsen, C., Fleming, S.M., and De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. Nat. Hum. Behav. *1*, 0002.
26. Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y.A., Shi, Z., Broberg, K., Safavi, S., Han, S., Nili Ahmadabadi, M., Frith, C.D., et al. (2015). Equality bias impairs collective decision-making across cultures. Proc. Natl. Acad. Sci. USA *112*, 3835–3840.
27. Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., and Frith, C.D. (2010). Optimally interacting minds. Science *329*, 1081–1085.
28. Myers, D.G., and Lamm, H. (1976). The group polarization phenomenon. Psychol. Bull. *83*, 602–627.
29. Bohman, J., and Rehg, W. (1997). Deliberative Democracy: Essays on Reason and Politics (MIT Press).
30. Fishkin, J.S., and Luskin, R.C. (2005). Experimenting with a democratic ideal: deliberative polling and public opinion. Acta Polit. *40*, 284–298.
31. Haidt, J., and Hersh, M.A. (2001). Sexual morality: the cultures and emotions of conservatives and liberals. J. Appl. Soc. Psychol. *31*, 191–221.
32. Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., and Cohen, J.D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science *293*, 2105–2108.
33. Mullen, E., and Skitka, L.J. (2006). Exploring the psychological underpinnings of the moral mandate effect: motivated reasoning, group differentiation, or anger? J. Pers. Soc. Psychol. *90*, 629–643.
34. Bostrom, N., and Yudkowsky, E. (2014). The ethics of artificial intelligence. In The Cambridge Handbook of Artificial Intelligence, W.K. Frankish, and W.M. Ramsey, eds. (Cambridge University Press), pp. 316–364.
35. Knoppers, B.M., and Chadwick, R. (2005). Human genetic research: emerging trends in ethics. Nat. Rev. Genet. *6*, 75–79.
36. Braun, M., Schickl, H., and Dabrock, P. (2018). Between Moral Hazard and Legal Uncertainty (Springer).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Data from Study 1 and Study 2	This paper; Mendeley Data	https://doi.org/10.17632/d87mx9s7f7.1#file-171d02a2-43cb-4750-b6cf-42b5e1665657
Software and Algorithms		
MATLAB codes to reproduce main figures	This paper; Mendeley Data	https://doi.org/10.17632/d87mx9s7f7.1#file-171d02a2-43cb-4750-b6cf-42b5e1665657

LEAD CONTACT AND MATERIALS AVAILABILITY

Requests for further information or materials associated with this study should be directed to the lead contact author, Dr. Mariano Sigman (msigman@utdt.edu). This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

In Study 1, a total of 5042 human participants performed the experiment (54.4% female, 6.1% aged [18–24] years, 28.2% aged [25–34] years, 52.1% aged [35–44] years, 11.9% aged [44-55] years, 1.7% aged over 60 years). In Study 2, a total of 1095 participants performed the experiment (37.1% female, 1.9% aged [18–24] years, 9.4% aged [25–34] years, 21.6% aged [35–44] years, 51.3% aged [44-55] years, 15.8% aged over 60 years). In the Control Experiment, 98 participants (62 females, mean age 19.8 years, range 18-27 years) performed the study. Data were completely anonymous, and participants gave written informed consent. The experimental protocol in this study was approved by the ethics committee of CEMIC (Centro de Educación Médica e Investigaciones Clínicas Norberto Quirno, Buenos Aires, Argentina).

METHOD DETAILS

Context

Study 1 took place during a TEDx event (http://www.tedxriodelaplata.org/) in Buenos Aires, Argentina and Study 2 took place during the 2017 TED conference (https://ted2017.ted.com/) in Vancouver, Canada. This was the fourth edition of an initiative called *TEDxperiments* (http://www.tedxriodelaplata.org/tedxperiments), aimed at constructing knowledge on human communication by performing behavioral experiments on large audiences. Previous editions studied the cost of interruptions on human interaction [13], the use of a competition bias in a "zero-sum fallacy" game [14], and the role of deliberation in the wisdom of crowds [15].

Materials

Research assistants handled one pen and one A4 paper to each participant. The A4 paper was folded on the long edge and had four pages. On page 1, participants were informed about their group number. The three stages of the experiment could be completed in pages 2, 3, and 4, respectively. On page 4, participants completed information about their age and gender.

Experimental procedure

The speakers (authors F.A.H., J.M.G., and P.A.G. in Study 1 and authors D.A. and M.S. in Study 2) announced that their section would consist in a behavioral experiment. Participants were informed that their participation was completely voluntary, and they could simply choose not to participate or withdraw their participation at any time. The structure of the experiment was identical to the procedure implemented in a previous experiment [15]. It consisted in three stages: individual ratings, deliberation and group ratings, and revised ratings.

First stage: Individual ratings

The speakers announced that participants would first listen to a set of moral scenarios, each of them involving a clear action. We asked participants to report how acceptable they found each action by providing a rating between 0 ("completely unacceptable") and 10 ("completely acceptable"). Participants were also asked to rate how confident they felt about their opinion between 0 ("completely uncertain") and 10 ("completely certain"). For each scenario, participants had 30 s to write down their answers. We made clear that decisions in this stage were individual.

Second stage: Deliberation and group ratings

In the second stage, we organized the crowd in groups and asked them to discuss these issues. First, they were instructed to find other members in their group according to a numerical code appearing in page 1. Each group had three participants, and all participants were seated next to each other in consecutive rows. Participants were instructed to deliberate about their opinions on each issue and, if possible, to reach consensus. In case they reached consensus, they needed to agree on a group rating and write it down. They were given a maximum of 2 minutes to reach consensus. If they could not agree on a group rating within that time window, they simply wrote down an 'X'. We explicitly made clear that there were no advantages nor disadvantages of reaching consensus. The speakers read all scenarios a second time and announced the moments in which time was over.

Third stage: Revised decisions

Finally, participants could revise their individual ratings and confidence. The speaker emphasized that this stage was, again, individual. We read all scenarios again and gave participants 30 s to write down their answers. This stage was present only in Study 1.

Control Experiment

Participants were undergraduate students recruited from two classrooms at Universidad Torcuato Di Tella. All of them were naive to the aim of the study and had never performed an experiment involving moral scenarios. To recreate as close as possible the experimental setting of Study 1, author JN read the four scenarios with the same words used in that experiment. Participants provided initial and revised ratings of acceptability and confidence, i.e., they performed stage 1 and stage 3 without deliberation or group decisions in between. Instead of performing stage 2, they performed an unrelated distractive task. This task consisted in five general-knowledge questions about the population of different cities. In all cases, the instruction was the same "Which of these two cities has the largest population"?. In all cases, the question was framed as a two-alternative choice between 1) Istanbul or Moscow, 2) Delhi or Sao Paulo, 3) Mexico City or London, 4) Hong Kong or Bogota, 5) Brasilia or Fortaleza. Participants responded individually to these questions. Right after providing their responses, we disclosed the correct answers to these questions (1: Istanbul; 2: Delhi; 3: Mexico City; 4: Hong Kong; 5: Fortaleza). Finally, we read again all moral scenarios and asked participants to provide revised ratings for their acceptability and confidence.

Selection criteria for moral scenarios

Our setting allowed us to have access to large sample sizes at the cost of having less control over a series of variables such as reaction times, contents of deliberation, and group compositions. Another aspect that we considered when designing these experiments is that we had a rather short slot (15 minutes at maximum) to run the entire experiment. This restricted the number of scenarios we could test to only four in Study 1 and two in Study 2. With that constraint in mind, we selected scenarios using three criteria. First, we wanted them to address a diverse set of issues so that we knew that our results would not depend on the specific details of a given scenario. Second, we wanted the scenarios to be grounded on previous literature of moral psychology. Third, we wanted these scenarios to generate a wide range of opinions in our population of study since our main aim was to identify the effect of conversations on people holding different views about these scenarios (Figure S1). Below, we describe each scenario and their relationship to previous work in moral psychology.

Study 1: Moral scenarios

There were four scenarios in Study 1. The first one (SIBLINGS) was: "Two siblings were home alone and decided to have sex just once. She is on the contraceptive pill and he used a condom. Both enjoyed it, never did it again, and kept it a secret. Their behavior was …." This scenario is an adaptation of the vignette "Julie and Mark" previously tested in the study of sexual morality and the phenomenon of moral dumbfounding (e.g., [31]).

The second scenario ("INVASION") was: "A family was trapped at home during a military invasion. To escape from the invading soldiers, both parents and their four children hided and took refuge in the basement. One of the children, a baby, suddenly started crying. Both parents decided to cover the baby's nose and mouth and provoked its death, since this was the only way to prevent the entire family from being discovered and killed. Their behavior was…." It was adapted from a high-conflict scenario (i.e., the "crying baby" dilemma) also tested in previous studies (e.g., [32]).

The third scenario ("ABORTION") was: "A woman and her boyfriend had sex and she got pregnant. Both live in a country where the voluntary interruption of pregnancy is legal and completely safe for the woman's health. During the 20th week of gestation, they decided to carry out the procedure. Their behavior was…." This scenario reflects a long-standing public debate about whether the voluntary termination of pregnancy is morally acceptable. This topic has been framed in many different ways in the literature including scenarios similar to the one tested here (e.g., [33]) and also through questions directly measuring moral attitudes toward the legalization of abortion (e.g., [16]).

The fourth and last scenario (A.I.) was: "A laboratory developed an artificial intelligence that is indistinguishable from human intelligence. The protocol states that every night researchers should delete the program and reboot it in a different version. According to the researchers, the program can maintain fluid conversation, report subjective states, and appears to be self-aware. One day, the program reports to be afraid and asks the researchers to please not be deleted. Researchers decided to follow the protocol as usual and deleted the program. Their behavior was…." This scenario digs into the debate of robot rights. This issue has led to an entire field

in the study of morality (see, for example [34],). The scenario tests people's moral attitudes toward the principle of substrate non-discrimination ("If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status") and the principle of ontogeny non-discrimination ("If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status").

Study 2: Moral scenarios

In Study 2, we asked participants to consider two scenarios. The first one was identical to the fourth and last scenario in Study 1 (A.I.). The second scenario (GENES) was: "*A company offers a service that takes a fertilized egg and produces millions of embryos with slight genetic variations. This allows parents to select their child's height, eye color, intelligence, social competence, and other non-health related features. The company's behavior is*…." This scenario raises the question of whether editing the human genome for non-health-related issues is morally acceptable. Transforming the genetic configuration of a human embryo has been previously argued to be ethically questionable (e.g., [35]) and previous research has examined and listed different arguments in favor of regulating and/or prohibiting this activity (for a map of ethical arguments, see [36]).

Exclusion criterion

At the end of the talk, we collected the answer sheets as participants exited the auditorium. After the event, data-entry clerks digitalized these data using a keyboard. Some of these groups had incomplete data due to at least one missing participant; we collected complete data from 1096 groups in Study 1 and 194 groups in Study 2. Following the same procedure used in a previous study [15], we removed from the analysis all data proceeding from incomplete groups. This is because there are several reasons why we could have incomplete data in a given group. For example, a seat might have been empty at the time of the experiment resulting in a smaller group. Alternatively, it could also mean that someone who took part in the experiment simply forgot to return her/his answer sheet. Because it is impossible for us to distinguish between these alternatives after the event was finished, we adopted a strict conservative criterion, excluding from the analyses all participants proceeding from groups with missing data. This was the only data exclusion criterion used in this work. All findings reported here are based on all complete groups with 3288 individuals in Study 1 (54.6% female, 10.6% aged [18–24] years, 30.8% aged [25–34] years, 45.0% aged [35–44] years, 12.1% aged [44-55] years, 1.7% aged over 60 years) and 582 in Study 2 (36.9% female, 2.5% aged [18–24] years, 8.9% aged [25–34] years, 21.1% aged [35–44] years, 52.4% aged [44-55] years, 14.9% aged over 60 years).

QUANTIFICATION AND STATISTICAL ANALYSIS

Measuring the variability of opinions

One possible concern about the variability of opinions in our sample is that groups were formed by physical proximity. Because participants sitting in neighboring seats typically know each other, they could also potentially share similar opinions. To reduce the chances of this happening, we asked people to organize in groups across different rows. In previous studies [13, 14], we found that this manipulation effectively led to many interactions between people who did not know each other from before the event. To confirm that the diversity of opinions in the group was random relative to the initial ratings provided by the crowd, we performed a permutation analysis. We created surrogated data by randomly shuffling the initial ratings 1,000 times. We then computed the range of opinions for each group and each simulation. From these 1,000 simulations we chose the one with median range of opinions and compared it to our empirical observations. We found that the surrogated and real data had overlapping distributions (95% CI [2.43-8.84] for the surrogated data, 95% CI [2.07-8.62] for the actual data) with a negligible effect size (Cohen's $d = 0.17$), suggesting that the observed diversity of ratings was indistinguishable from random.

Descriptive statistics of ratings

The first scenario of Study 1 (SIBLINGS) produced a distribution ratings with the following properties: initial ratings, *mean* = 2.3, *s.d.* = 3.1., *Q1* = 0, *median* = 0, *Q3* = 5; group ratings, *mean* = 2.5, *s.d.* = 2.7, *Q1* = 0, *median* = 2, *Q3* = 5; revised ratings: *mean* = 3.0, *s.d.* = 3.4, *Q1* = 0, *median* = 2, *Q3* = 5.

The second scenario of Study 1 (INVASION) produced a distribution ratings with the following properties: initial ratings, *mean* = 4.1, *s.d.* = 3.7., *Q1* = 0, *median* = 4, *Q3* = 7; group ratings, *mean* = 4.5, *s.d.* = 2.8, *Q1* = 2, *median* = 5, *Q3* = 6; revised ratings: *mean* = 4.6, *s.d.* = 3.5, *Q1* = 1, *median* = 5, *Q3* = 8.

The third scenario of Study 1 (ABORTION) produced a distribution ratings with the following properties: initial ratings, *mean* = 5.0, *s.d.* = 3.9., *Q1* = 0, *median* = 5, *Q3* = 9; group ratings, *mean* = 4.6, *s.d.* = 3.5, *Q1* = 1, *median* = 5, *Q3* = 8; revised ratings: *mean* = 4.7, *s.d.* = 4.2, *Q1* = 0, *median* = 5, *Q3* = 8.

The fourth scenario of Study 1 (AI) produced a distribution ratings with the following properties: initial ratings, *mean* = 6.6, *s.d.* = 3.4., *Q1* = 4, *median* = 8, *Q3* = 10; group ratings, *mean* = 7.2, *s.d.* = 2.9, *Q1* = 5, *median* = 8, *Q3* = 10; revised ratings: *mean* = 6.8, *s.d.* = 3.7, *Q1* = 5, *median* = 8, *Q3* = 10.

The first scenario of Study 2 (AI) produced a distribution ratings with the following properties: initial ratings, *mean* = 7.2, *s.d.* = 2.9, *Q1* = 6, *median* = 8, *Q3* = 10; group ratings, *mean* = 8.1, *s.d.* = 1.9, *Q1* = 8, *median* = 8, *Q3* = 9.

The second scenario of Study 2 (GENES) produced a distribution ratings with the following properties: initial ratings, *mean* = 3.6, *s.d.* = 3.2, *Q1* = 1, *median* = 3, *Q3* = 6; group ratings, *mean* = 2.9, *s.d.* = 2.7, *Q1* = 1, *median* = 2, *Q3* = 5.

Modeling the probability of consensus

We estimated the likelihood of reaching consensus by measuring the fraction of groups that provided a group rating. To study the factors underlying this probability we fitted logistic models with different predictor variables. For example, the line in Figure 2A is based on the best-fitting logistic model with range $\delta = r_{max} - r_{min}$ as predictor variable.

To study the effect of symmetry on consensus (Figure 2B), we measured the absolute non-parametric skewness,

$$\gamma_1 = \frac{|\text{mean}(\mathbf{r}) - r_{med}|}{\text{std}(\mathbf{r})}, \quad \text{(Equation 1)}$$

where **r** is a vector with all three ratings in the group and r_{med} is the intermediate opinion.

We then defined a variable called "symmetry" as

$$S = \max(\gamma_1) - \gamma_1, \quad \text{(Equation 2)}$$

where, in our data, $\max(\gamma_1) = .484$ is the absolute non-parametric skewness of the most skewed groups (i.e., those with ratings 0-1-10 or 0-9-10). Note that, for those groups, *S* is equal to zero, whereas for completely symmetric configurations (e.g., 0-5-10, 1-5-9, etc.) it takes its maximum value $(S = .484)$. The complete logistic model that accounts for the effect of range and symmetry was

$$\log\left(\frac{p(\text{consensus})}{1 - p(\text{consensus})}\right) = \alpha + \beta_\delta \cdot \delta + \beta_S \cdot S, \quad \text{(Equation 3)}$$

where α is an intercept, and β_δ and β_S are respectively the weights of range and symmetry. All findings reported in our main text and figures are based on mixed-effects models with random intercepts for each scenario. Supplementary figures are based on models fitted for each scenario separately. To evaluate the effect of range and symmetry across all groups with different ranges we fitted Equation 3 to our data.

To quantify the interaction between the effects of confidence and disagreement on *p*(consensus), we performed an additional logistic regression with δ and confidence i as predictor variables. We found that the effect of range on consensus was substantially reduced when groups had an intermediate person with high confidence (interaction between range and confidence distribution, Study 1: $\beta = 0.019 \pm 0.004$, p = 0.001; Study 2: $\beta = 0.05 \pm 0.01$, p = 0.002).

To visualize the effect of different variables (e.g., symmetry, confidence, gender, age, etc.) on consensus, we plot in Figures 2 and S3 the change in the probability of reaching consensus relative to baseline, i.e., Δp(consensus). This quantity measures the probability of reaching consensus for a given condition minus the probability of reaching consensus in general, without considering any specific condition. Given an experiment with *N* groups, let assume that *n* of them reached consensus. Assuming that such experiment has a condition which applies to N_c groups, and that *c* of them reached consensus, then we estimate Δp as

$$\Delta p(\text{consensus}) = \frac{c}{N_c} - \frac{n}{N}. \quad \text{(Equation 4)}$$

Finally, we show in the Results and Discussion that the range and skewness of group ratings modulates the probability of reaching consensus. Since these two quantities are highly correlated with the second and third moments of the distribution of ratings, one might wonder if the fourth moment (kurtosis) of ratings also influenced the likelihood of consensus. However, we did not find any evidence that the sample non-parametric kurtosis modulated the probability of reaching consensus (Study 1: $\beta = -.07 \pm .29$, $p = .79$, Study 2: $\beta = .05 \pm .99$, $p = .95$).

Modeling group ratings

To study the relationship between collective ratings (c) and the initial ratings of the three individuals in the group, we performed the following multivariate regression,

$$c = \alpha + \beta_{\text{min}} \cdot r_{\text{min}} + \beta_{\text{med}} \cdot r_{\text{med}} + \beta_{\text{max}} \cdot r_{\text{max}}, \quad \text{(Equation 5)}$$

where r_{min}, r_{med}, and r_{max} are the ratings of participants p_{min}, p_{med}, and p_{max}, respectively, and α is an intercept. The weights β_{min}, β_{med}, and β_{max} and their corresponding SEM are displayed in Figures 2E and S5. A BIC analysis revealed that including the three weights appearing in Equation 4 was preferable to having a single coefficient modulating the average rating in the group ($\Delta BIC = 27$, $\Delta df = 2$, log-likelihood ratio test, $L = 42.9$, p = 10^{-10}).

Analysis of the control experiment

We performed two positive controls (i.e., sanity checks) to verify that the Control Experiment produced revised ratings that were consistent with a second sample taken from probability distributions. First, we observed regression to the mean: the opinion shift was negatively correlated with the initial rating ($r = -.15$, p = 0.002). Low initial values of acceptability shifted toward higher ratings (shift for participants with initial ratings lower than 5: *mean* ± SEM = 0.27 ± 0.10, t test against 0, $t(192) = 2.7$, p = 0.007) and vice versa, high initials values of acceptability shifted toward lower ratings (shift for participants with initial ratings higher

than 5: *mean* ± SEM = -.15 ± 0.08, t test against 0, $t(192) = 2.1$, p = 0.04). We also found that the absolute shift between the revised and initial ratings was negatively correlated with confidence ($r = -.14$, p = 0.005), which is consistent with the idea that high values of confidence are associated with tighter distributions of beliefs [17].

We then performed a different control analysis involving groups with different symmetry. We reasoned that if people indeed resampled from their distribution of beliefs, one should see a reduction in the range of opinions of virtual groups. However, in symmetric groups, this reduction should be more likely to happen than in asymmetric groups because only two of the three individuals need to regress toward the mean (the third is already in the middle). So, statistically speaking, convergence is expected to be more likely in symmetric groups than in non-symmetric groups even in the absence of interaction. Consistent with this prediction, symmetric virtual groups that started with maximum range showed greater convergence than asymmetric groups (reduction in range for virtual triads, high symmetry: ΔR = 0.79 ± 0.01, low symmetry: ΔR = 0.70 ± 0.02, t(27984) = 3.5, $p = 10^{-4}$). This effect was small (Cohen's d = 0.09) but it is still consistent with what we observed in Figure 2C -if individuals resample their distributions of beliefs, symmetric groups are more likely to converge than asymmetric groups. This suggests that the effect of symmetry on consensus could be, at least partially, explained by resampling.

However, our findings also indicate that several of our results could not be replicated using virtual groups (e.g., Figure 4A and Table S1). For example, we did not find any evidence for greater convergence if virtual groups had a confident intermediate participant ($t(27984) = 1.3$, p = 0.17), suggesting that the effect of confidence on consensus for moderate individuals (Figure 2D) is only present in interacting groups. Second, we found that extreme participants rarely changed their mind in the absence of social influence while their probability to do so after deliberation was substantially higher (without interaction: 12.8%, with interaction: 45.5%, Chi-square test for equal proportions, $\chi^2(1) = 61.9$, $p = 10^{-15}$). This suggests that resampling alone cannot explain why so many individuals changed their mind. Altogether, these two findings indicate that the convergence to intermediate ratings could not be simply be explained by regression to the mean.

Analysis of revised private ratings

Following deliberation, participants provided revised acceptability and confidence ratings. Due to time constrains, revised ratings could not be acquired in Study 2, so we focused this analysis in the data of Study 1. We observed that, after social influence, there was a significant increase in confidence, both for participants who reached consensus (Δc = 0.52 ± 0.02, $t(9497) = 19.8$, $p = 10^{-85}$) and for participants who did not reach consensus (Δc = 0.21 ± 0.04, $t(3548) = 5.1$, $p = 10^{-7}$). However, this effect was larger for those who succeeded at reaching consensus (unpaired t test, $t(13045) = 6.2$, $p = 10^{-10}$). This is consistent with the idea that revised ratings were based on both private and social information.

We used these data to probe a prediction of the mediation hypothesis. We tested if revised ratings at one extreme (i.e., RE, revised r_{min} or r_{max}) were partially explained by the initial ratings at the other extreme (i.e., IO, initial r_{max} or r_{min}, respectively) after controlling by the initial ratings at that extreme (i.e., IE, initial r_{min} or r_{max}, respectively). To this end, we performed the following bivariate linear regression

$$RE = \alpha + \beta_E \cdot IE + \beta_O \cdot IO, \qquad \text{(Equation 6)}$$

where α is an intercept and the "relative weights" in Figure 4A are β_E and β_O normalized so that both numbers add to 1. This analysis was performed separately for virtual triads from the Control Experiment (NI condition in Figure 4A), real groups that did not reach consensus (NC condition in Figure 4A), and real groups that reached consensus (C condition in Figure 4A). The residuals of RE and IO plotted in Figure 3B are the result of performing a univariate linear regression of RE against IE (i.e., residuals of RE) and of IO against IE (i.e., residuals of IO) for condition C only. The correlation between these two residuals represent the partial correlation between RE and IO after controlling by IE.

To formally test for a difference between conditions, we pooled together the data from the Control Experiment and Study 1 and modeled RE as a linear combination of IE and IO while adding three dummy variables. The first one indicates whether the participant belonged to a group where there was interaction without consensus (X_{int}), the second one indicates whether that group reached consensus (X_{cons}) and the third one indicates whether there was a high-confident intermediate participant in that group (X_{conf}). Overall, the considered linear model is:

$$\begin{aligned} RE = {} & \beta_0 + (\beta_1\, IE + \beta_2\, IO + \beta_3\, IE \cdot X_{conf} + \beta_4\, IO \cdot X_{conf}) + \\ & (\beta_9\, IE + \beta_{10}\, IO + \beta_{11}\, IE \cdot X_{conf} + \beta_{12}\, IO \cdot X_{conf}) \cdot X_{cons} \\ & (\beta_5\, IE + \beta_6\, IO + \beta_7\, IE \cdot X_{conf} + \beta_8\, IO \cdot X_{conf}) \cdot X_{int} + \end{aligned} \qquad \text{(Equation 7)}$$

This model with dummy variables decomposes the effect of IE and IO on RE using the control condition as reference group. The best-fitting estimates of the model described in Equation [6] are reported in Table S1.

Attitude shifts

To evaluate the presence of group polarization or depolarization, we defined a variable called "attitude shift" (ζ) as:

$$\zeta = (\bar{r}_{post} - \bar{r}_{pre}) \frac{(\bar{r}_{pre} - 5)}{|\bar{r}_{pre} - 5|}, \qquad \text{(Equation 8)}$$

where $\bar{r}_{pre}$ and $\bar{r}_{post}$ respectively are the average pre-interaction and post-interaction ratings in a given group. The second term in the equation is a sign function relative to the mid-point of the scale (i.e., from 0 to 10), and we define $\zeta = 0$ if $\bar{r}_{pre} = 5$. This quantity was calculated for each group and scenario with positive values indicating a shift in the average rating away from the mid-point of the scale (group polarization) and negative values meaning a shift toward the mid-point of the scale (group depolarization). Figure 4C display a histogram of attitude shifts across all debates.

DATA AND CODE AVAILABILITY

The data and codes to reproduce our main figures are available in https://doi.org/10.17632/d87mx9s7f7.1#file-171d02a2-43cb-4750-b6cf-42b5e1665657.