Anger, evidence, & trending opinions: we trust consensus when we believe it reflects genuine persuasion

Emory Richardson & Frank C. Keil

Yale University

Address for correspondence : Emory Richardson
Department of Psychology
Yale University
Box 208205
New Haven, CT, 06520-8205
Email : emory.richardson@yale.edu
Word Count : 3,328 (Main Text); 3, 844 (Methods & Results); 231 (Abstract)
Date : 22.8.25_v1.0

Acknowledgments

This research was supported by NSF grant DRL 1561143 awarded to Frank C. Keil.

Author Contributions

ER developed the study concept; ER designed the experiments with input from FCK; ER collected and analyzed the data. ER drafted the manuscript, and FCK provided critical revisions. All authors approved the final version of the manuscript for submission.

Abstract

Social learners frequently treat the consensus judgment of multiple informants as evidence favoring one option over another. However, in populations of social learners, learners' informants are also learning from each other, making consensus a product of complex informational dependencies between informants. These interdependencies can make consensus more accurate when informants are reliably able to identify accurate testimony, or less accurate when they are not. But do learners consider whether their sources are reliable judges of each others' testimony when deciding whether to trust their consensus judgment? Here, participants in four experiments rated their confidence in a consensus judgment, and then saw that consensus change when one faction "converted" informants from an opposing faction during a meeting (e.g., consensus grew from 6v4 to 9v1 or fell from 9v1 to 6v4), and again rated their confidence. Critically, we manipulate affective cues during the meeting: the focal faction attempts to convert the opposing faction either by shouting angrily at them or expressing surprise. We find an asymmetric effect of affective cues on trust in consensus: while confidence in the judgment of surprised factions *rises* when they *gain* endorsers and *falls* when they *lose* endorsers, confidence in angry factions *falls* when they *lose* endorsers, but does *not rise* even when they *gain* endorsers. Trust in non-independent consensus may therefore depend on reasoning about the extent to which informants' belief-forming processes are reliable.

Introduction

Imagine ten people individually trying to solve a problem: most favor one answer, but some favor a second. Which answer would you expect to be more accurate? Like many species capable of learning from each other, humans tend to treat majority judgment as more reliable (Claidière & Whiten, 2012). Now suppose that when the ten people meet, one faction "converts" several people from the other faction after expressing surprise at their answer. Does your confidence in the consensus change? It seems reasonable to infer that the converts were convinced by the evidence provided in the discussion. Thus, even though you were unable to evaluate their evidence directly, you may interpret their conversions as evidence-by-proxy for the solutions' accuracy. But what if instead, the converts had changed their answer after the emotional faction had shouted angrily at them: would your confidence still change? You may worry that the "converts" were simply intimidated into conformity instead of being convinced by evidence. Indeed, you may even feel *less* confident in the angry faction's answer, even though more people endorse it than before the meeting.

We expect readers to find these inferences commonsensical. But, we suggest that they also reveal something important about how people evaluate socially-transmitted information. Since several informants' judgments changed as a result of the meeting, the post-meeting consensus no longer consists of independent judgments. If your confidence changed along with the change in consensus, then it follows that you have treated your informants' collective evaluations of higher-order evidence (such as each others' testimony, reputation, or reasoning) as *more* reliable than their independent evaluations of firsthand evidence. Of course, this may or may not have been a wise strategy: you made this inference even though you yourself had no way to evaluate the evidence "on the merits". But, it wasn't a blind strategy: if the angry shouting led you to write off the change of consensus as an artifact of intimidation, you displayed some epistemic vigilance. Again, your skepticism may or may not have been appropriate: after all, the information available to you as an observer provided no way of knowing if the anger was justified. However, regardless of whether or not your inferences were justified, they may have implications not only for your ability to learn from consensus and reason about others' beliefs, but also for the reliability of consensus as a measure of collective intelligence. We return to these points in the general discussion.

Consensus-based learning strategies are a double-edged sword. On the one hand, as long as informants' judgments are statistically independent from one another, increasing the number of informants makes the strength of consensus increasingly likely to reflect the average competence of the crowd (Condorcet, 1994/1785; Dietrich & Spiekermann, 2013). Roughly speaking, consensus in a competent crowd amplifies accurate judgments, but consensus in an

incompetent crowd amplifies inaccurate judgments, simply because uncorrelated errors cancel each other out. On the other hand, if individuals *are* able to influence each other, their own ability to solve a problem alone matters much less than their ability to recognize when someone *else's* judgment is accurate: consensus can amplify the judgment of the "best member" when the average informant is able to recognize it, but can also drown out the best member when they are not (Laan, Madirolas, & de Polavieja, 2017; Kao, Miller, Torney, Hartnett, & Couzin 2014; Mercier & Claidière, 2022). In other words, regardless of whether or not informants' judgments are statistically independent, learning from consensus will require learners to make some judicious inferences about their informants' competence.

Informational dependencies complicate consensus-based learning strategies in the real world in two ways. First, our informants are typically only partially informed as individuals. Thus, even competent informants will sometimes make mistakes simply because they lack critical information. But, sharing information could have helped them avoid these mistakes, as long as their strategies for learning from each other are reliable. Second, widespread social learning in the biological world makes assuming that informants' judgments are independent notoriously implausible; even shared perceptual and cognitive biases are sufficient to compromise statistical independence, to say nothing of shared culture, environment, or motivated reasoning. Thus, real world consensus is almost certain to involve some redundancies — informants whose judgments depend on the same information. But again, whether such a consensus is more or less accurate than an "independent" consensus depends on the reliability of learners' strategies for evaluating the information they share with each other. While ineffective strategies can corrupt consensus, effective strategies can make it more accurate (Toyokawa, Whalen, & Laland, 2019). How learners themselves estimate consensus' reliability is therefore fundamental to understanding how reliable consensus will actually be.

One approach to assessing consensus may be to evaluate the reliability of the processes that produced it (Dunn, 2019; Goldman, 2014). This can include evaluating how robust the *evidence itself* is to the kinds of lost-in-translation distortions that are intrinsic to testimony (Bartlett, 1932), but also evaluating how robust our *informants themselves* are to the kinds of social pressures that might distort their testimony (or improve it). For instance, while individuals are normally reliable judges of their own perceptual experience, perception is intrinsically firsthand: details from the testimony of a specific eyewitness that are lost in telephone-style transmission chains typically cannot be recovered downstream by someone else. In contrast, details from a forecast based on a mathematical formula may also be lost in transmission, but inconsistencies are not only much more detectable, conclusions can be corrected or even improved by any mathematically competent learner. Of course, most of what we learn will

neither be as vulnerable as eyewitness testimony or as robust as mathematical equalities. Yet, just as collective judgments are less robust when evidence is more vulnerable to random decay, they can also be corrupted by informants' motivational and cognitive biases. This can occur even when the informants themselves believe what they're saying. Moreover, cultural, linguistic, and ideological divides ensure that informants' judgments will not always be truth-seeking. In short, considering the processes that produced consensus may lead you to prefer that your informants give maximally independent judgments in some cases, but in other cases, tradeoffs may favor the reverse.

Adults and even children as young as 6 appear to distinguish between contexts in which there is more or less chance that informants' judgments or their evidence *could* be distorted — or improved — by social influences (Miton & Mercier, 2019; Richardson & Keil, 2022; Desai, Xie, & Hayes, 2022). Here, we examine people's evaluations of social influence more directly, by presenting them with informants who actually *reverse* their judgments after meeting with peers who disagree, in a context in which there is potentially something to be learned from discussion. Critically, we use affective cues to manipulate the reliability of the social process that produced the final consensus.

Thin-slice evaluations of two orthogonal dimensions, warmth and competence, account for most of the variance in person-perception (Fiske, Cuddy, & Glick, 2007), and people display a "benevolence bias" leading to greater trust in "nice" informants. Asking children to choose between testimony from mean-but-smart and nice-but-ignorant informants suggests that learning to trust competence over benevolence emerges later in development for both moral judgments and matters of scientific fact (Danovitch & Keil, 2007; Landrum, Mills, & Johnston, 2013; Johnston, Mills, & Landrum, 2015). Yet, despite a large literature examining the effects of emotion on individual decision making and interpersonal negotiations, recent reviews note a dearth of studies asking how emotion influences *group* processes and perceptions of groups (Lerner et al., 2015; but see Goldenberg, Saguy, & Halperin, 2014). Existing research has focused on participants' reactions to political violence and moral outrage (Simpson et al., 2018; Teixeira et al., 2020; Steinert-Threlkeld et al., 2022; Brady et al., 2017). However, these contexts make it difficult to separate outrage that spreads simply because people who share political sympathies are outraged by similar things from outrage that spreads because people treat outrage as an epistemic signal (cf. Burton, Cruz, & Hahn, 2021).

To the extent that our affective responses to social interactions carry information about our evaluations of *others*' beliefs and desires as well as our own, observers could use those responses to evaluate changes in consensus, even without any other information about the merits of the opposing beliefs. If learners expect group members to respond to benevolence and hostility as

they themselves do, affective cues like anger and surprise could help learners evaluate social influences on consensus judgements by suggesting whether consensus is *genuine* or *forced*. If affective signals suggest that group members changed their opinion because they were genuinely convinced by the evidence presented in the discussion, observer confidence may shift to reflect the final degree of consensus. If affective signals suggest that group members were *forced* to conform instead of being convinced by evidence, then observer confidence may instead track the degree of consensus prior to the shouting or even turn against the consensus.

General Method

In each experiment, participants are told about students learning to make rockets at a science camp (Fig 1). Each student needs to design their own rocket, but they have opportunity to discuss in groups of 10 what they've learned from the engineers at camp. Participants first see the students' *PreMeeting* opinions (what each individual thought before finding out what others thought), which reveals just two opinion-based factions in each group. However, answers are color-coded; thus, participants only know the "vote" (i.e., how many endorsements each option received) but nothing about the content of the answers. This initial vote share is always either 6v4 or 9v1. Participants are then asked for a *PreMeeting* rating of which opinion is more likely to be accurate. Next, during the group meeting, 3 students "convert" after the emotional faction

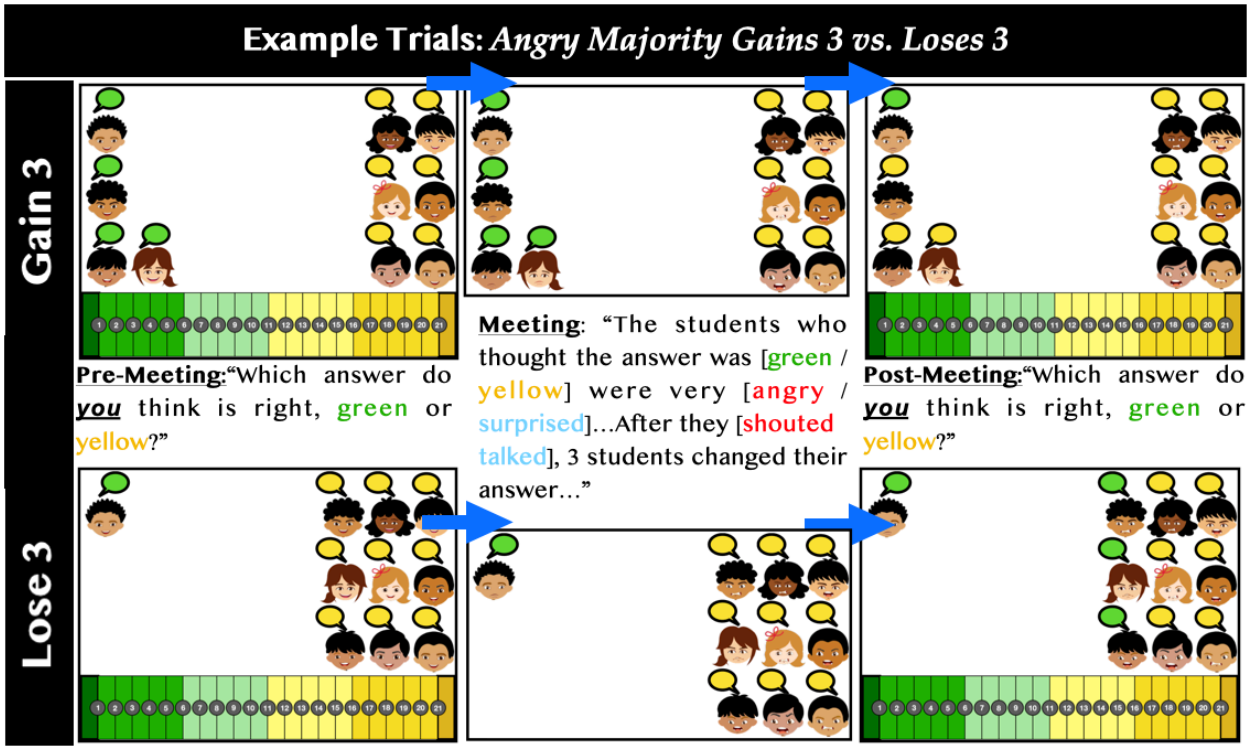


Figure 1. Example procedure for *Anger* trials (Exp 1: *MajorityGain* vs *MajorityLose)*. After being told the possible topics and seeing the initial "vote", participants were told that some students changed their answer during the meeting, and shown the final "vote".

either shouts at or talks with the opposing faction, and participants make a *PostMeeting* accuracy rating. Each participant completes two pre-post trials: one in which the emotional faction expresses *Anger*, and one in which the emotional faction expresses *Surprise*. Ratings are made on a 1-21 confidence scale. Participants are asked to explain their ratings after each trial. Though we made no explicit hypotheses about the precise frequency of different kinds of explanations, explanations were coded post-hoc by a research assistant, and we present descriptive results in supplemental materials.

In Experiments 1-2, the EmotionalFaction is either the initial Majority or initial Minority, and they either *Gain3* or *Lose3* endorsers. In Experiment 3, the initial 4v6 Minority always Gains3 endorsers, flipping "control" of the vote to become the new majority. In Experiment 4, we dissociate informants' feeling of anger from their rhetorical use of anger by manipulating whether the angry faction *Expresses* their anger (i.e., shouts at the opposing faction, as in Exps 1-3) or *Restrains* their anger and simply talks with them; this tests whether participants penalize the use of anger as a means of influence, rather than simply mistrusting the angry faction's beliefs because of their affective state. Since we expected similar effect sizes for all experiments, the sample sized determined by power analysis for Experiment 1 was used for all experiments. For all experiments, preregistrations of hypotheses and methods, as well as all materials, data, power analysis, and analysis scripts can be found at the first author's OSF repository. (https://osf.io/t45am/?view_only=0619aa1ffdf84ac29575733c77316ecc). All studies were approved by Yale University's IRB.

Experiment 1

In Experiment 1, the majority either shouted angrily at the minority or was surprised and talked with them, after which 3 students converted to the other side. In the Gain3 condition, participants saw the majority grow from 6v4-to-9v1 after the meeting; in the Lose3 condition, participants saw the majority shrink from 9v1-to-6v4.

Participants. We recruited 80 participants from MTurk; due to a platform error, one failed to submit data, leaving n=40 in MajorityGain & n=39 in MajorityLoss. Each participant made a pre-meeting and post-meeting in an Anger trial and in a Surprise trial (order counterbalanced).

Materials & Procedure. Participants were told that an engineer at a science camp had taught students about how to design a model rocket that would fly well, using a set of materials and computer code. Each student's goal was to design a rocket that would fly as high as possible; but they would first have to talk together about how to build a rocket that would fly the highest. Students were depicted as clipart faces rather than real models in order to enable us to (A) show schematic changes in facial expression and gaze direction during the discussion using simple

photoshop edits while avoiding face-specific competence inferences (e.g., Todorov et al., 2015), and (B) reuse a similar method with children in future work. Participants first saw "*what each student thought was best*" as a color-coded speech-bubble, with the faces looking straight ahead and smiling; color-coding ensures that while participants' confidence ratings could rely on common social learning heuristics like majority rule or opinion dynamics, they would be unable to evaluate the content of the answers firsthand. After participants' made their Pre-Meeting rating, they were told that, e.g., "*students who thought the answer was blue were* ***very angry*** *that the other students thought the answer was green",* at which point the students in the majority faction were shown to be looking over at the opposing faction and shouting angrily, with the minority looking submissive and frowning silently. Participants were then told that "*after the students who answered blue* ***shouted at*** *the students who answered green, some of the students changed their answer*", at which point 3 students speech-bubbles changed color (e.g., in the Lose3 condition, this meant that three students from the majority who had been shouting flipped to endorse the minority answer; in the Gain3 condition, three students from the minority flipped to endorse the majority answer). Participants then made their Post-Meeting ratings and were asked to write 1-2 sentences explaining their ratings. Participants then completed a second trial with the other emotion. Note that the anger and surprise trials used different colors (e.g., blue & green or purple & yellow) in order to prevent participants from referencing 'global' consensus across the two trials (e.g., inferring that one color was more likely to be to be optimal regardless of emotion because it was endorsed by 6v4 majority in both groups).

Results. Our primary question was whether people's reactions to the discussions differed by the direction of *Change* and *Emotion*. We computed a difference score by subtracting the Pre-meeting ratings from the Post-meeting ratings for *Anger* and *Surprise* trials, producing negative numbers for a decrease in confidence in the emotional faction and positive numbers for an increase in confidence. In each condition, the shift in confidence was consistent with our predictions. When the majority faction was *Angry* at the minority view, confidence in majority accuracy *fell* significantly when the majority *Lost* three endorsers to the minority view, but *did not rise* significantly even when they *Gained* three endorsers from the minority view (*Anger*: β_{Lose3} = -5.13, SE = .59, p<.001; β_{Gain3} = 0.075, SE = .59, p = .90). In contrast, when the majority faction was *Surprised* at the minority view, confidence in majority accuracy *fell* significantly when the majority *Lost* three endorsers to the minority view and *rose* significantly when they *Gained* three endorsers from the minority view (*Surprise*: β_{Lose3} = -4.23, SE = .59, p<.001; β_{Gain3} = 3.65, SE = .59, p<.001).

We next asked whether participants' ratings favored majorities. Past work has suggested that people trust majority over minority judgement, and that confidence increases with the size of

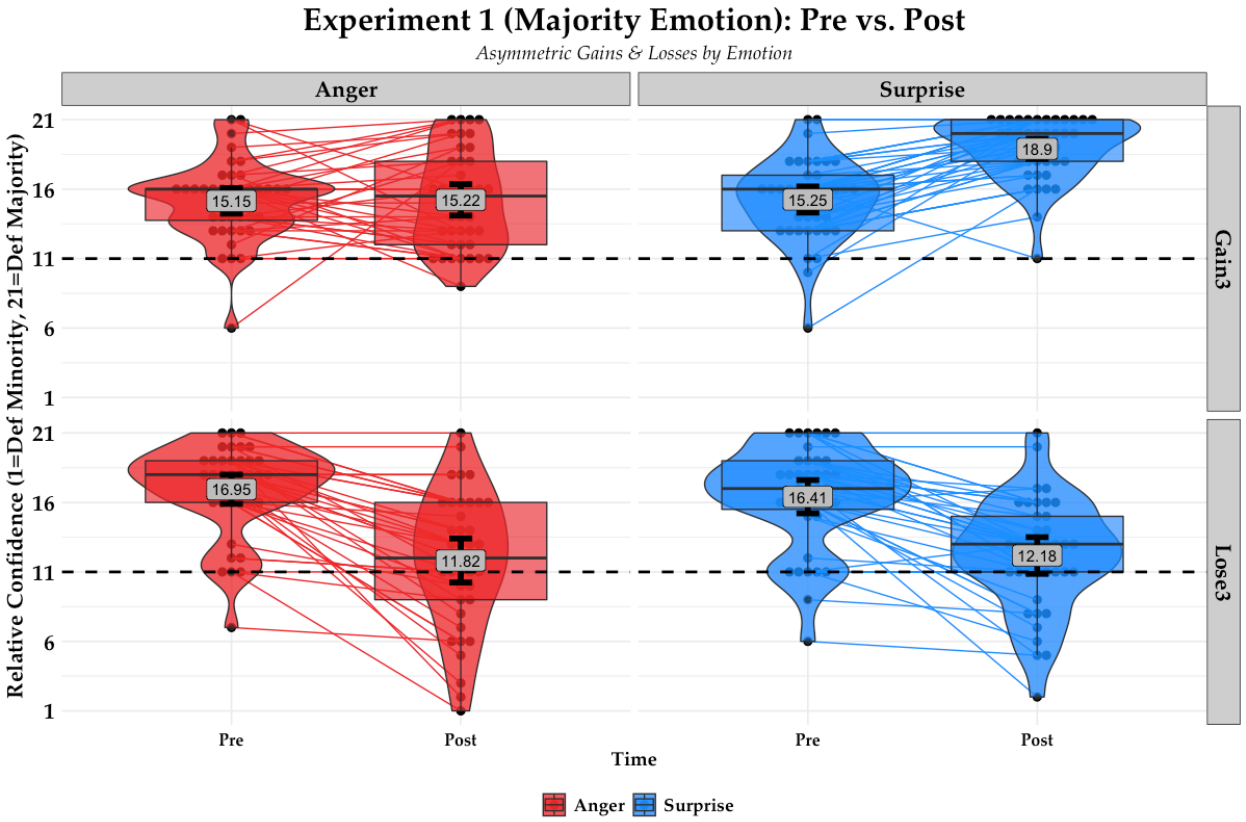


Figure 2. Distribution of responses for Experiment 1. Each participant rated their relative confidence in the two answers on a 1-21 scale on one *Anger* trial and one *Surprise* trial. In the *Gain3* condition, consensus grew from 6v4 Pre-Meeting to 9v1 Post-Meeting; in the *Lose3* condition, consensus fell from 9v1 Pre-Meeting to 6v4 Post-Meeting. With the exception of the asymmetry for angry gains, confidence ratings rose or fell as consensus gained or lost endorsers. Grey labels show means; lines join individual participants' pre- and post-meeting ratings. Error bars are 95% CIs.

the majority. Our data was broadly consistent with this pattern. Prior the meeting, participants favored both the 6vs4 majority (*MajorityGain3*: M_{Anger} = 15.15, t(39) = 9.18, p<.001; $M_{Surprise}$ = 15.25, t(39) = 9.14, p<.001) and the 9vs1 majority (*MajorityLose3*: M_{Anger} = 16.95, t(39) = 11.5, p<.001; $M_{Surprise}$ = 16.41, t(39) = 9.14, p<.001). Importantly, the pre-ratings for Anger and Surprise did not differ (as no emotion had been expressed); however, the pre-ratings for the initial 9vs1 majority were significantly higher than the initial 6vs4 majority (InitialVote: F(1,154)=8.45, p= .004), with no Emotion*InitialVote interaction. Post-meeting ratings suggested effects of opinion dynamics beyond the predicted asymmetry. While participants in the 6v4-to-9v1 condition continued to favor the majority after the meeting for (*MajorityGain3*: M_{Anger} = 15.22, t(39) = 7.58, p<.001; $M_{Surprise}$ = 18.90, t(39) = 22.1, p<.001), trust in the new majority in the 9v1-to-6v4 condition was no longer significantly different from the midpoint (*MajorityLose3*: M_{Anger} = 11.82, t(39) = 1.05, p=.30; $M_{Surprise}$ = 12.18, t(39) = 1.81, p=.079), contrary to the prediction that participants would continue to favor the majority even if it shrinks from 9v1 to 6v4.

Finally, we asked whether confidence reflected majority strength when pooling ratings according to vote share across measurement times. We pooled ratings for each Emotion according to vote proportion (i.e., ObservedVote_6v4 or ObservedVote_9v1) with the single exception that, per our prediction of an asymmetry for angry gains, the Anger_Gain3_Post rating was treated as a 6v4 vote instead of 9v1 (PenalizedVote_6v4). A one-way ANOVA showed an effect of the Asymmetry, but no effect of Emotion, so we removed Emotion. Bonferroni-corrected post-hoc comparisons suggested that while confidence in the majority was significantly higher in ObservedVote_9v1 than both ObservedVote_6v4 and PenalizedVote_6v4, the difference between the latter two was also significant (M_{Obs9v1} = 17.43, SE=.34; M_{Obs6v4} = 13.62, SE=.30; M_{Pen6v4} = 15.22, SE=.59; Obs6v4 vs Obs9v1: t(313) = -8.415, *SE*=.453, *p*<.0001; Pen6v4 vs Obs9v1: t(313) = -2.21, *SE*=.681, *p*<.0040; Pen6v4 vs Obs6v4: t(313)=1.60, *SE*=.659, *p*= .046*).* Why did the two 6v4 ratings differ? We consider possible causes in the General Discussion; to foreshadow, we attribute them to the influence of opinion dynamics combined with unusually high ratings for the 6v4 pre-meeting judgments in the Gain3 condition, as compared both to pilot data and all other 6v4 pre-meeting ratings in subsequent experiments.

Experiment 2

In Experiment 2, we used the same procedure to examine observers' evaluations of *Minority* affect. As underdogs, minority factions tend to have less power to either force majority conversion or convince them. Thus, people may be less certain whether a majority-to- minority conversion is due to the threat of minority anger instead of a genuine belief change, making minority gains less suspect than majority gains (and minority losses more suspect). Alternatively, affective signals could work similarly in any relationship, regardless of power: to the extent that conflict is costly to both sides, even a majority may convert simply to avoid conflict. As in Experiment 1, our critical prediction is that confidence in the *Angry* faction would fall when they lost endorsers, but would *not* rise even when they gained endorsers, because participants infer that converts are no longer endorsing their genuine beliefs. We can also check for an underdog effect that favors minority factions by comparing Experiments 1 and 2: if reasoning about affective signals integrates relative power, then the predicted effect of *Anger* may be weaker in Experiment 2 than in Experiment 1.

Participants. We recruited 80 participants from MTurk (n=40 in MajorityGain & n=40 in MajorityLoss; one additional participant was excluded prior to participation for failing basic comprehension checks about the instructions twice in a row). Each participant made a pre-meeting and post-meeting in an Anger trial and in a Surprise trial (order counterbalanced).

Materials & Procedure. Materials and procedure were identical to Experiment 1, but with the minority as the emotional faction.

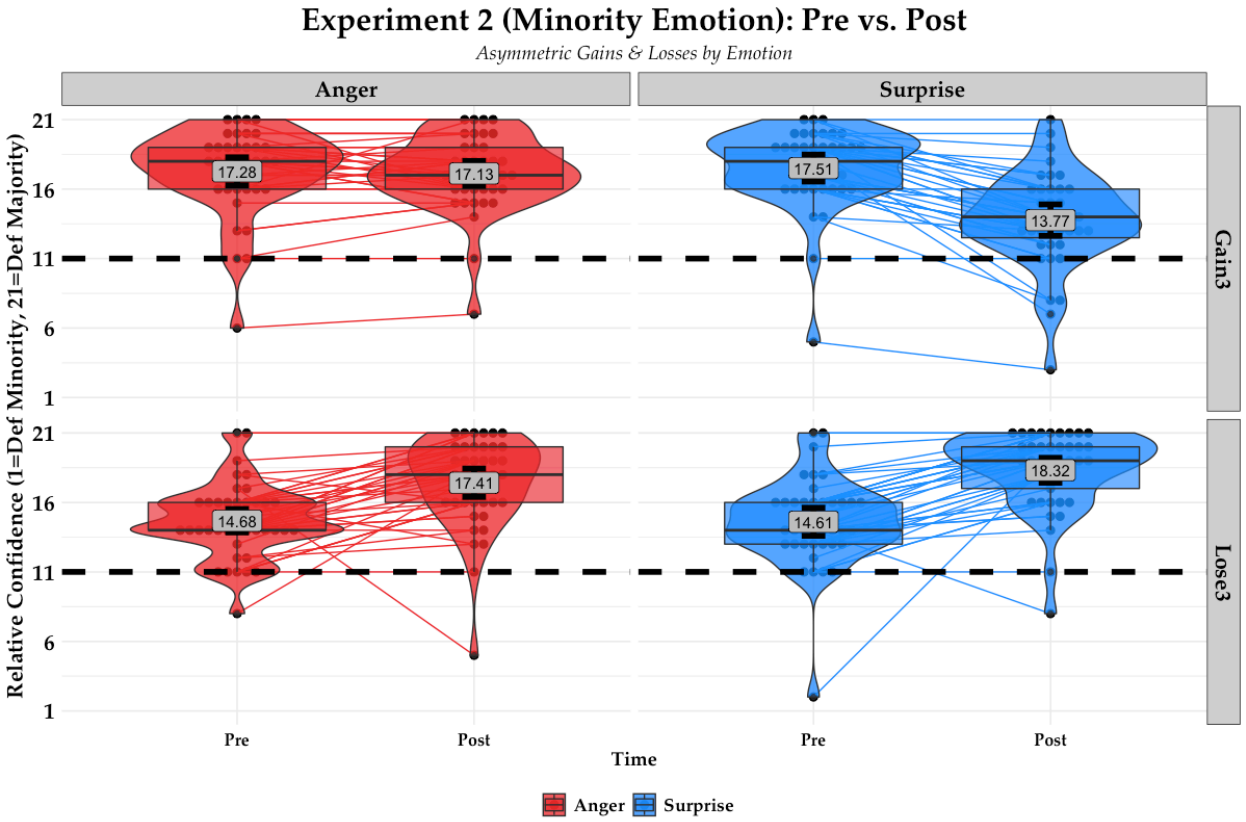


Figure 3. Distribution of responses for Experiment 2. Each participant rated their relative confidence in the two answers on a 1-21 scale on one *Anger* trial and one *Surprise* trial. In the *Gain3* condition, consensus grew from 1v9 Pre-Meeting to 4v6 Post-Meeting; in the *Lose3* condition, consensus fell from 4v6 Pre-Meeting to 1v9 Post-Meeting. With the exception of the asymmetry for angry gains, confidence ratings rose or fell as consensus gained or lost endorsers. Grey labels show means; lines join individual participants' pre- and post-meeting ratings. Error bars are 95% CIs.

Results. As in Experiment 1, our primary question was whether people's reactions to the discussions differed by the direction of *Change* and *Emotion*. We again computed the difference in Pre-Post ratings for the Anger and Surprise trials; however, we reverse-scored these differences so that the negative numbers indicate a decrease in confidence in the emotional minority and positive numbers indicate an increase in confidence. In each condition, the shift in confidence was consistent with our predictions. When the minority faction was *Angry* at the majority view, confidence in minority accuracy *fell* significantly when the minority *Lost* three endorsers to the majority view, but *did not rise* significantly even when they *Gained* three endorsers from the majority view (*Anger*: $\beta_{Lose3} = -2.73$, SE = .45, $p<.001$; $\beta_{Gain3} = 0.154$, SE = .46, $p = .74$). In contrast, when the minority faction was *Surprised* at the majority view, confidence in minority accuracy *fell* significantly when the minority *Lost* three endorsers to the majority view and *rose* significantly when they *Gained* three endorsers from the majority view (*Surprise*: $\beta_{Lose3} = -3.71$, SE = .45, $p<.001$; $\beta_{Gain3} = 3.74$, SE = .46, $p<.001$).

We next asked whether participants' ratings favored majorities. Past work has suggested that people trust majority over minority judgement, and that confidence increases with the size of the majority. Our data was broadly consistent with this pattern. Prior the meeting, participants favored both the 6vs4 majority (*MinorityLose3*: M_{Anger} = 14.68, t(40) = 8.87, p<.001; $M_{Surprise}$ = 14.61, t(40) = 7.22, p<.001) and the 9vs1 majority (*MinorityGain3*: M_{Anger} = 17.28, t(38) = 12.6, p<.001; $M_{Surprise}$ = 17.51, t(38) = 13.6, p<.001). Importantly, the pre-ratings for Anger and Surprise did not differ (as no emotion had been expressed); however, the pre-ratings for the initial 9vs1 majority were significantly higher than the initial 6vs4 majority (InitialVote: F(1,156)=33.71, p < .001), with no Emotion*InitialVote interaction. Moreover, participants continued to favor the majority after the meeting in both the 6v4-to-9v1 condition (*MinorityLose3*: M_{Anger} = 17.41, t(40) = 12.7, p<.001; $M_{Surprise}$ = 18.32, t(40) = 16.5, p<.001), and in the 9v1-to-6v4 condition (*MinorityGain3*: M_{Anger} = 17.13, t(38) = 13.9, p< .001; $M_{Surprise}$ = 13.77, t(38) = 4.95, p < .001).

Finally, we asked whether confidence reflected majority strength when pooling ratings according to vote share. We pooled ratings for each Emotion according to vote proportion (i.e., ObservedVote_6v4 or ObservedVote_9v1) with the single exception that, per our prediction of an asymmetry for angry gains, the Anger_Gain3_Post rating was treated as a 6v4 vote instead of 9v1 (PenalizedVote_6v4). A one-way ANOVA showed an effect of the Asymmetry, but no effect of Emotion, so we removed Emotion from the model. Bonferroni-corrected post-hoc comparisons suggested that while confidence in the majority was significantly lower in ObservedVote_6v4 than both ObservedVote_9v1 and PenalizedVote_9v1, the difference between the latter two was not significant (M_{Obs9v1} = 17.64, SE=.24; M_{Obs6v4} = 14.36, SE=.28; M_{Pen9v1} = 17.13, SE=.49; Obs9v1 vs Obs6v4: t(317) = -8.93, *SE*=.367, p<.0001; Pen9v1 vs Obs6v4: t(317) = -4.93, *SE*=.56, p<.0001; Pen9v1 vs Obs9v1: t(317)=-0.51, *SE*=.54, *p= ns).*

Notably, the genuine changes in consensus were followed by corresponding shifts in confidence for 87% of participants in Experiments 1-2, versus only 35% when the consensus "forced" their opponents to convert. Moreover, even when the majority lost endorsers, only ~10% of participants responded to these changes in consensus by doubting both solutions equally (i.e., retreating to the midpoint of the confidence scale), consistent with the proposal that in evaluating consensus, people treat informants' collective evaluations of higher-order evidence as *more* reliable than their evaluations of firsthand evidence. In Experiment 3, we ask whether people treat consensus as informative even if the same number of informant conversions results in a different solution garnering the majority of endorsements.

Experiment 3

In Experiments 1-2, the changes in consensus never altered which solution garnered a majority of votes. Given people's tendency to trust majority opinion over minority opinion, is it

possible that changes in consensus made participants more or less confident in the reliability of the majority heuristic per se, without influencing their beliefs about the minority judgment? In other words, seeing the majority *gain* endorsers could make you more confident that the majority heuristic is reliable and the majority is therefore correct; but seeing the majority *lose* endorsers would simply make you doubt that the majority heuristic is reliable, leaving you unsure which answer is more accurate. However, our account makes an alternative prediction: people will treat the change of consensus as evidence in favor of the converts' new belief regardless of who convinced who. On this account, people's confidence will follow the consensus even if it comes to endorse a different solution than it did initially.

Participants. We planned to recruit 80 participants from MTurk, but one additional participant was included due to a coding error in the recruitment platform (n=40 in MajorityLose3 & n=41 in MinorityGain3; three additional participants were excluded prior to participation for failing basic comprehension checks about the instructions twice in a row). Each participant made a pre-meeting and post-meeting in an Anger trial and in a Surprise trial (order counterbalanced).

Materials & Procedure. All participants saw a pre-vs-post meeting shift of 6v4-to-3v7. In the MajorityLose3 condition, the 6-person majority was the emotional faction, while in the MinorityGain3 condition, the 4-person minority was the emotional faction. Thus, we recycled the materials for the initial consensus and meeting from Experiments 1 and 2, and simply reversed the direction of the conversions after the meeting. In other words, the materials for the MajorityLose3 condition of Experiment 3 are identical to the MajorityGain3 condition of Experiment 1 except that 3 of 6 emotional majority members join the minority instead of 3 of the 4 minority members joining the emotional majority; the materials for the MinorityGain3 condition of Experiment 3 are identical to the MinorityLose3 condition of Experiment 2, except that instead of 3 of the 4 emotional minority members joining the majority, 3 of the 6 majority members join the emotional minority.

Results. Participants' judgments confidence judgments categorically reversed from pre-meeting to- post-meeting in all conditions except the one in which we had predicted an asymmetry (MinorityGain3_Anger). As in Experiments 1 and 2, we computed the difference in Pre-Post ratings for the Anger and Surprise trials; differences are scored so that so negative numbers indicate decreased confidence in the majority, regardless of the emotional faction. When the emotional *Majority* faction *Lost* three endorsers to the minority view, confidence in the majority *fell* significantly regardless of whether it was *Angry* or *Surprised* (*MajorityLose3*: β_{Anger} = -5.37, SE = .72, p<.001; $\beta_{Surprise}$ = -7.17, SE = .72, p<.001). When the emotional *Minority* faction *Gained* three endorsers from the majority view, confidence in the majority *fell* significantly when

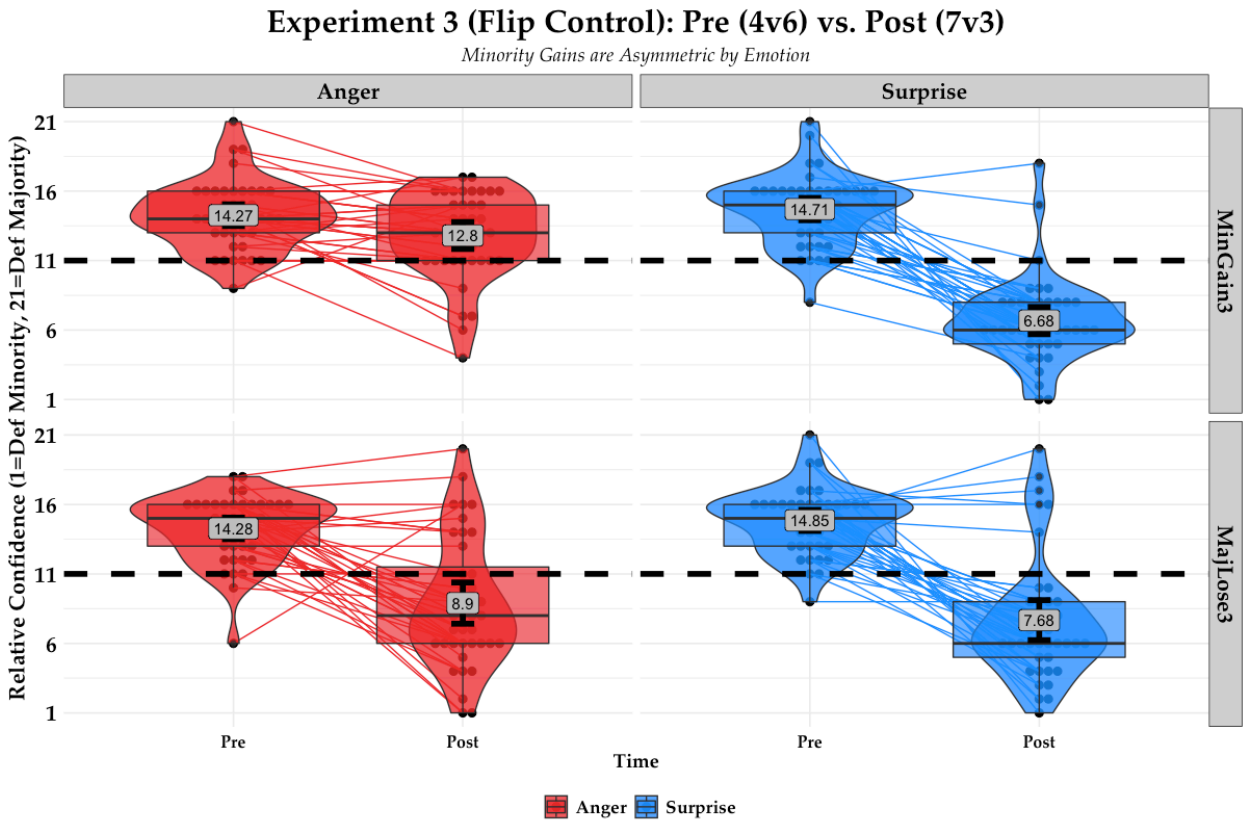


Figure 4. Distribution of responses for Experiment 3. Each participant rated their relative confidence in the two answers on a 1-21 scale on one *Anger* trial and one *Surprise* trial. In both the *MinorityGain3* condition and the *MajorityLose3* condition, consensus flipped from 6v4 Pre-Meeting to 3v7 Post-Meeting. With the exception of the asymmetry when the an angry faction gained endorsers (i.e., MinGain3), confidence ratings rose or fell as consensus gained or lost endorsers. Grey labels show means; lines join individual participants' pre- and post-meeting ratings. Error bars are 95% CIs.

the minority expressed *Surprise*, but also (contrary to prediction), when they expressed *Anger* (*MinorityGain3*: $\beta_{Surprise}$ = -8.02, SE = .71, p<.001; β_{Anger} = -1.46, SE = .71, p<.041).

Despite the angry minority's unexpected success in shifting participants' confidence in their judgment post-meeting, participants continued to favor the initial majority overall — whereas in all other conditions, participants flipped to trusting the new majority post-meeting (*MajLose3*: M_{Anger} = 8.90, t(39) = -2.86, p<.007; $M_{Surprise}$ = 7.68, t(39) = -4.66, p<.001; *MinGain3:* $M_{Surprise}$ = 6.68, t(40) = -8.88, p<.001; M_{Anger} = 12.80, t(40) = 3.76, p<.001). Moreover, only the asymmetric (MinorityGain3_Anger) post-meeting ratings differed significantly from any of the others, and all of the pre-meeting ratings indicated greater trust in the initial majority of 6v4 than the minority, with no differences between conditions (*MajLose3*: M_{Anger} = 14.28, t(39) = 8.57, p<.001; $M_{Surprise}$ = 14.85, t(39) = 10.30, p<.001; *MinGain3:* M_{Anger} = 14.27, t(40) = 8.47, p<.001; $M_{Surprise}$ = 14.71, t(40) = 9.34, p<.001).

Experiment 4

What drives the asymmetry in confidence changes for angry gains? One possibility is that people treat anger per se as signaling a relative lack of competence; indeed, the benevolence bias in children and adults emerges after a failure early in development to distinguish between warmth and competence. We propose an alternative: people only reject anger as a means of influencing others. On this account, using anger to force consensus change means that observers cannot be sure that the converts' judgments reflect their genuine evaluations of the evidence; however, if the angry informants restrain their anger while attempting to convince others, there is less reason to doubt that the converts' genuine judgments were compromised by social pressures, and thus changes in consensus are informative.

Participants. We again aimed for a sample of 40 participants per condition. Because the *Display* manipulation made no new predictions for *Surprised* groups or *Lose3* groups, participants were only shown an Anger trial, and assigned to one of the four between-subjects conditions in which the emotional faction gains endorsers: *Express_MajGain3* (n=39), *Express_MinGain3* (n=40), *Restrain_MajGain3* (n=42), *Restrain_MinGain3* (n=39).

Materials & Procedure. The materials used in the previous experiments were modified in two ways. First, students were shown with a wall separating the two factions in order to allow participants' to see the students emotional reactions, but emphasize that the students themselves were unaware of the other faction's emotions. Second, the text was modified to say that while the emotional faction was very angry about the other students answers before the meeting, they either (Restrain) "*did not show that they were angry when they met to talk as a group. But after they talked with [the other students]…*" or (Express) "*were still very angry when they met to talk as a group. But after they shouted at [the other students]…*".

Results. We first computed the difference in Pre-Post ratings for the Anger trials; differences are scored so that negative and positive numbers indicate decreased or increased confidence in the emotional faction. When the angry faction gained three endorsers after shouting at them, they once again failed to gain any confidence from participants, regardless of whether the angry faction was the minority or majority (*Express*: $\beta_{MajGain3}$ = -0.85, SE = .50, p=.093; $\beta_{MinGain3}$ = 0.60, SE = .50, p=.227). However, when the angry faction restrained their anger and gained three endorsers after talking with them, participants' confidence shifted accordingly, with no Display*EmotionalFaction interaction (*Restrain*: $\beta_{Restrain}$ = 2.99, SE = .70, p<.001; $\beta_{MajGain3*Restrain}$ = .76, SE = .99, p=.769).

We next asked whether participants' ratings favored majorities. Prior the meeting, participants favored both the 6vs4 majority (*MinGain3*: $M_{Restrain}$ = 18.36, t(38) = 20.83, p<.001; $M_{Express}$ = 18.43, t(39) = 18.48, p<.001) and the 9vs1 majority (*MajGain3:* $M_{Restrain}$ = 14.43, t(41) =

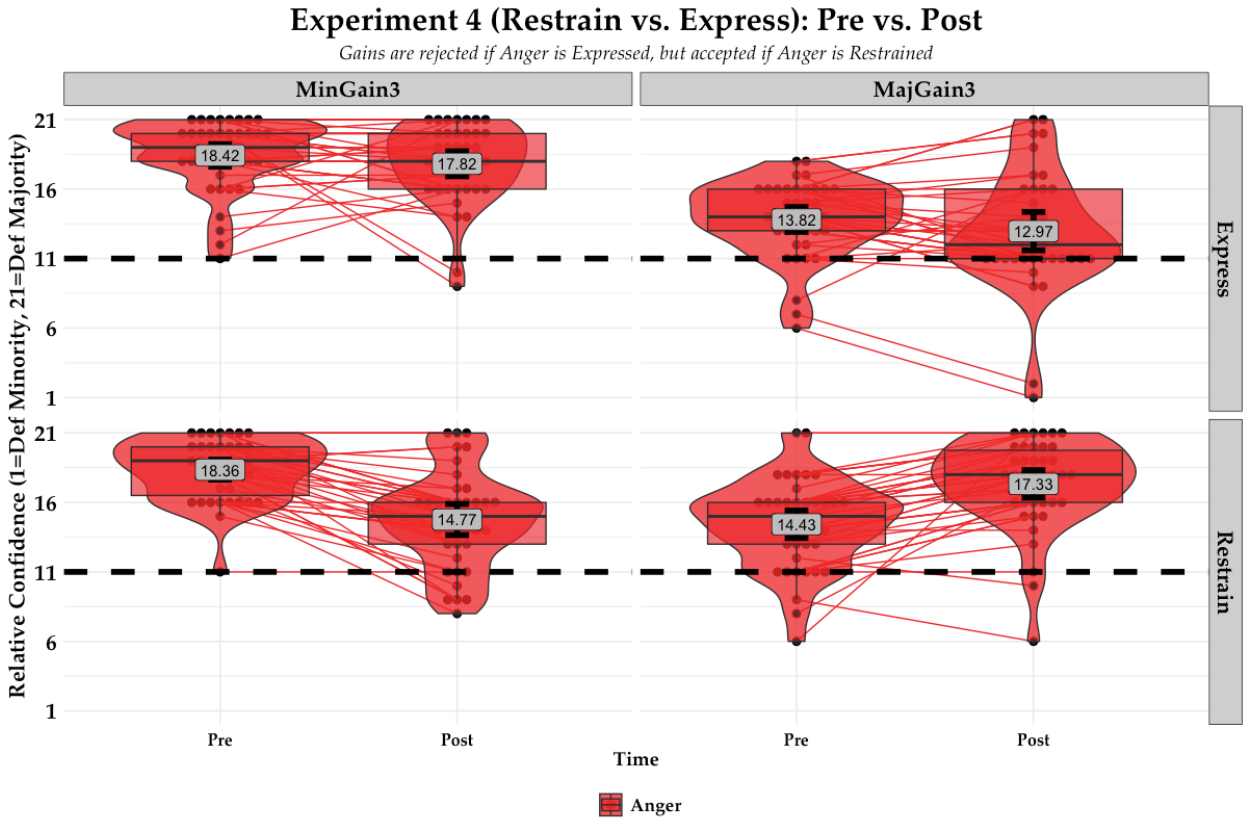


Figure 5. Distribution of responses for Experiment 4. Each participant rated their relative confidence in the two answers on a 1-21 scale on one *Anger* trial (since there were no new predictions for *Surprise*, no *Surprise* trial was run). In both the *MinorityGain3* condition and the *MajorityLose3* condition, consensus shifted from 6v4 Pre-Meeting to 9v1 Post-Meeting, but either the Majority or Minority expressed emotion. Confidence ratings rose or fell as consensus gained or lost endorsers when the faction restrained their anger, but not when they expressed it. Grey labels show means; lines join individual participants' pre- and post-meeting ratings. Error bars are 95% CIs.

7.02, $p<.001$; $M_{Express}$ = 13.82, t(38) = 6.40, $p<.001$). As expected, the pre-meeting ratings for *Express* and *Restrain* did not differ; an ANOVA suggested higher confidence in the 9vs1 pre-meeting vote of the MinGain3 condition than the 6v4 pre-meeting vote of the MajGain3 condition (F(1,156)=99.62, $p < .001$), with no effect of Display or interaction, suggesting that participants' pre-meeting ratings were a function of the degree of consensus. However, while participants post-meeting judgments continued to favor the majority in both the 6v4-to-9v1 condition (*MinGain3*: $M_{Restrain}$ = 14.77, t(38) = 6.80, $p<.001$; $M_{Express}$ = 17.83, t(39) = 15.37, $p<.001$), and in the 9v1-to-6v4 condition (*MajGain3:* $M_{Restrain}$ = 17.33, t(41) = 12.93, $p<.001$; $M_{Express}$ = 12.97, t(38) = 2.89, $p=.006$), an ANOVA revealed a significant *Display*EmotionalFaction* interaction (F(1,156)=45.85, $p < .001$), but no effect of *Display* and a marginal effect of *EmotionalFaction* (*Display*: F(1,156)=1.62, $p = .21$; *EmotionalFaction*: F(1,156)=3.91, $p = .0498$). Multiple-comparisons of participants' confidence ratings suggested that they had ignored the post-meeting consensus when the emotional faction had *Expressed* their anger, but accepted the post-meeting consensus

when they *Restrained* their anger. The post-meeting ratings for a 9v1-to-6v4 minority gain in *Express* were not significantly different than the post-meeting ratings for a 6v4-to-9v1 majority gain in *Restrain* (t(156) = 0.643, *SE*=0.765, *p=ns)*, and the post-meeting ratings for a 6v4-to-9v1 majority gain in *Express* were not significantly different than the post-meeting ratings for a 9v1-to-6v4 minority gain in *Restrain* (t(156) = -2.29, *SE*=0.784, *p*=.14*);* however, all other comparisons differed as would be predicted by the Express-reject Restrain-accept account (*Express: MinGain3 vs. MajGain3*: t(156) = 6.23, *SE*=.779, *p*<.0001; *Restrain: MinGain3 vs. MajGain3*: t(156) = -3.33, *SE*=.770, *p*<.001; *MajGain3: Express vs. Restrain*: t(156) = -5.66, *SE*=.770, *p*<.0001; *MinGain3: Express vs. Restrain*: t(156) = 3.92, *SE*=.770, *p*<.001*).*

General Discussion

These experiments document two general patterns. First, far from discounting any post-meeting change in consensus as an illusion created by pernicious social influences, changes in consensus prompted proportional changes in participants' confidence — as long as participants were given no reason to doubt the authenticity of their informants' judgments. Second, the asymmetry for angry gains suggests that participants were neither following the consensus judgment blindly nor treating anger as a sign of incompetence. On the contrary, participants grew more confident in an angry faction that gained endorsers, as long as the faction didn't express their anger during the meeting. And while they rejected an angry faction's influence on others, they treated the angry informants' *own* beliefs as informative despite their anger.

Our results suggest that participants' trust in consensus is driven by intuitions about higher-order evidence. In other words, people interpret each other's beliefs as evidence of evidence — and "evidence of evidence *is* evidence" (Feldman, 2009; Christensen, 2009; Dorst, 2020). If Alice tells Bob that she knows exactly how high her rocket will fly, her belief in her own knowledge provides Bob with higher-order evidence about her calculation's accuracy, even if he doesn't actually know what her result is. Hearing Alice report the calculations doesn't always give Bob new first-order evidence (after all, he may have run the same calculations and gotten a different result). But testimony of any kind always generates new higher-order evidence (e.g., Bob can now weigh his trust in Alice's result or her competence against his own, or consider the plausibility of the height-thrust ratios). And if discussion is the only means of resolving disagreement (as in our experiments), then even if Alice alleges new facts that would constitute first-order evidence if true, Bob's decision to accept or reject those facts ultimately depends on trust. Moreover, reasoning about how the kind of evidence available to Bob has shaped his beliefs makes his beliefs a rich source of higher-order evidence for *us* as well. Of course, evidence is *just evidence*: Alice may be wrong, and Bob's trust or distrust in her may be misplaced. But regardless of Alice and Bob's reliability, our willingness as observers to treat

each others' beliefs as higher-order evidence can inform research on both consensus and our capacity for reasoning about other minds.

What can participants' inferences tell us about human mentalizing? Even young children expect individual agents to rationally update their beliefs when presented with new evidence from the world around them (Magid et al., 2018; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). These rational agent studies ask participants to infer an agent's updated beliefs from their prior beliefs and incoming evidence. Our studies reverse this reasoning: people infer the strength of evidence available to their informants by reasoning about the extent to which informants' belief-forming processes are reliable. Since changing our beliefs under the influence of someone else's risks inheriting their errors, one might have expected participants to treat the discussion as undermining the reliability of their informants' belief changes (Yousif et al., 2019). Indeed, in the extreme case of informants *blindly imitating* each other, any belief changes make consensus less reliable (Condorcet, 1785). However, if one's informants *are* evaluating each other's reliability, then failing to trust them suggests either skepticism of their evaluations, or a failure to engage in theory-of-mind reasoning at all. In our studies, the anger asymmetry clearly demonstrates theory-of-mind reasoning. But despite the risk of inherited errors, participants inferred that genuine discussion made the consensus judgment *more* reliable. In other words, even when an informant is convinced by discussion that their own independent judgment is less reliable than someone else's, we don't simply disregard them as an unreliable source. We trust them to weigh their own judgment against others'. This suggests that reasoning about how people process evidence collectively may differ from reasoning about how they process evidence in isolation. However, limitations in our design mean that more work is needed to understand what these differences are.

People may believe that collectives to have higher information-processing capacities or lower error rates than individuals. Similarly, participants may believe that resolving *inter*personal disagreement simply requires more rigor or effort than resolving *intra*personal uncertainty, and thereby produces more reliable judgments. These intuitions may be especially salient in the engineering context we studied. Engineering not only implies a ground truth, but affords fairly objective methods of demonstrating errors and resolving disagreements. Other contexts may differ. For instance, adjudicating eyewitness testimony is ultimately a matter of trust even when conflicting memories all refer to a single ground truth (Aboody et al., 2022). In ethical decisions or matters of subjective preference, there may either be no ground truth at all or fundamental antinomies underlying disagreement. Future work could examine contexts in which evidence might be expected to be less robust than the engineering contest in our studies (e.g., eyewitness memory or fair division of resources). Another possibility is that participants inferred that a

discussion would simply allow the group to identify their "best member" and defer to their judgment. While participants in our studies had no reason to assign more evidential weight to one informant's judgments than another's *prior* to the meeting, the *post*-meeting congruence between participants' confidence and the "true" consensus is consistent with several different inferences about competence and belief change. Participants may have simply recomputed the "vote" while continuing to give equal evidential weight to each informant; or, they may have grown proportionally more confident in the informants' who were able to convince their opponents. Future work could address these possibilities by asking how participants' reasoning about consensus' reliability changes in mixed-competence groups, or by asking participants to directly infer informants' competence from their influence on others (Kawakatsu, Chodrow, Eikmeier, & Larremore, 2021).

What can participants' inferences tell us about people's ability to learn from consensus? After all, our fictional scenarios and color-coded answers provided no ground truth from which to evaluate accuracy. However, for learners whose informants are learning from each other, a reliable informant is simply one that can reliably identify reliable informants, even when they are unable to solve the problem themselves. The more reliably learners can identify such informants, the more those learners can amplify accurate information for each other; the less reliably they do so, the more they may contribute to "illusions" of consensus (Yousif, et al., 2019). Thus, clarifying what people *think* makes consensus reliable is critical to understanding how reliable consensus is actually likely to be. Moreover, converging lines of evidence from biology, psychology, and philosophy suggest that under certain conditions it would not be unwise to infer that discussion would make the consensus more reliable than the informants' independent judgments. First, group discussions frequently outperform the majority vote of individual learners, and in some cases even the group's best member (Laughlin, Bonner, & Miner, 2002; Mercier & Claidière, 2021; Navajas et al., 2018). This is especially common for tasks with *demonstrably* better and worse solutions, such as the engineering context we studied here (Laughlin & Ellis, 1986; Bonner et al., 2021). Second, simulations of animal groups suggest that consensus judgments are more reliable when the informants pool their total collective evidence before making a decision than when they each process their evidence individually and decide by simply voting (Kao et al., 2014). Indeed, the potential for informational dependencies to improve consensus judgment under certain conditions can also be shown analytically (Barnett, 2019; Pilditch, Hahn, Fenton, & Lagnado, 2020), and managing informational dependencies may be critical for groups of partially informed agents to be able to form reliable beliefs (Dunn, 2019; Goldman, 2014). People's preference for *more* or *less* independence between their informants evinces some understanding of how informational dependencies can affect the reliability of

consensus judgments. This understanding appears as early as age six and is sensitive to the demonstrability of the task (Richardson & Keil, 2022; Aboody et al., 2022; Yousif et al., 2019, Exp 5). In short, while we can't say whether participants' trust in the post-meeting consensus was wise or unwise, understanding how people evaluate informational dependencies between their informants is an essential part of understanding what makes consensus more reliable in some contexts than others.

Finally, the asymmetric evaluation of angry groups in our findings raises questions about the influence of affective signals on the diffusion of beliefs in social networks (Brady et al., 2017; cf. Burton et al., 2019). While there are important contextual differences between debates in an "engineering contest" in our experiments and the coordinated broadcasts of political messages people are responding to in observational studies of social networks like Twitter, our results suggest that anger per se may dramatically reduce *belief* diffusion even if it amplifies information itself. This finding could help explain the polarizing effect of moral outrage (Crockett, 2017): while anger sharply limits belief diffusion when observers have no way to evaluate a belief directly, asking people to evaluate angry rhetoric about controversies they *already* hold opinions about may produce radically different results. Indeed, to the extent that having similar beliefs increases the probability that two people will get angry at similar things, it would be quite surprising if Bob were *not* more likely to respond angrily to news that made his friend Alice angry than news that didn't. Thus, to the extent that people are more likely to share news that provokes a strong emotional reaction, anger-inducing news may be able to quickly saturate homophilous networks. However, this does not mean that Bob is angry about the news *because* Alice was angry, nor that he gives it more credence *because* Alice believed it. On the contrary, our results suggest that even if angry tweets are more likely to be retweeted by "allies", an aversion to angry rhetoric among neutral observers will limit how far through a network the news can travel from the original source. In other words, angry rhetoric may still reach a much smaller audience than it could have reached if it had been packaged differently. Anger may also limit the direct influence a speaker's argument has on public opinion, particularly in contexts with strong norms against affect-laden rhetoric.

More concerning is the potential for algorithms that monetize engagement to shift norms by reinforcing any behavior that increases audience engagement (Brady, McLoughlin, Doan, & Crockett, 2021). A brazen lie or mocking comment may ruin your chances of convincing your interlocutor, but if it convinces even a fraction of your wider audience, you may still have more to gain than lose. However, audiences are often well aware of the performative nature of their informants' public speech; indeed, speakers may be counting on it. For instance, accusations of moral violations may seem *less* credible if the accuser is not angry, but anger can only be

expressed by certain people and in certain ways; identifying parody can be difficult if the speaker is unknown (Roudakova, 2017). Research on misinformation and moral outrage could benefit from examining how theory-of-mind reasoning about our informants' rhetorical strategies in public disputes informs our trust in their testimony, both online and offline.

Social learning allows individuals to learn far more than they could individually, and trust in consensus is a widely used learning strategy among both humans and other animals. Yet, like other social learning strategies, consensus can expose us to popular delusions based on limited evidence if taken at face value (Harvey et al., 2018; Yousif et al., 2019). Our results suggest that people's trust in consensus is not blind. Rather, people may trust consensus to the extent that they believe their informants are reliable judges of higher-order evidence; indeed, given that even the fairly mundane judgments we make in daily life often require us to integrate information from partially-informed informants whose judgments are based on secondhand information at best, this trust in higher-order evidence may be a rational strategy. Indeed, the depth and breadth of technical specialization that characterize human societies would be impossible if we were unwilling to learn from each others' secondhand judgments. Comparing people's trust in each others' evaluations of higher-order evidence with ground truth may be critical to understanding both misinformation and illusions of consensus as well as our extraordinary success as social learners.

Context

Trusting majority judgment more than minority judgment is not only a common strategy for learning from others; it is demonstrably more reliable than minority judgment given certain assumptions. However, those assumptions — such as the statistical independence of individuals' judgments — are implausible in the biological world. Moreover, consensus is notoriously vulnerable to maladaptive herd behavior when judgments are not independent. Why do we trust non-independent consensuses nevertheless? Our experiments were motivated by the intuition that we treat our social networks as "epistemic filters". In other words, we trust each other's ability to evaluate testimonial evidence. In many cases, we may even expect to be *more* reliable in collectively evaluating each other's judgment than in evaluating firsthand evidence on our own. Our results contribute to a growing literature in philosophy and the cognitive sciences on collective belief formation.

References

1. Aboody, R., Yousif, S. R., Sheskin, M., & Keil, F. C. (2022). Says who? Children consider informants' sources when deciding whom to believe. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001198
2. Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4). https://doi.org/10.1038/s41562-017-0064
3. Barnett, Z. (2019). Belief dependence: How do the numbers count? *Philosophical Studies*, *176*(2), 297–319. https://doi.org/10.1007/s11098-017-1016-0
4. Bartlett, F. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.
5. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114
6. Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641
7. Bonner, B. L., Shannahan, D., Bain, K., Coll, K., & Meikle, N. L. (2021). The Theory and Measurement of Expertise-Based Problem Solving in Organizational Teams: Revisiting Demonstrability. *Organization Science*, orsc.2021.1481. https://doi.org/10.1287/orsc.2021.1481
8. Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01133-5
9. Claidière, N., & Whiten, A. (2012). Integrating the study of conformity and culture in humans and nonhuman animals. *Psychological Bulletin*, *138*(1), 126–145. https://doi.org/10.1037/a0025868

10. Christensen, D. (2009). Disagreement as Evidence: The Epistemology of Controversy. *Philosophy Compass*, *4*(5), 756–767. https://doi.org/10.1111/j.1747-9991.2009.00237.x
11. Condorcet, M. (1785 / 1994). Essay on the application of probability analyses to decisions returned by a plurality of people. In I. McLean & F. Hewitt (Eds. & Trans.), *Condorcet: Foundations of social choice and political theory* (pp. 11–36). Brookfield, VT: Edward Elgar. (Original work published 1785)
12. Connor Desai, S., Xie, B., & Hayes, B. K. (2022). Getting to the source of the illusion of consensus. *Cognition*, *223*, 105023. https://doi.org/10.1016/j.cognition.2022.105023
13. Danovitch, J. H., & Keil, F. C. (2007). Choosing between hearts and minds: Children's understanding of moral advisors. *Cognitive Development*, *22*(1), 110–123. https://doi.org/10.1016/j.cogdev.2006.07.001
14. Dietrich, F., & Spiekermann, K. (2013). Epistemic Democracy with Defensible Premises. *Economics and Philosophy*, *29*, 34. https://doi.org/doi:10.1017/S0266267113000096
15. Dorst, K. (2020). Evidence: A Guide for the Uncertain. *Philosophy and Phenomenological Research*, *100*(3), 586–632. https://doi.org/10.1111/phpr.12561
16. Dunn, J. (2019). Reliable group belief. *Synthese*. https://doi.org/10.1007/s11229-018-02075-8
17. Feldman, R. (2009). Evidentialism, Higher-Order Evidence, and Disagreement. *Episteme*, 19. https://doi.org/10.3366/E1742360009000720
18. Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005
19. Goldenberg, A., Saguy, T., & Halperin, E. (2014). How group-based emotions are shaped by collective emotions: Evidence for emotional transfer and emotional burden. *Journal of Personality and Social Psychology*, *107*(4), 581–596. https://doi.org/10.1037/a0037462

20. Goldman, A. I. (2014). Social Process Reliabilism. In J. Lackey (Ed.), *Essays in Collective Epistemology* (pp. 11–41). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199665792.003.0002
21. Harvey, J. A., van den Berg, D., Ellers, J., Kampen, R., Crowther, T. W., Roessingh, P., Verheggen, B., Nuijten, R. J. M., Post, E., Lewandowsky, S., Stirling, I., Balgopal, M., Amstrup, S. C., & Mann, M. E. (2018). Internet Blogs, Polar Bears, and Climate-Change Denial by Proxy. *BioScience*, *68*(4), 281–287. https://doi.org/10.1093/biosci/bix133
22. Johnston, A. M., Mills, C. M., & Landrum, A. R. (2015). How do children weigh competence and benevolence when deciding whom to trust? *Cognition*, *144*, 76–90. https://doi.org/10.1016/j.cognition.2015.07.015
23. Kao, A. B., Miller, N., Torney, C., Hartnett, A., & Couzin, I. D. (2014). Collective Learning and Optimal Consensus Decisions in Social Animal Groups. *PLoS Computational Biology*, *10*(8), e1003762. https://doi.org/10.1371/journal.pcbi.1003762
24. Kawakatsu, M., Chodrow, P. S., Eikmeier, N., & Larremore, D. B. (2021). Emergence of hierarchy in networked endorsement dynamics. *Proceedings of the National Academy of Sciences*, *118*(16), e2015188118. https://doi.org/10.1073/pnas.2015188118
25. Laan, A., Madirolas, G., & de Polavieja, G. G. (2017). Rescuing Collective Wisdom when the Average Group Opinion Is Wrong. *Frontiers in Robotics and AI*, *4*. https://doi.org/10.3389/frobt.2017.00056
26. Landrum, A. R., Mills, C. M., & Johnston, A. M. (2013). When do children trust the expert? Benevolence information influences children's trust more than expertise. *Developmental Science*, *16*(4), 622–638. https://doi.org/10.1111/desc.12059
27. Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, *66*(1), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043
28. Magid, R. W., Yan, P., Siegel, M. H., Tenenbaum, J. B., & Schulz, L. E. (2018). Changing minds: Children's inferences about third party belief revision. *Developmental Science*, *21*(2), e12553. https://doi.org/10.1111/desc.12553

29. Mercier, H., & Claidière, N. (2022). Does discussion make crowds any wiser? *Cognition*, *222*, 104912. https://doi.org/10.1016/j.cognition.2021.104912
30. Mercier, H., Dockendorff, M., Majima, Y., Hacquin, A.-S., & Schwartzberg, M. (2020). Intuitions about the epistemic virtues of majority voting. *Thinking & Reasoning*, 1–19. https://doi.org/10.1080/13546783.2020.1857306
31. Mercier, H., & Miton, H. (2019). Utilizing simple cues to informational dependency. *Evolution and Human Behavior*, *40*(3), 301–314. https://doi.org/10.1016/j.evolhumbehav.2019.01.001
32. Mercier, H., Trouche, E., Yama, H., Heintz, C., & Girotto, V. (2015). Experts and laymen grossly underestimate the benefits of argumentation for reasoning. *Thinking & Reasoning*, *21*(3), 341–355. https://doi.org/10.1080/13546783.2014.981582
33. Pilditch, T. D., Hahn, U., Fenton, N., & Lagnado, D. (2020). Dependencies in evidential reports: The case for informational advantages. *Cognition*, *204*, 104343. https://doi.org/10.1016/j.cognition.2020.104343
34. Richardson, E., & Keil, F. C. (2022). The potential for effective reasoning guides children's preference for small group discussion over crowdsourcing. *Scientific Reports*, *12*(1), 1193. https://doi.org/10.1038/s41598-021-04680-z
35. Roudakova, N. (2017). *Losing Pravda: Ethics and The Press in Post-Truth Russia*. Cambridge University Press.
36. Simpson, B., Willer, R., & Feinberg, M. (2018). Does Violent Protest Backfire? Testing a Theory of Public Reactions to Activist Violence. *Socius: Sociological Research for a Dynamic World*, *4*, 237802311880318. https://doi.org/10.1177/2378023118803189
37. Steinert-Threlkeld, Z. C., Chan, A. M., & Joo, J. (2022). How State and Protester Violence Affect Protest Dynamics. *The Journal of Politics*, *84*(2), 798–813. https://doi.org/10.1086/715600
38. Teixeira, C. P., Spears, R., & Yzerbyt, V. Y. (2020). Is Martin Luther King or Malcolm X the more acceptable face of protest? High-status groups' reactions to low-status groups' collective action. *Journal of Personality and Social Psychology*, *118*(5), 919–944. https://doi.org/10.1037/pspi0000195

39. Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

40. Toyokawa, W., Whalen, A., & Laland, K. N. (2019). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, *3*(2), 183–193. https://doi.org/10.1038/s41562-018-0518-x

41. Yousif, S. R., Aboody, R., & Keil, F. C. (2019). The Illusion of Consensus: A Failure to Distinguish Between True and False Consensus. *Psychological Science*, *30*(8), 1195–1204. https://doi.org/10.1177/0956797619856844

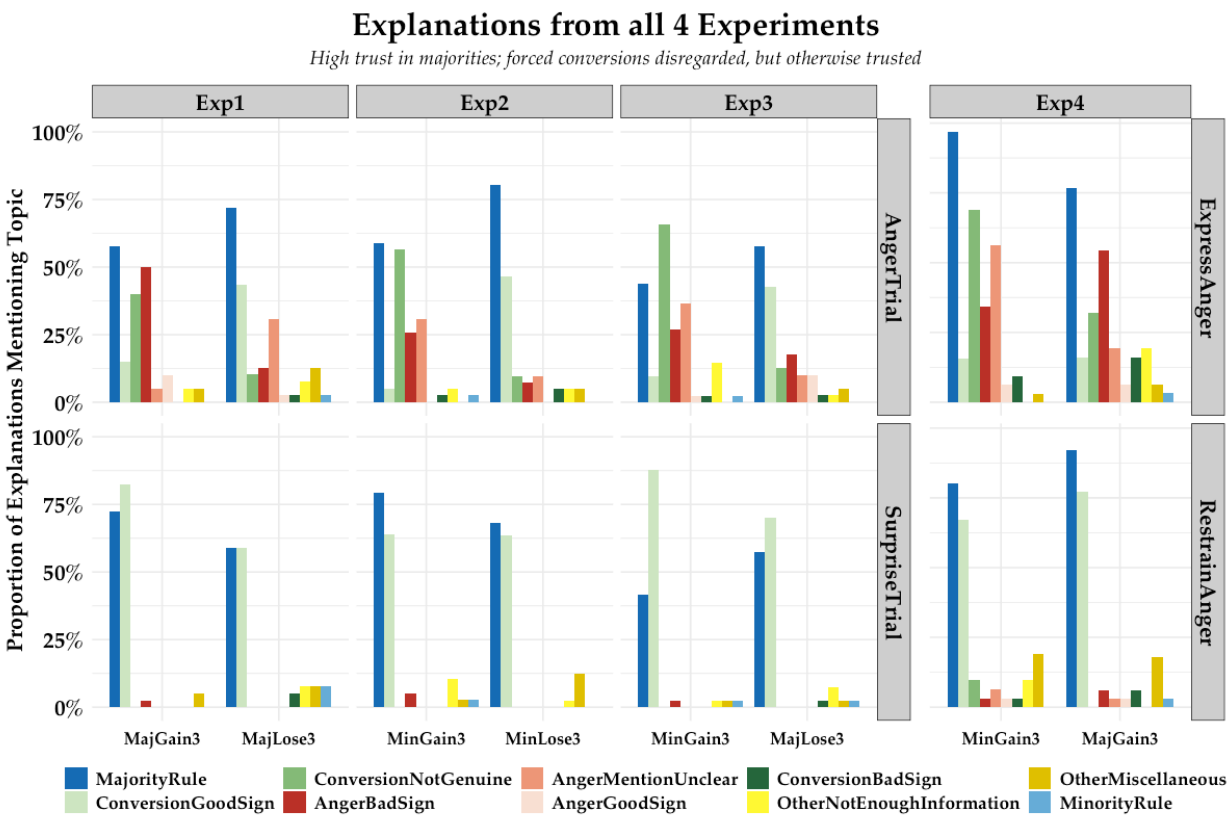
Supplemental Materials

Though we made no explicit hypotheses about the frequency of different kinds of explanations, our general expectation was that participants' explanations would match their ratings. Thus, for instance, we expected participants to infer that informants had been genuinely convinced in the *Surprise* trials (and the "Restrain Anger" condition in Experiment 4), and treat the conversion as positive evidence in favor of the informants' new answer. In contrast, we expected participants to infer that conversions were not genuine in the *Anger* trials, and treat anger per se as a bad sign.

We developed a coding scheme to categorize participants explanations. A research assistant was trained using a few examples, and coded each explanation on multiple dimensions, as presented in the outline below.

In all experiments, participants' spontaneous explanations in the *Surprise* trials (and the "Restrain Anger" condition in Experiment 4, in which the angry faction simply talked with the other faction without showing that they were angry) overwhelmingly referred to both trusting the majority judgment and treating the conversion as positive evidence in favor of the informants' new answer. In Experiment 3, where three conversions flipped majority control from 4v6-to-7v3, positive mention of conversion was even higher than mention of trusting the majority rule.

Participants also referred to trusting the majority judgment in the *Anger* trials, but only mentioned the conversions as positive evidence in favor of the informants' new judgment in the *Lose3* conditions, where it was members of the angry faction *themselves* who converted (i.e., *Exp1_MajLose3; Exp2_MinLose3; Exp3_MajLose3*). In the *Gain3* conditions, in which the informants who converted did so after being shouted at, participants often explicitly mentioned that they had not genuinely changed their beliefs (*Exp1_MajGain3; Exp2_MinGain3; Exp3_MinGain3; Exp4_MajGain3; Exp4_MinGain3*).



		ANGER	SURPRISE
Exp1	**MajLose3**	*I chose blue at first since 9 out of 10 chose blue. After talking, 4 out of the 10 students chose green even though every one of the 9 initial blue students were angry at the green student which suggests the green student was persuasive to overcome the negative emotions.*	*Well, it seems like yellow had a novel idea. It was novel enough that made some of the purples change their answer so it becomes more ambiguous on what the best way is. Still leaning towards purple since there were a few more there.*
Exp1	**MajGain3**	*I went with the majority. Even though the students were bullied into changing their minds the majority stayed the same.*	*The students who thought purple were able to convince those who thought yellow at first. I'm inclined to side with them since they were able to convince those who thought something different.*
Exp2	**MinLose3**	*More students overall chose blue in the beginning. After the other group was ANGRY, they still ended up choosing blue so that made me really confident that blue was the better option*	*I thought purple in the beginning because more students overall chose purple. Once they discussed it even more students chose purple so I felt more confident that purple was the better choice*

Exp2	**MinGain3**	*I figured there might be room for the yellow comment to be pretty persuasive, so I didn't go all in for purple. The fact that even a few were persuaded after being part of such a huge majority leaves the door open even further for the possibility that yellow might have some great points to consider.*	*I believe blue would make the rocket fly highest, even though some students changed their minds but they changed by force or fear.*
Exp3	**MajLose3**	*With no knowledge of the skill level of these students, I take their original answer at face value - that it actually is pretty close. So I chose slightly blue. After discussions, the blue students were acting immature by yelling which makes me trust them less. And some changed their mind from the majority vote to the minority vote. So I think they realized they were wrong.*	*More people had purple as their answer in the beginning so I chose that. Then after the discussion people switched to yellow so I thought it more likely that was the right answer.*
Exp3	**MinGain3**	*In the first phase more kids picked blue so I thought it would probably be the best way to build the rocket. However in the second phase the kids only changed their minds because they were yelled at. I still think blue is probably the best choice because the kids only changed their minds because they felt intimidated by the green people yelling at them.*	*The yellow's argument seems to be pretty convincing, as three different people changed their answers to yellow. No one changed their answer to purple.*
		Express Anger	**Restrain Anger**
Exp4	**MinGain3**	*I felt, after the first round, that blue is by far the most likely, correct answer. After the second round, when the students talked, I still felt that blue is by far the most likely, correct answer because few students changed their mind about blue and they did so under duress.*	*At first the great majority thought that blue was best, and after they talked there was still a majority that thought blue was best, but not as much. I kept my answer as blue but I moved my sureness down the scale a bit.*
Exp4	**MajGain3**	*Before I sided with the majority as the most likely answer. Following the blue students shouting, I lost some confidence in their correctness as you shouldn't have to yell to prove you are right/ intimidate others into following your idea.*	*I think that since most initially said that blue would fly better then that is most likely the best way to go, and after they talked and stated their opinions, most that originally went with green as the best answer then changed to blue, so blue has to be the best option.*

I. Coding Scheme
   A. Trust
      1. Majority Rule vs. Minority rule
         a) which side does participant expect to have a better answer (e.g., unconventional-, or expertise-based for minority, or "more people said it" for majority)
   B. Emotion
      1. Anger_BadSign
         a) intimidation means they did ***not*** change the others' belief
         b) anger means they were wrong or insecure or forcing groupthink
      2. Anger_GoodSign
         a) anger was justified, for whatever reason
      3. Anger_MentionUnclear
         a) anything that mentions anger but it's not clear how they're evaluating it
   C. Persuasion / Conversion
      1. Conversion_GoodSign
         a) (the converts were genuinely convinced; evidence; good reasons)
      2. Conversion_BadSign
         a) (the converts were just uncertain; peer pressure)
      3. Conversion_NotGenuine
         a) participant explicitly denies that the converts were convinced
   D. Refuseniks
      1. Participants who say they don't have enough information, or that majorities aren't always right (but don't say that means that minorities are right)
   E. Miscellaneous
      1. Explanations that have aspects that don't seem to fit any of the categories above

II. Coding Instructions

A. For each, you're basically answering two questions: "does the participant mention the concept?", and "is it a good sign or bad sign for them?". For instance, Person A and B might mention that one group was angry while Person C doesn't; but Person A interprets one faction's anger as a sign that the other side said something stupid (for example) — in other words, anger was justified, and a "good sign" for angry side — while Person B interprets a faction's anger as a sign that the angry faction was wrong or insecure or forcing conformity, but they definitely did **not** *convince their opponents — in other words, anger was a bad sign for the angry side. In this case, Person A, B, and C would be in different categories: positive mention, negative mention, and no mention. Additionally, Person B might be in an orthogonal category, "Persuasion".*

III. Code Labels

A. Anger_GoodSign
B. Anger_BadSign
C. Anger_MentionUnclear
D. Conversion_GoodSign
E. Conversion_BadSign
F. Conversion_NotGenuine
G. MajorityRule
H. MinorityRule
I. Other_NotEnoughInformation
J. Other_Miscellaneous