

RESEARCH ARTICLE

Whether groups value agreement or dissent depends on the strength of consensus

James C. Mellody *

Center for Information Technology Policy, Princeton University, Princeton, New Jersey, United States of America

* jamesmellody@princeton.edu



Abstract

I investigate the conditions under which groups value agreement versus dissent in collective decision-making processes. I argue that which kind of contribution a group values more will depend on the strength of the consensus. As a consensus evolves from weak to moderate to strong, I predict that groups will prefer agreement, dissent, and then agreement again. These predictions are in line with a multi-phase decision-making process in which groups pursue sequential goals reflected in the evolving consensus: establishing an initial consensus, exploring alternative perspectives, and settling on a final decision. I test these predictions with data from the Reddit community *r/AmltheAsshole*, in which people make normative judgments of social situations. I find support for the predicted pattern, with one caveat: when the consensus is strong, groups exhibit no preference for agreement or dissent.

OPEN ACCESS

Citation: Mellody JC (2025) Whether groups value agreement or dissent depends on the strength of consensus. *PLoS One* 20(12): e0334850. <https://doi.org/10.1371/journal.pone.0334850>

Editor: Edmund Cheng, City University of Hong Kong, HONG KONG

Received: February 25, 2025

Accepted: October 2, 2025

Published: December 4, 2025

Copyright: © 2025 Mellody. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: Data is available through the Harvard Dataverse. DOI is here: <https://doi.org/10.7910/DVN/CDPIVS>. Replication code here: <https://github.com/jmellody/aita>.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Through collective intelligence, the aggregation of inputs from many people can produce solutions and decisions that outperform individual efforts [1–4]. Whereas collective intelligence once emerged only locally, through co-located groups and teams [5], the continued proliferation of the internet has extended the reach of collective intelligence to a wide variety of tasks—including evaluation [6], prediction [7], and innovation [8]—that now benefit from global networks of individuals offering thoughts, ideas, opinions, and judgments.

A large body of work has explored the role of social influence in collective intelligence—mostly investigating how, and under what conditions, social influence either inhibits the diversity of individual judgments and contributes to premature or suboptimal convergence [9,10] or alternatively offers a means of learning that facilitates convergence on optimal judgments from the crowd [11–15].

In addition to the process of making and updating judgments, another key process in collective intelligence is how the collective weights, or values, different kinds of judgments [16]. Much of the research on this process compares different weighting strategies to see which might be the most optimal [17–21]. Given the importance of social influence in the processes

of making and updating judgments, it is likely that social influence also plays an important role in the way that collectives value different kinds of judgments in practice.

In this article, I investigate the role of social influence in the process of valuing contributions to the collective. I use the terms ‘group,’ ‘collective,’ or ‘crowd’ to refer to a set of individuals contributing to a single collective decision. I examine how new judgments relate to existing judgments collected by the group and whether this relationship matters for the valuation of these new judgments. More specifically, I ask the following research question: how does the value that a group places on an individual judgment depend on whether the judgment agrees with, and thus bolsters, or disagrees with, and thus questions, the existing group consensus?

There is evidence that groups value both agreement and dissent—as both can facilitate decision making in different ways [22]. Groups value agreement because it fuels convergence on a decision, which is the ultimate goal of many collective intelligence tasks [23–27]. Additionally, groups may value agreement as a means not only of reaching a decision, but also of fostering cohesion, which can veer into groupthink when the pursuit of cohesion incentivizes conformity and crowds out alternative perspectives [28–31]. On the other hand, groups may value dissent—especially when they approach decision making as problem solving—because it represents diversity of thought, which can bolster decision quality [32,33]. Whether a particular group values agreement or dissent may depend on the nature of the task (e.g., where the task falls on the speed-accuracy tradeoff [34]) as well as group norms around consensus and criticality [35].

While particular groups may have a tendency to value agreement or dissent more, it is also possible that, within the course of a single decision-making process, conditions evolve such that groups alternate between valuing agreement more and valuing dissent more. Research on group decision making finds that decision making occurs in phases, across which the motivation of the group can shift. Much of this work suggests a multi-phase process, in which the group’s motivation shifts from general concerns around orienting to a problem, to exchanging ideas, and finally to settling on a solution [36–38]. While there is evidence that the specific dynamics of these phases depend on features of the task and the group [39–41], the broader goals (and thus phases) of establishing and orienting to the problem, exploring potential options, and moving toward a decision, are consistent across different decision-making sequences [40]. These broad phases thus offer a theoretical foundation for hypothesizing about how groups may value different kinds of contributions as a decision-making process evolves.

Further research suggests a link between a group’s motivation and how the group values different kinds of contributions, finding that an approaching deadline can induce a “need for closure,” in turn leading groups to reject dissenting opinions in favor of opinion uniformity [42–44]. However, prior work has not examined how a group’s valuation of different kinds of contributions may change dynamically across the phases of decision making. In this article, I address this by investigating whether and how the strength of group consensus serves as a social reference point, shaping whether individuals prioritize agreement or dissent. I argue that the strength of the group consensus offers a social reference point enabling individuals in the collective to track the phase of the decision-making process, such that whether a group places greater value on agreement or dissent will depend on the strength of the consensus.

More specifically—and drawing from research on the benefits of agreement and dissent as well as the phases of decision making—I predict a curvilinear relationship between the strength of consensus and the relative value placed on agreement versus dissent. I predict that when the consensus is weak, groups will value agreement more, in line with a goal of strengthening an initial consensus [23,24]; when the consensus is moderate, groups will value dissent more, in line with a desire to explore diverse perspectives and pressure test the

emerging consensus [32,33,36]; and when a consensus is strong, groups will once again value agreement more, in line with a need for closure [36,42–44].

The norms literature offers a potential alternative hypothesis, suggesting a monotonic relationship, such that dissenters are penalized more as group consensus strengthens [22,45]. However, because the present study examines how groups value contributions in specific decision-making contexts, rather than how dissenters are evaluated relative to broader societal norms, I predict that such increased punishment of dissenters will not apply here, and that instead the response to dissenters will evolve in a curvilinear fashion relative to the strength of the consensus.

Additionally, I examine whether the predicted pattern varies by the topic of the decision. Doing so offers an initial exploration into whether the pattern varies by features of the task at hand, which research suggests can shift the dynamics of the phases in decision-making processes [39,40]. While I predict that the general pattern should hold across topics, exploring heterogeneity offers additional insight into how these dynamics may vary, and can suggest potential paths forward for further research on the importance of decision context and topic to how groups value different kinds of contributions.

Setting and data

I test these predictions in the domain of online evaluations—an increasingly prominent context of collective intelligence [6]. I leverage data from the Reddit community *r/AmItheAsshole*—a forum that crowd-sources normative evaluations of social situations. In each thread on the forum, the original poster (OP) will post a contentious social situation from their life, and then other people make comments judging who was the “asshole” in the situation. Commenters make one of the following judgments: “you’re the asshole (YTA)” denoting the poster as the asshole, “not the asshole (NTA)” denoting the other party as the asshole, “everyone sucks here (ESH)” denoting all parties as assholes, “no assholes here (NAH)” denoting no parties as assholes, and finally “not enough info (INFO)” to convey that more information is needed to make a judgment. Fig 1 displays descriptive information about the community—including number of posts and users over time, and the distribution of judgments across posts over time.

I focused the analysis on comments made on posts that ultimately received a judgment of either YTA or NTA, because these two judgments were unambiguous. Posts judged as YTA and NTA made up the bulk of posts, as seen in Fig 1. I analyzed only comments that contributed a judgment of YTA or NTA, because these comments had an unambiguous relationship of dissent or agreement relative to the post judgments. I only analyzed top-level comments with judgments, as commenters are supposed to make judgments using top-level comments, according to the subreddit rules.

The rules of the subreddit assign a final post judgment based on the judgment of the comment that received the highest score 18 hours after the post was made. For the purposes of the analyses here, I measured the final judgment in two ways, and then limited the sample to posts for which these two measures agreed (97% of all posts). I did this to ensure that posts in the analyses had an unambiguous judgment, and to capture a more complete picture of the judgment based on the contributions of all in the community, rather than just a single comment. First, I captured the judgment of the single comment that received the highest score. Second, I added up the scores received by comments of each judgment type (“NTA”, “YTA”, “NAH”, “ESH”, or “INFO”) to see which judgment received the highest score (see *S1 Appendix*).

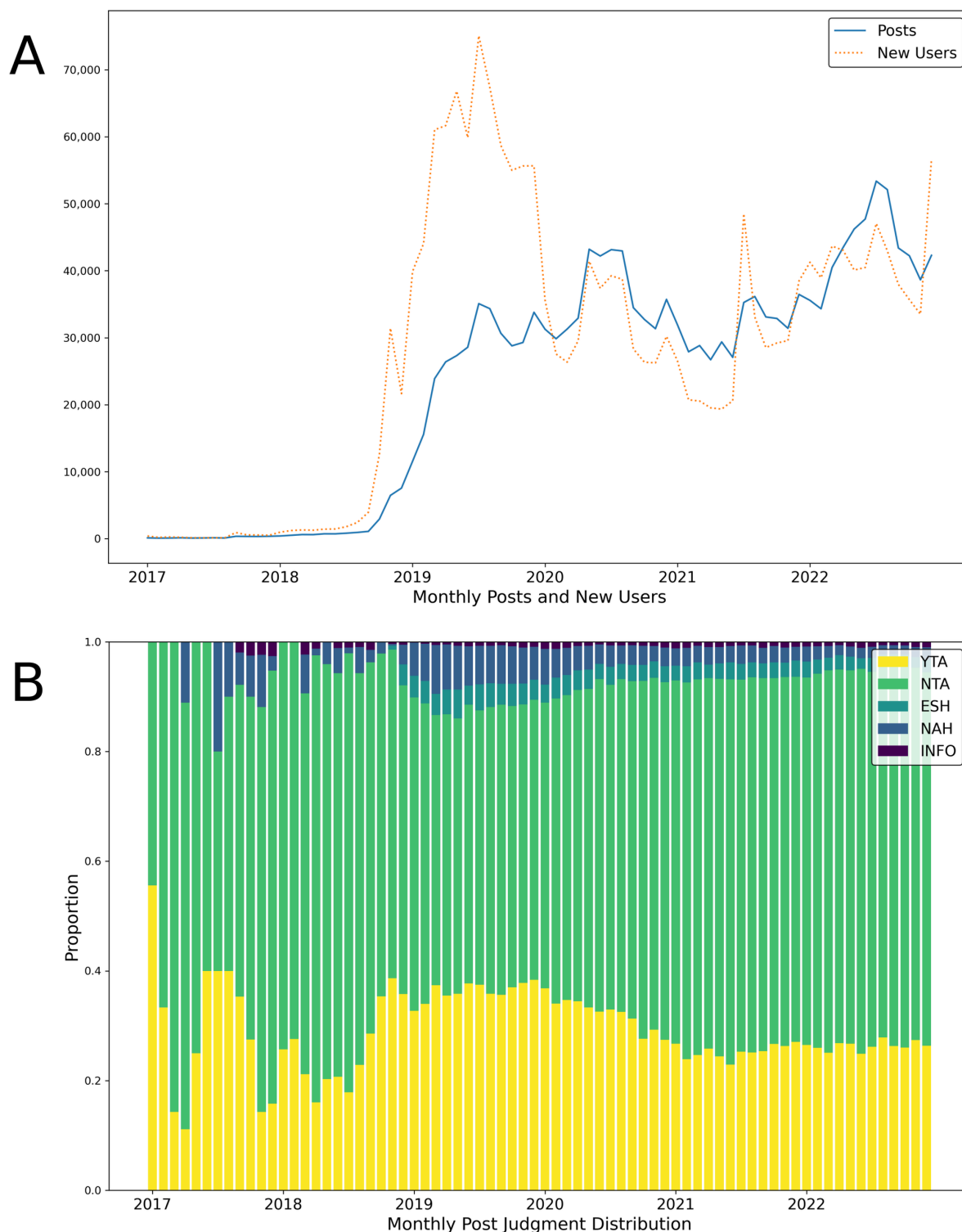


Fig 1. Information about r/AmItheAsshole. Data is from between 2017, before which subreddit activity was limited, and 2022. (A) Number of posts and users each month. The subreddit rapidly increased in popularity in late 2018. (B) Proportion of posts with judgments YTA, NTA, ESH, NAH, and INFO each month (limited to posts with > 14 comments).

<https://doi.org/10.1371/journal.pone.0334850.g001>

I then tracked the evolving strength of the consensus judgment of the post—a measure I call *Consensus Strength*. I measured this as a function of both the score-weighted proportion of comments at time t that expressed the majority judgment (I tested the robustness with an unweighted measure—i.e., proportion of comment counts—and found similar results; see *S5 Appendix*), and the logged number of comments with judgments at time t to capture whether the proportion variable reflects many or few comments. In addition to data limitations which exclude the possibility of measuring the evolving score of a single comment, I measured the consensus strength across all comments with judgments in order to capture the evolving consensus of the full group, rather than focusing on the single comment that received the most attention. Doing so captures a more complete picture of the evolving consensus of the group.

In the analysis, I modeled the logged number of upvotes that each comment received—a measure of how the group valued the comment—as a function of the strength of the consensus at the time the comment was made as well as a binary variable indicating whether the focal comment dissented or agreed with the current consensus, along with a number of other control variables. See *Materials and Methods* for more information on the data, measures, and analyses.

Results

Main result

The results of the main models offer evidence in support of the prediction that the relative value placed on dissent versus agreement would vary with the strength of consensus, as well as evidence largely in line with the more specific predictions for how this relationship would vary as the consensus evolved. The main result is a curvilinear relationship between the strength of the consensus and the relative value placed on dissent versus agreement. [Fig 2](#) displays the average marginal effect of dissenting with the current consensus on the predicted score (log) of a comment, across different consensus strengths, using a quadratic and then a cubic term to model the evolution of the strength of the consensus. Both models corroborate the first two phases of the core theoretical prediction: a shift from valuing agreement when consensus is weak, to valuing dissent when consensus is moderate. The models diverge in the third phase, in which the quadratic model shows a shift back toward valuing agreement, whereas the cubic model shows a shift toward indifference between dissent and agreement at the highest levels of consensus. This result from the cubic model goes against the prediction for the third phase, which was that groups would prefer agreement in order to close out the decision-making process.

Model diagnostics show identical AIC and BIC values (1.45×10^7), suggesting both models offer a similar fit versus complexity tradeoff, but I will detail the cubic results here, given that the model appears to be picking up an additional statistically significant inflection point in the data. When the consensus strength was very weak (i.e., two standard deviations below the mean), dissenting comments were discounted, receiving an average score that was 14.20% lower ($p < 0.001$) than the average score of agreeing comments. As the consensus got a bit stronger (i.e., one SD below the mean) the relationship reversed dramatically: dissenting comments received an average score that was 33.97% higher ($p < 0.001$) than the average score of agreeing comments. At the mean, dissent was still highly favored: dissenting comments received an average score that was 28.06% higher ($p < 0.001$) than the average score of agreeing comments. As the strength of the consensus further increased (i.e., one SD above the mean), dissent was still slightly favored: dissenting comments received an average score that was 4.49% higher ($p < 0.001$) than the average score of agreeing comments. Finally, once the consensus was very strong (i.e., two SD above the mean), there was no significant difference

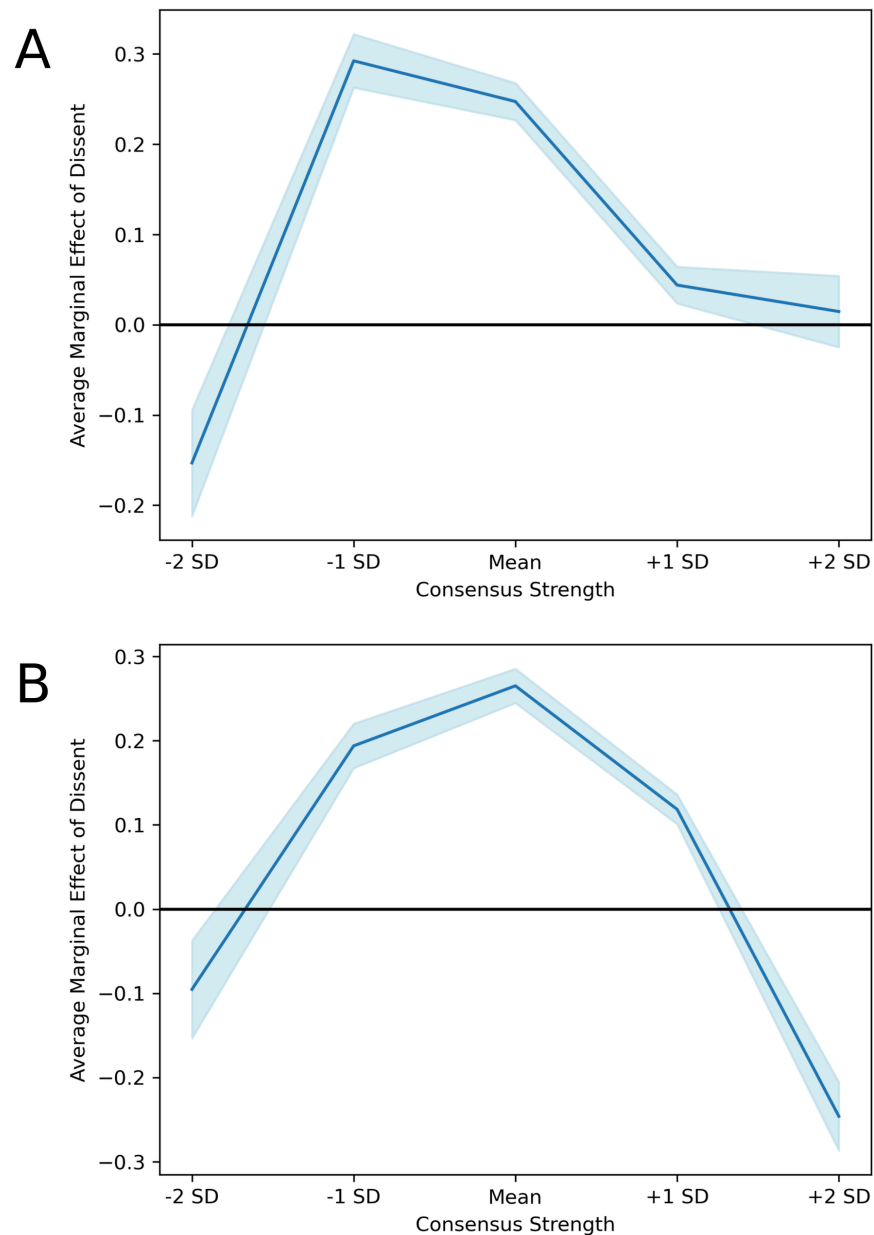


Fig 2. Main result. The main results from A) a model with a cubic term for consensus strength and B) a model with a quadratic term for consensus strength. Results are robust in both models through the second phase, after which they diverge. In the cubic model, there is no preference in this third phase, while the quadratic model shows a shift back to valuing agreement.

<https://doi.org/10.1371/journal.pone.0334850.g002>

in preference for dissenting versus agreeing comments. This main pattern was largely robust across a variety of criteria for filtering the sample (see *S5 Appendix*). See *S4 Appendix* for full regression results from the main models.

I conducted an additional analysis to examine whether the relationship between comment length and comment score differs by agreement with the consensus and the strength of

that consensus. Comment length—often a proxy for quality or the depth of explanation—was interacted with consensus strength (cubed) and a binary dissent indicator.

Marginal effects reveal that length matters most when consensus is weakest, and much less when consensus is strong. At a consensus strength 2 SD below the mean, doubling the length of an agreeing comment is associated with a 28.1% increase in score, while a dissenting comment sees a 21.6% increase. At the mean consensus level, these effects drop sharply to 2.0% for agreeing comments and 3.7% for dissenting comments. Once consensus reaches 2 SD above the mean, the boost from doubling length is <1% for agreeing comments and statistically negligible for dissenting comments.

These results suggest that explanations and elaboration play their largest role when consensus is not yet firmly established—possibly because they help sway opinion during this formative stage. As consensus solidifies, such persuasive efforts have diminishing returns. Notably, the pattern also shifts across the consensus trajectory: length appears to matter more first for agreeing comments, then for dissenting comments, before becoming largely irrelevant in the final phase.

Regression discontinuity in time

To investigate the causal relationship between the strength of consensus and how agreement and dissent are valued, I conducted a regression discontinuity in time (RDiT) analysis [46]. This is a regression discontinuity design in which time operates as the running variable. The goal of this analysis was to see if the kind of contributions the group values would shift after a shock to the strength of the consensus. I used instances when the original poster (OP) of a post made a comment on their own post as a shock to the strength of the consensus. The logic is that the OP commenting on their own post provides new information, painting a more complete picture of the situation and thus lending strength to the consensus. See the *S7 Appendix* for graphical evidence of the jump in the strength of consensus when OP posts a comment. The shift from pre- to post-OP making a comment is not a transition into a fully established consensus, but rather a transition from a first phase of building an initial consensus to a second phase of beginning to explore alternative perspectives. The expectation is that there should be a change in the kind of contribution valued by the group: away from valuing agreement and toward valuing dissent.

I used the time difference between the focal comment and the first comment made by the original poster (OP) as the running variable, and I limited the analysis to comments made 150 minutes before and after the OP's first comment, as RDiT estimates a local treatment effect around the event [46]. Fig 3 shows the discontinuity around the first OP comment: the y-axis displays the mean difference in log score between dissenting and agreeing comments. Before OP made a comment, agreement was valued more, whereas after, there was a shift toward valuing dissent, in line with a move toward exploration. Predictive margins from the RDiT regression offer statistical support in line with these graphical results. Before OP made a comment, dissenting comments received an average log score of 0.74, which was 51% lower ($p < 0.001$) than the average log score of agreeing comments (1.51). After the OP made a comment, dissenting comments received an average log score of 1.48, which was 3.10% higher ($p > 0.001$) than the average log score of agreeing comments (1.44). These results offer evidence that a shock to the strength of the consensus can shift the kind of contribution valued by the group, and more specifically, that an increase in the strength of the consensus can shift the group away from valuing agreement and toward valuing dissent.

A potential concern with this RDiT analysis is that the OP's decision to intervene and comment is dependent on the amount of negativity expressed toward the OP. In other words,

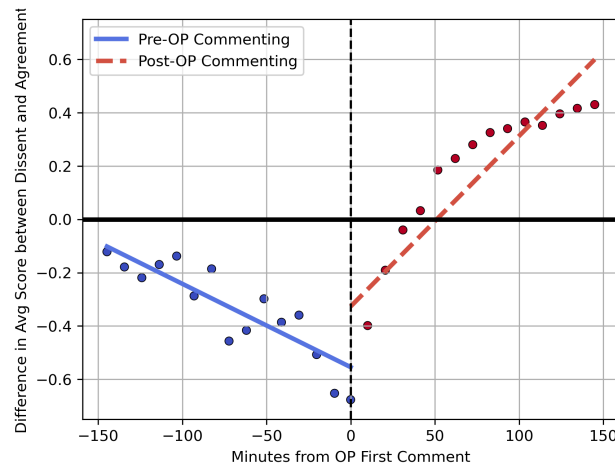


Fig 3. Local discontinuity. This graph shows the local discontinuity in the outcome of interest (difference in average log score between comments that dissented versus agreed with the consensus). This represents the local effect of the original poster's (OP's) first comment based on a small bandwidth around the time of treatment: 150 minutes before and after the OP first commented on the post. Outcomes are binned for each 10-minute interval. Before the OP's first comment, agreement was valued more, whereas after, there was a shift toward valuing dissent.

<https://doi.org/10.1371/journal.pone.0334850.g003>

the OP may be more likely to comment if the consensus is going against them. To examine whether this pre OP-commenting level of negativity matters for the main effect, I reran the RDIT analysis controlling for the level of pre OP-commenting negativity. To do this, I tracked comments expressing YTA as +1 and comments expressing NTA as -1, weighted by the score of these comments, to produce a score where a larger number reflects greater negativity toward the OP. I also reran the analysis controlling for the strength of the consensus just before the OP commented. For both of these robustness checks, the core finding remains the same: dissent is valued more after the OP comments compared to before, and the strength of this finding remains similar.

Exploring heterogeneity across topics

The nature of the data from r/AmItheAsshole enables further testing of the robustness of the main result. Specifically, I can test whether the decision-making pattern depends on the topic of the post, or whether the pattern is consistent across topics. To explore these possibilities, I broke the analysis down by post topic. I ran these models with data from the full set of years (i.e., 2013–2022), in order to increase the sample size for each topic. I also leveraged cubic models here, to more flexibly model potential heterogeneity in the patterns. The goal here was to test the generalizability of the main result across different substantive topic areas. To identify the topics of posts, I ran a machine learning model, Top2Vec [47], which makes use of joint document and word embeddings to identify topic vectors. See *Materials and Methods* for more information. The 10 primary topics identified cover a wide range of social situations: health, neighbors, money, events, chores, dating, bigotry, social media, school, and work. Fig 4A displays a two-dimensional UMAP representation of the post embeddings, colored by primary topic.

Fig 4B displays the average marginal effects of dissenting, from the models with the data sub-setted to each of the 10 identified topics. The graphs show a relatively stable pattern across these 10 topics. Variation comes mainly at the extreme ends of the consensus strength

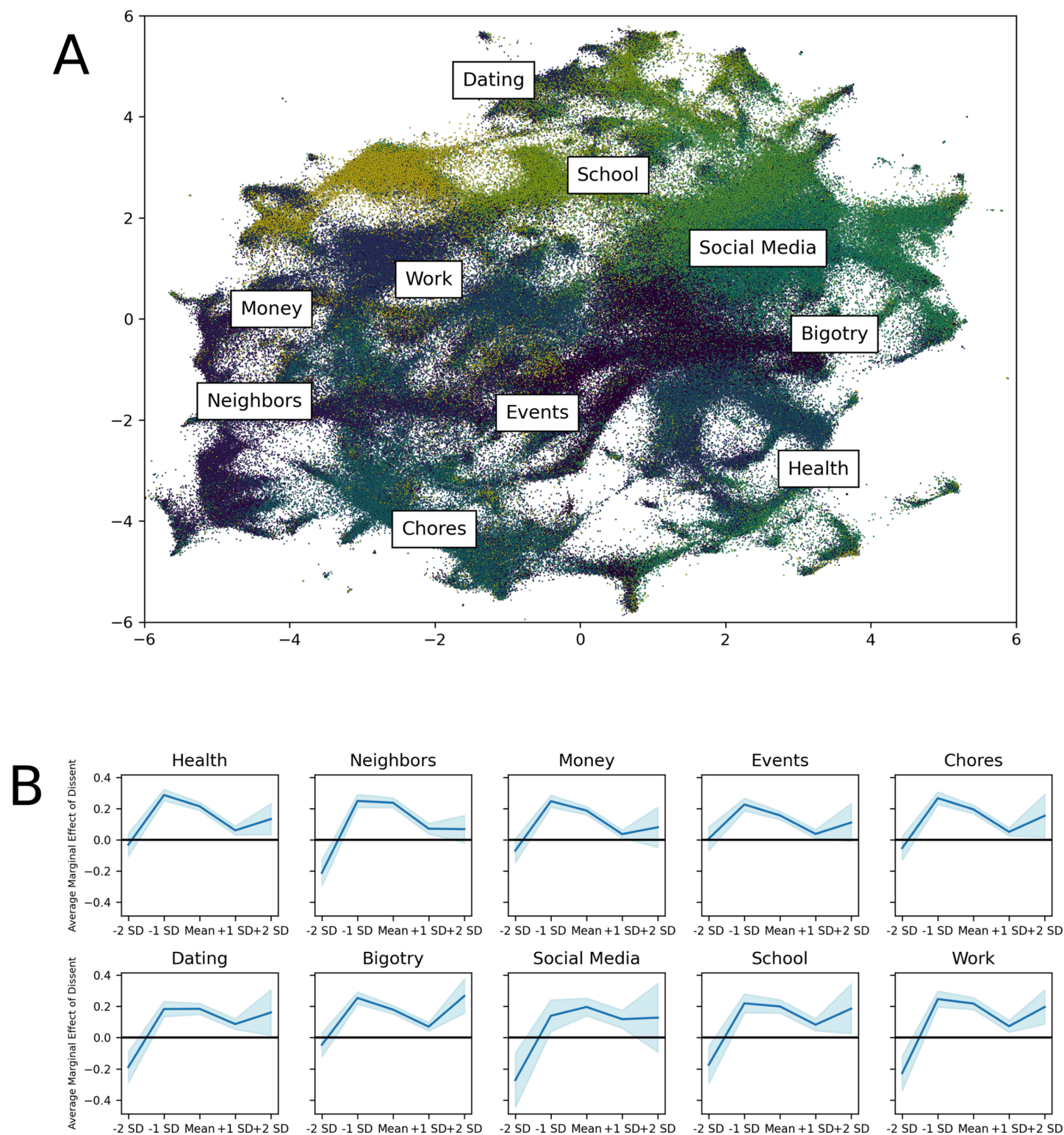


Fig 4. Post topics. (A) 2-dimensional UMAP projection of post embeddings generated through Top2Vec model.
(B) Average marginal effects of dissent across topics.

<https://doi.org/10.1371/journal.pone.0334850.g004>

spectrum. For some topics (i.e., neighbors, dating, social media, school, and work), agreeing comments were valued more when the consensus was weak—in line with the main pattern, whereas for other topics (i.e., health, money, events, chores, and bigotry), there was not a statistically significant difference between the scores received by dissenting and agreeing comments when the consensus was weak. This suggests that the initial phase of the decision-making process may differ depending on the topic—sometimes the focus of this phase may be valuing agreement to strengthen the consensus, and other times the focus may be on gathering all kinds of contributions.

Similarly, there was variation in how groups valued dissent versus agreement in the final phase: for most topics, in line with the main result, groups exhibited no preference between agreeing and dissenting comments, suggesting that contribution type became less salient once the decision trajectory was established. For a few topics—most significantly bigotry—there was a movement back toward valuing dissent more at the strongest levels of consensus. This finding suggests that the salience of dissenting perspectives in final-phase decision making may depend on the complexity or stakes of the judgment domain, with some contexts requiring ongoing scrutiny even after apparent consensus emergence.

Discussion

The present study investigated how groups value agreement and dissent throughout the process of making a decision. Leveraging data from the Reddit community *r/AmItheAsshole*, I found evidence supporting the prediction that the relative value placed on agreement versus dissent would vary with the strength of the consensus. Specifically, I found evidence of a curvilinear relationship between the strength of the consensus and the relative value placed on dissent versus agreement. Groups valued agreement over dissent when a consensus was very weak, placed greater value on dissent when the consensus was moderate, and trended toward not having a preference as the consensus solidified.

This study offers two main contributions. First, this study contributes to our understanding of social influence in collective intelligence. Whereas past work has focused on the role of social influence in how individuals make and update their judgments [9–15], here I focus on the role that social influence plays in how groups value different kinds of contributions. Specifically, I find that the group's existing consensus operates as a reference point, such that individuals in the group value new contributions based on how these contributions relate to the existing consensus. Together with prior research on social influence in collective intelligence, this helps to paint a more complete picture of exploration and convergence in collective intelligence. Whether, and how, groups explore diverse perspectives and/or converge on particular judgments depends both on how judgments are made and updated, and also on how the group values different kinds of judgments. These processes are interrelated in collective intelligence, as individuals are motivated to take actions that they see receiving recognition [48].

The robustness of main result was further confirmed through additional functional form testing. Cubic and quadratic specifications of consensus strength yielded substantively similar results for the core theoretical prediction—groups initially valued agreement, then shifted to valuing dissent as consensus moved from weak to moderate levels. However, the cubic model revealed additional nuance in the final phase: rather than returning to strong agreement preference, groups with very strong consensus showed indifference between agreement and dissent contributions. This finding suggests that late-stage decision making may be characterized by reduced sensitivity to contribution type, rather than by active agreement-seeking.

Second, this study contributes to research on the phases of group decision making [36–38]. The results offer evidence in support of a three-phase decision-making process, in which groups move from orienting to the problem and building an initial opinion, to evaluating and considering different opinions, and finally to settling on a decision [36]. The present study shows that these different phases have implications for how groups value dissent versus agreement. This matters because decision quality relies on a group's ability to consider diverse perspectives and avoid converging on suboptimal solutions or ideas.

These findings also connect with and offer insight to theories on the role of disagreement in democratic decision making. Classic political theory, such as Mansbridge's *Beyond Adversary Democracy* [49], distinguishes between consensus-oriented and adversarial modes of democratic practice, raising the question of when agreement should be sought versus when disagreement should be encouraged. More recent empirical and theoretical work suggests that moderate levels of disagreement can enhance deliberation by fostering problem-solving, prediction, and positive dissensus [50], while too little or too much disagreement can undermine decision quality [51]. The curvilinear pattern observed here—in which groups shift from valuing agreement, to dissent, and then to reduced preference at very strong consensus—aligns with arguments for a dynamic, phase-sensitive approach to democratic deliberation [52]. In this view, the value of dissent is not fixed but depends on where a group is in its decision trajectory, echoing calls for institutions and facilitation strategies that adapt to the deliberative phase [53].

While the specific patterns of valuing agreement and dissent identified in this study may generalize to other online evaluation contexts with similar conditions, it is an open question how far-reaching these patterns are in other settings. The topic-level heterogeneity analyses provide an initial lens on this question. Two noteworthy variations emerged.

First, there was heterogeneity in the initial phase of the decision-making process. Topics that were primarily interpersonal in nature (e.g., neighbors, dating, social media, school, work) followed the main pattern, with groups initially favoring agreement. In contrast, topics that were less directly interpersonal (e.g., health, money, events, chores, bigotry—the last of which is both interpersonal and individual/cognitive) showed no significant early preference. One potential explanation for this divergence draws on structuration theory as applied to small group contexts (e.g., [54–56]). This perspective suggests that decisions about interpersonal matters may follow a logic of “assumed consensus” [55], such that groups are inclined to seek early agreement as a way to reaffirm shared norms and social structure, effectively “grounding” the discussion in a collectively recognized framework before ironing out the details.

Second, in the final phase of consensus evolution, decisions about bigotry diverged from the main pattern. Here, groups tended to value dissent even at high levels of consensus strength, perhaps because of the heightened sensitivity and potential consequences of the topic, leading participants to probe and test agreement more rigorously before closure.

Future research could explore the conditions shaping group preferences for agreement versus dissent at both the early and late stages of consensus building. One such condition may be the group's entativity—the extent to which members perceive themselves as a unified entity versus a loose collection of individuals [57–59]. In online forums such as Reddit, entativity may be relatively low compared to in-person groups, which could influence receptiveness to different kinds of contributions. Relatedly, group boundaries, their permeability, and the group's task orientation [35,60,61] may moderate the evolution of preferences for agreement and dissent. Investigating these factors could clarify both the generalizability of the present findings and the mechanisms through which they arise. Another area for future research is

examining whether different patterns of alternating between valuing agreement and dissent are more or less optimal for achieving accurate or otherwise effective decisions.

Materials and methods

Data and sample

Data was gathered through data dumps provided by Pushshift.io [62]—available from the start of the subreddit in 2013 through the end of 2022, and in accordance with applicable data terms and conditions. For the main analyses, I used matched comments on posts made in 2022—consisting of 6,799,071 comments—because some rules of the subreddit have changed over time and I wanted to capture a period of subreddit stability and maturity, and this is the last full year of Reddit data available. I conducted a robustness check leveraging comments on posts from all years (i.e., 2013–2022), finding largely consistent results (see *S5 Appendix* for this and other robustness checks). Because the goal of the study was to look at the valuation of contributions throughout the course of a decision-making process, I limited the sample to posts which received a minimum of 15 comments with judgments so that each post had a critical mass of judgments. Findings were robust to a different choice of minimum comment threshold. In the main model, I included comments from posts whose main text was deleted by the author or removed by the moderator team, and found the results to be consistent if I excluded these comments. Note that it was a relatively common occurrence for posters to remove their post eventually, but the Reddit AutoModerator bot would repost the original post's content so that people could keep commenting. In the main analyses, I also removed comments made when the existing consensus was fully balanced (i.e., the proportion variable was equal to 0.5 exactly). This is because the relationship between these comments and the existing consensus was ambiguous. I ran robustness checks including these comments (demarcating them as being in agreement with the existing consensus because a comment made when the consensus is fully balanced effectively establishes a new consensus), and found largely consistent results. There was one important difference to note: when including these comments, at low consensus strength, there was no difference in valuing dissent vs. agreement, possibly because these comments tended to occur when the consensus was weak—because a balanced consensus often occurs early on with only a few comments, meaning the consensus strength is weak. The unclear nature of how these comments relate to the existing consensus likely adds noise at this spectrum of consensus strength, thus why I exclude these comments from the main analysis.

I also limited the analyses to only include comments that received scores equal to or greater than 1. A comment score can take on any integer value. Comments start out with a score of 1, and then can move to a score of 0 and into negative scores if the comment receives more downvotes than upvotes. I limited the sample in this way because the rules of r/AmItheAsshole state that downvoting is reserved not for expressing disapproval of a comment's judgment, but instead for signaling that a comment is off topic. This means that comments receiving more downvotes than upvotes (i.e., those with a score of 0 or lower) were likely to be irrelevant or off-topic comments, which I wanted to exclude from the analyses.

Measures

The dependent variable is the score received by a comment. I log-transformed this variable because it is right-skewed. The score was not captured at the exact time the comment was made, but rather after a time lag [62]. Note that the Pushshift project routinely re-retrieved old data from Reddit, such that all comment scores here effectively reflect the final comment

score. This is beneficial because a time lag is required for comments to accrue upvotes and downvotes. Given that research shows that attention to social media posts and comments decays exponentially after the time of posting [63], the comment score provides the best available approximation of the comment's reception close to its original creation time.

In measuring the strength of the consensus, I created a composite measure taking two elements into account. First, in line with measurements of consensus in discrete choice scenarios [64], I measured the strength of the current majority consensus. To measure this, I captured a running tally of the number of comments with the judgment "YTA" and a second running tally of the number of comments with the judgment "NTA." I then captured the proportion of comments belonging to the majority judgment—meaning a score from (0.5, 1]. This was an unweighted measure of consensus. I also constructed a weighted measure of consensus, as users can upvote a comment if they agree with it, and so the strength of consensus can reflect not just the count of different judgments, but also the number of upvotes that the comments expressing these judgments have received. To construct this weighted measure of consensus, I multiplied each comment by its score, using the sum of these weights for all "YTA" and "NTA" comments to represent the total weighted opinion for each. Note that the weighted and unweighted measures yield similar results, suggesting that the finding is robust to different ways of capturing the consensus strength. I then calculated the proportion of weighted comments contributing to the majority—again a score from (0.5, 1]. I used the weighted measure in the main analysis, and conducted a robustness check with the unweighted measure, finding consistent results (see *S5 Appendix*). Second, I then multiplied the weighted or unweighted proportion variable by the logged number of other comments made at the time of the focal comment. The goal here was to contextualize the proportion variable with information on the volume of comments contributing to that proportion. See the *S2 Appendix* for more information on these measures.

I also measured whether a comment dissented or agreed with the current consensus: *Consensus Dissent*. This is a binary variable—0 if the comment agreed with the consensus, and 1 if it dissented. To capture whether a comment dissented or agreed with the consensus, I measured the judgment of the comment and whether that judgment agreed with the judgment that was currently in the majority under the current consensus. So, for example, if "YTA" currently had a 65% majority, and the focal comment expressed "YTA," then it was assigned a 0 for this variable, denoting agreement. If the focal comment expressed "NTA," then it was assigned a 1, denoting dissent.

I included several control variables that likely impacted the score received by a comment. I included the length of the comment based on the number of characters: *Comment Length (ln)*, the time in minutes since the original post: *Minutes Since Post (ln)*, the number of other comments in the thread at the time of the focal comment (as a measure of competition for attention)—I added 1 to this variable before taking the log: *Comment Competition (ln)*, the score of the comment author's other comments made previously in the subreddit (as a measure of author skill at expressing a judgment)—I used STATA's *Inskew0* command to take the log of this variable which ranged from negative to positive values: *Author Score (ln)*, and finally, indicators to capture time-invariant characteristics of the post in which the comment was made, the month of the year in which the comment was made, the hour of the day in which the comment was made, and the day of the week on which the comment was made. See the *S3 Appendix* for descriptive characteristics and correlations for all variables.

Analyses

I ran linear regressions predicting the score (log) received by each comment. I estimated the following regression equation:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 CS_i + \beta_3 CS_i^2 + \beta_4 CS_i^3 + \beta_5 D_i CS_i + \beta_6 D_i CS_i^2 + \beta_7 D_i CS_i^3 + \delta X_i + \alpha_i + \alpha_m + \alpha_d + \alpha_h + \epsilon_i$$

Where y_i is the logged score of the focal comment, D_i is an indicator for whether the comment dissents with the existing consensus, CS_i is the current strength of the consensus. β_7 is the main quantity of interest: the interaction effect between the dissent variable and the cubed consensus strength term, to account for possible non-linearity. X_i is a vector of covariates and δ a vector of accompanying coefficients, α_i is a fixed effect for the post in which the focal comment appears, α_m is a month fixed effect, α_d is a day-of-week fixed effect, α_h is an hour-of-day fixed effect, and ϵ_i is the error term. To simplify the interpretation of the interaction effect of interest, I estimated the average marginal effect of dissenting with the consensus, at different consensus strengths, using the STATA margins command.

In order to move toward a causal interpretation of the findings, I took two additional steps. First, I used a coarsened exact matching procedure [65,66] to generate weights for each observation, which I then included in the main regression. This has the dual benefit of improving covariate balance between comparison units (here comments that dissent versus agree with the consensus) and also of operating as a regression pre-processing step that reduces model dependence [67]. I coarsened four variables (i.e., the strength of the consensus, the time since the post in minutes, the score of the author's other comments in the subreddit, and the length of the comment) into 12 equally sized quantile bins, and a fifth variable (i.e., the number of other comments) into 25 equally sized quantile bins. This process generated weights for each observation, based on how many times each control observation (i.e., agreeing comment) was used as a match for a treated observation (i.e., dissenting comment). I included these weights in the main regression. See S6 Appendix for balance table with pre- and post-matching covariate balance, which shows a significant improvement from matching.

Second, I conducted a regression discontinuity in time analysis [46]. I estimated the following equation:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 OP_i + \beta_3 D_i OP_i + \gamma f(TOP_i) + \delta X_i + \alpha_i + \alpha_y + \alpha_m + \alpha_d + \alpha_h + \epsilon_i$$

Where y_i is the logged score of the comment, D_i is an indicator for whether the comment dissents with the existing consensus, OP_i is an indicator for whether the original poster (OP) has posted a comment yet or not, and TOP_i is the running variable (i.e., minutes from first OP comment). Identification comes from the assumption that the potentially endogenous relationship between ϵ_i and the running variable TOP_i is captured by the flexible function $f(\cdot)$ [68]. To estimate this, I included in the model a three-way interaction between TOP_i , D_i , and OP_i , as well as the two-way interactions between TOP_i and each of these other two variables. X_i is a vector of covariates and δ a vector of accompanying coefficients, α_i is a fixed effect for the post in which the focal comment appears, α_y is a year fixed effect, α_m is a month fixed effect, α_d is a day-of-week fixed effect, α_h is an hour-of-day fixed effect, and ϵ_i is the error term. To simplify the interpretation of the interaction effect of interest, I estimated the marginal effect of dissenting with the consensus, before and after the OP's first comment. I used data from all years for this analysis because data in the small time threshold (150 minutes before/after OP commented) is limited in any given year.

For the heterogeneity in topic analyses, I identified post topics using the Top2Vec algorithm [47]. Top2Vec calculates topic vectors using the following steps: 1) generate joint document and word embeddings—I used the Doc2Vec algorithm here, 2) create lower dimensional embeddings with UMAP, 3) identify dense areas—topics—using HDBSCAN, 4) then calculate the centroid of document vectors in each dense area—this is the topic vector for each dense area, 5) locate the n-closest word vectors which comprise the resulting topic words. I then ran a hierarchical topic reduction algorithm, to reduce the number of topics to ten, and assigned a topic to each post based on the best fit topic. I labeled the topics by inspecting representative topic words from each. See the *S8 Appendix* for example representative topic words from each of the ten topics. I then ran ten regressions, subsetting to each of the topics in turn.

Supporting information

S1 Appendix. Information on how judgments were captured from comments.

(PDF)

S2 Appendix. Measuring consensus strength.

(PDF)

S3 Appendix. Descriptive statistics and correlations.

(PDF)

S4 Appendix. Regression table for main results.

(PDF)

S5 Appendix. Regression tables for robustness checks.

(PDF)

S6 Appendix. Matching balance table.

(PDF)

S7 Appendix. RDiT details.

(PDF)

S8 Appendix. Topic analysis details.

(PDF)

Acknowledgments

Thank you to the Editor and three Reviewers for very helpful suggestions in improving this paper. Thank you to Roni Luo for helpful discussions throughout the development of this paper.

Author contributions

Conceptualization: James Corbett Mellody.

Data curation: James Corbett Mellody.

Formal analysis: James Corbett Mellody.

Investigation: James Corbett Mellody.

Methodology: James Corbett Mellody.

Project administration: James Corbett Mellody.

Resources: James Corbett Mellody.

Software: James Corbett Mellody.

Supervision: James Corbett Mellody.

Validation: James Corbett Mellody.

Visualization: James Corbett Mellody.

Writing – original draft: James Corbett Mellody.

Writing – review & editing: James Corbett Mellody.

References

1. Galton F. One vote, one value. *Nature*. 1907;75(1948):414–414. <https://doi.org/10.1038/075414a0>
2. Lorge I, Fox D, Davitz J, Brenner M. A survey of studies contrasting the quality of group performance and individual performance, 1920–1957. *Psychological Bulletin*. 1958;55(6):337.
3. Surowiecki J. *The wisdom of crowds*. 2005.
4. Larrick RP, Soll JB. Intuitions about combining opinions: misappreciation of the averaging principle. *Management Science*. 2006;52(1):111–27. <https://doi.org/10.1287/mnsc.1050.0459>
5. Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW. Evidence for a collective intelligence factor in the performance of human groups. *Science*. 2010;330(6004):686–8. <https://doi.org/10.1126/science.1193147> PMID: 20929725
6. Dellarocas C. The digitization of word of mouth: promise and challenges of online feedback mechanisms. *Management Science*. 2003;49(10):1407–24. <https://doi.org/10.1287/mnsc.49.10.1407.17308>
7. Arrow KJ, Forsythe R, Gorham M, Hahn R, Hanson R, Ledyard JO. *The promise of prediction markets*. 2008.
8. Boudreau KJ, Lacetera N, Lakhani KR. Incentives and problem uncertainty in innovation contests: an empirical analysis. *Management Science*. 2011;57(5):843–63. <https://doi.org/10.1287/mnsc.1110.1322>
9. Lorenz J, Rauhut H, Schweitzer F, Helbing D. How social influence can undermine the wisdom of crowd effect. *Proc Natl Acad Sci U S A*. 2011;108(22):9020–5. <https://doi.org/10.1073/pnas.1008636108> PMID: 21576485
10. Muchnik L, Aral S, Taylor SJ. Social influence bias: a randomized experiment. *Science*. 2013;341(6146):647–51. <https://doi.org/10.1126/science.1240466> PMID: 23929980
11. Becker J, Brackbill D, Centola D. Network dynamics of social influence in the wisdom of crowds. *Proc Natl Acad Sci U S A*. 2017;114(26):E5070–6. <https://doi.org/10.1073/pnas.1615978114> PMID: 28607070
12. Jayles B, Kim H-R, Escobedo R, Cezera S, Blanchet A, Kameda T, et al. How social information can improve estimation accuracy in human groups. *Proc Natl Acad Sci U S A*. 2017;114(47):12620–5. <https://doi.org/10.1073/pnas.1703695114> PMID: 29118142
13. Bernstein E, Shore J, Lazer D. How intermittent breaks in interaction improve collective intelligence. *Proc Natl Acad Sci U S A*. 2018;115(35):8734–9. <https://doi.org/10.1073/pnas.1802407115> PMID: 30104371
14. Becker J, Porter E, Centola D. The wisdom of partisan crowds. *Proc Natl Acad Sci U S A*. 2019;116(22):10717–22. <https://doi.org/10.1073/pnas.1817195116> PMID: 31085635
15. Almaatouq A, Noriega-Campero A, Alotaibi A, Krafft PM, Moussaid M, Pentland A. Adaptive social networks promote the wisdom of crowds. *Proc Natl Acad Sci U S A*. 2020;117(21):11379–86. <https://doi.org/10.1073/pnas.1917687117> PMID: 32393632
16. Galton F. The ballot-box. *Nature*. 1907;75(1952):509–10. <https://doi.org/10.1038/075509f0>
17. Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD. Optimally interacting minds. *Science*. 2010;329(5995):1081–5. <https://doi.org/10.1126/science.1185718> PMID: 20798320
18. Davis-Stober CP, Budescu DV, Dana J, Broomell SB. When is a crowd wise?. *Decision*. 2014;1(2):79–101. <https://doi.org/10.1037/dec0000004>
19. Budescu DV, Chen E. Identifying expertise to extract the wisdom of crowds. *Management Science*. 2015;61(2):267–80. <https://doi.org/10.1287/mnsc.2014.1909>

20. Mahmoodi A, Bang D, Olsen K, Zhao YA, Shi Z, Broberg K, et al. Equality bias impairs collective decision-making across cultures. *Proc Natl Acad Sci U S A*. 2015;112(12):3835–40. <https://doi.org/10.1073/pnas.1421692112> PMID: 25775532
21. Palley AB, Satopää VA. Boosting the wisdom of crowds within a single judgment problem: weighted averaging based on peer predictions. *Management Science*. 2023;69(9):5128–46. <https://doi.org/10.1287/mnsc.2022.4648>
22. Jetten J, Hornsey MJ. Deviance and dissent in groups. *Annu Rev Psychol*. 2014;65:461–85. <https://doi.org/10.1146/annurev-psych-010213-115151> PMID: 23751035
23. FESTINGER L. Informal social communication. *Psychol Rev*. 1950;57(5):271–82. <https://doi.org/10.1037/h0056932> PMID: 14776174
24. Schachter S. Deviation, rejection, and communication. *The Journal of Abnormal and Social Psychology*. 1951;46(2):190.
25. Moscovici S, Mugny G, Van Avermaet E. Perspectives on minority influence. Cambridge University Press; 1985.
26. Nemeth CJ. Differential contributions of majority and minority influence.. *Psychological Review*. 1986;93(1):23–32. <https://doi.org/10.1037/0033-295x.93.1.23>
27. Whyte G. Groupthink reconsidered. *The Academy of Management Review*. 1989;14(1):40. <https://doi.org/10.2307/258190>
28. Asch SE. Opinions and social pressure. *Sci Am*. 1955;193(5):31–5. <https://doi.org/10.1038/scientificamerican1155-31>
29. Gero A. Conflict avoidance in consensual decision processes. *Small Group Behavior*. 1985;16(4):487–99. <https://doi.org/10.1177/104649648501600405>
30. Marques JM, Paez D. The ‘Black Sheep Effect’: social categorization, rejection of ingroup deviates, and perception of group variability. *European Review of Social Psychology*. 1994;5(1):37–68. <https://doi.org/10.1080/14792779543000011>
31. Janis IL. Groupthink. *IEEE Engineering Management Review*. 2008;36(1):36.
32. Stasser G, Stewart D. Discovery of hidden profiles by decision-making groups: solving a problem versus making a judgment.. *Journal of Personality and Social Psychology*. 1992;63(3):426–34. <https://doi.org/10.1037/0022-3514.63.3.426>
33. Schulz-Hardt S, Brodbeck FC, Mojzisch A, Kerschreiter R, Frey D. Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *J Pers Soc Psychol*. 2006;91(6):1080–93. <https://doi.org/10.1037/0022-3514.91.6.1080> PMID: 17144766
34. Wickelgren WA. Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*. 1977;41(1):67–85. [https://doi.org/10.1016/0001-6918\(77\)90012-9](https://doi.org/10.1016/0001-6918(77)90012-9)
35. Postmes T, Spears R, Cihangir S. Quality of decision making and group norms. *J Pers Soc Psychol*. 2001;80(6):918–30. <https://doi.org/10.1037/0022-3514.80.6.918> PMID: 11414374
36. BALES RF, STRODTBECK FL. Phases in group problem-solving. *J Abnorm Soc Psychol*. 1951;46(4):485–95. <https://doi.org/10.1037/h0059886> PMID: 14880365
37. Tuckman BW. Developmental sequence in small groups. *Psychol Bull*. 1965;63:384–99. <https://doi.org/10.1037/h0022100> PMID: 14314073
38. Fisher BA. Decision emergence: phases in group decision-making. *Speech Monographs*. 1970;37(1):53–66. <https://doi.org/10.1080/03637757009375649>
39. Poole MS. Decision development in small groups, III: a multiple sequence model of group decision development1. *Communication Monographs*. 1983;50(4):321–41. <https://doi.org/10.1080/03637758309390173>
40. Poole MS, Roth J. Decision development in small Groups IV a typology of group decision paths. *Human Comm Res*. 1989;15(3):323–56. <https://doi.org/10.1111/j.1468-2958.1989.tb00188.x>
41. Poole MS, Roth J. Decision development in small Groups V test of a contingency model. *Human Comm Res*. 1989;15(4):549–89. <https://doi.org/10.1111/j.1468-2958.1989.tb00199.x>
42. Kruglanski AW, Webster DM. Group members’ reactions to opinion deviates and conformists at varying degrees of proximity to decision deadline and of environmental noise. *J Pers Soc Psychol*. 1991;61(2):212–25. <https://doi.org/10.1037/0022-3514.61.2.212> PMID: 1920063
43. Kruglanski AW, Webster DM. Motivated closing of the mind: “seizing” and “freezing”. *Psychol Rev*. 1996;103(2):263–83. <https://doi.org/10.1037/0033-295x.103.2.263> PMID: 8637961
44. Kruglanski AW, Pierro A, Mannetti L, De Grada E. Groups as epistemic providers: need for closure and the unfolding of group-centrism. *Psychol Rev*. 2006;113(1):84–100. <https://doi.org/10.1037/0033-295X.113.1.84> PMID: 16478302
45. Wenzel M, Woodyatt L. The power and pitfalls of social norms. *Annu Rev Psychol*. 2025;76(1):583–606. <https://doi.org/10.1146/annurev-psych-020124-120310> PMID: 39322435

46. Hausman C, Rapson DS. Regression discontinuity in time: considerations for empirical applications. *Annu Rev Resour Econ*. 2018;10(1):533–52. <https://doi.org/10.1146/annurev-resource-121517-033306>
47. Angelov D. Top2vec: Distributed representations of topics. arXiv preprint 2020. <https://doi.org/10.48550/arXiv.2008.09470>
48. Dumas TM, Maxwell-Smith M, Davis JP, Giulietti PA. Lying or longing for likes? Narcissism, peer belonging, loneliness and normative versus deceptive like-seeking on Instagram in emerging adulthood. *Computers in Human Behavior*. 2017;71:1–10. <https://doi.org/10.1016/j.chb.2017.01.037>
49. Mansbridge JJ. Beyond adversary democracy. University of Chicago Press; 1983.
50. Landemore H, Page SE. Deliberation and disagreement: problem solving, prediction, and positive dissensus. *Politics, Philosophy & Economics*. 2015;14(3):229–54.
51. Esterling KM, Fung A, Lee T. How much disagreement is good for democratic deliberation?. *Political Communication*. 2015;32(4):529–51. <https://doi.org/10.1080/10584609.2014.969466>
52. Karpowitz CF, Mansbridge J. Disagreement and consensus: the importance of dynamic updating in public deliberation. *The deliberative democracy handbook: Strategies for effective civic engagement in the twenty-first century*. 2005. p. 237–53.
53. Stromer-Galley J, Bryant L, Bimber B. Context and medium matter: expressing disagreements online and face-to-face in political deliberations. *Journal of Deliberative Democracy*. 2015;11(1). <https://doi.org/10.16997/jdd.218>
54. Giddens A. Elements of the theory of structuration. *Practicing history: new directions in historical writing after the linguistic turn*. 1984.
55. Seyfarth B. Structuration theory in small group communication: a review and agenda for future research. *Annals of the International Communication Association*. 2000;23(1):341–80. <https://doi.org/10.1080/23808985.2000.11678977>
56. Meyers RA, Seibold DR, Kang P. Examining argument in a naturally occurring jury deliberation. *Small Group Research*. 2010;41(4):452–73. <https://doi.org/10.1177/1046496410366308>
57. Lickel B, Hamilton DL, Wierzchowska G, Lewis A, Sherman SJ, Uhles AN. Varieties of groups and the perception of group entitativity. *J Pers Soc Psychol*. 2000;78(2):223–46. <https://doi.org/10.1037/0022-3514.78.2.223> PMID: 10707331
58. Crawford MT, Sherman SJ, Hamilton DL. Perceived entitativity, stereotype formation, and the interchangeability of group members. *J Pers Soc Psychol*. 2002;83(5):1076–94. <https://doi.org/10.1037/0022-3514.83.5.1076> PMID: 12416913
59. Spencer-Rodgers J, Hamilton DL, Sherman SJ. The central role of entitativity in stereotypes of social categories and task groups. *J Pers Soc Psychol*. 2007;92(3):369–88. <https://doi.org/10.1037/0022-3514.92.3.369> PMID: 17352598
60. Sasaki T, Granovskiy B, Mann RP, Sumpter DJT, Pratt SC. Ant colonies outperform individuals when a sensory discrimination task is difficult but not when it is easy. *Proc Natl Acad Sci U S A*. 2013;110(34):13769–73. <https://doi.org/10.1073/pnas.1304917110> PMID: 23898161
61. Kameda T, Toyokawa W, Tindale RS. Information aggregation and collective intelligence beyond the wisdom of crowds. *Nat Rev Psychol*. 2022;1(6):345–57. <https://doi.org/10.1038/s44159-022-00054-y>
62. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The pushshift reddit dataset. *ICWSM*. 2020;14:830–9. <https://doi.org/10.1609/icwsml.v14i1.7347>
63. Sangiorgio E, Di Marco N, Etta G, Cinelli M, Cerqueti R, Quattrocioni W. Evaluating the effect of viral posts on social media engagement. *Sci Rep*. 2025;15(1):639. <https://doi.org/10.1038/s41598-024-84960-6> PMID: 39753870
64. Riley MW, Riley JW, Toby ML. The Measurement of Consensus. *Social Forces*. 1952;31(2):97–106. <https://doi.org/10.2307/2573391>
65. Blackwell M, Iacus S, King G, Porro G. CEM: coarsened exact matching in stata. *The Stata Journal: Promoting communications on statistics and Stata*. 2009;9(4):524–46. <https://doi.org/10.1177/1536867x0900900402>
66. Iacus SM, King G, Porro G. Causal inference without balance checking: coarsened exact matching. *Polit anal*. 2012;20(1):1–24. <https://doi.org/10.1093/pan/mpr013>
67. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit anal*. 2007;15(3):199–236. <https://doi.org/10.1093/pan/mpj013>
68. Anderson ML. Subways, strikes, and slowdowns: the impacts of public transit on traffic congestion. *American Economic Review*. 2014;104(9):2763–96. <https://doi.org/10.1257/aer.104.9.2763>