Article

new media & society
1–20
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14614448241287510
journals.sagepub.com/home/nms
Sage

Beyond dislike counts: How YouTube users react to the visibility of social cues

Maggie Mengqing Zhang
University of Illinois Urbana-Champaign, USA

Yee Man Margaret Ng
University of Illinois Urbana-Champaign, USA

Abstract

This study investigates the impact of YouTube's 2021 policy, which hides dislike counts and limits a form of negative social feedback. It examines how this change affects social media herding behavior—the tendency of users to align with the majority opinion. We adopted a mixed-method approach, incorporating an online experiment that simulates the YouTube interface and an Interrupted Time Series analysis of real-world user reactions, to assess how the policy affects user engagement. Specifically, we looked at how the absence of one-sided digital cues, combined with content characteristics and individual user predispositions, influences user behavior. Our findings suggest that YouTube's initiative to boost platform positivity had limited success: user responses were more influenced by their ideological leanings than by visible digital cues; Hiding dislikes reduced commenting frequencies and inadvertently increased negative expression. These results highlight the stronger role of ideological beliefs over social cues in shaping engagement, challenging the presumed impact of audience conformity and the negativity bias on social media dynamics.

Keywords

Social cues, YouTube, user engagement, opinion expression, social influence bias, herding behavior

Corresponding author:
Maggie Mengqing Zhang, Institute of Communications Research, University of Illinois Urbana-Champaign, 235C Armory Building, Urbana, IL 61820, USA.
Email: mz44@illinois.edu

Introduction

The concept of "collective intelligence" has gained significant importance in the realm of group decision-making (Bonabeau, 2009). It suggests that by aggregating individual opinions, we can create a collective wisdom that surpasses the understanding of any single person or leader (Landemore, 2017). While the concept seems promising, it relies on individuals forming opinions independently and expressing thoughts freely (Leimeister, 2010). However, in reality, opinion formation and decision-making are often susceptible to the influence of others, leading to "social influence bias" (Muchnik et al., 2013: 647).

In today's interconnected world, primarily facilitated by social networking sites (SNSs), individuals are exhibiting a heightened degree of interdependence (Rainie and Wellman, 2012). Digital platforms offer users a wide range of affordances, such as like/dislike counts, view counts, and more metrics, each serving as different social cues that users can leverage. Nevertheless, in this era of information overload, individuals often rely on majority ratings and pre-existing opinions as cognitive shortcuts to filter information and form judgments (Modgil et al., 2021). Thus, these seemingly lightweight technological affordances challenge the foundational premise of collective intelligence.

The reliance on social cues also opens the possibility for malicious actors to exploit the platform. One notable example is the phenomenon of "dislike mobs" on YouTube, where abusive users weaponize the dislike feature and hijack the platform's technical infrastructure for their benefit, turning YouTube into an intermediary that amplifies racist expressions and harassment (Matamoros-Fernández, 2017).

To counter this, YouTube implemented a new measure on 10 November 2021, hiding the dislike count of videos, to protect creators from "dislike mobs" (YouTube, 2021). However, while this strategy appears to be a tactical move, its actual effectiveness has not been empirically tested. On one hand, hiding the dislike count might elevate the prominence of the like count, potentially reducing disliking behavior if audiences follow herding patterns. On the other hand, when various platform cues coexist, simply removing the dislike numbers might have a limited impact. Especially on YouTube, where the comment section occupies a significant portion of the page, people's expression of opinions may be more influenced by the tone of pre-existing comments rather than the quantified metrics of the dislike count.

Existing research on the impact of dislikes has often overlooked the presence of multiple types of cues and their interactions. This article highlights the need to conduct a thorough investigation into the independent influence of dislikes, likes, and comments, as well as their potential interactions, on user perceptions and behaviors. In this two-fold study, we first (1) leveraged the natural experiment of YouTube's policy change to examine how the removal of negative social cues affected people's engagements using an interrupted time series (ITS) design. We then (2) conducted an online experiment using a YouTube-resembled web page interface to systematically vary the dislike-to-like ratio and the tone of pre-existing comments, aiming to understand the independent and collective effects of likes, dislikes, and comments on user behavior. We also explored how ideological alignment of comments influences this process. Through these two complementary analyses, our study seeks to triangulate results and offer a comprehensive understanding of how social cues impact user behavior.

Literature review

Social cues, platform affordance, and audience conformity

Almost all SNSs, including Facebook, Reddit, and YouTube, provide quantified metrics, such as aggregated like counts, as social cues to indicate user engagement with content. These social cues are usually distinct features that are separate from media content but provide contextual information that shapes and mediates social dynamics.

Sundar's (2008) MAIN (Modality-Agency-Interactivity-Navigability) model emphasizes the role of technological affordances in triggering cognitive heuristics pertaining to users' evaluation. According to the MAIN model, aggregated ratings instigate the "bandwagon heuristic," wherein users align their opinions with the prevailing majority, thinking, "if others think that this is a good story, then I should think so too" (Sundar, 2008: 83).

The bandwagon effect is also related to audience conformity, the tendency of individuals to align their behavior with that of others (Cialdini and Goldstein, 2004). Previous studies have demonstrated that individuals often conform to the actions of their peers, even when these actions may contradict their personal beliefs (Asch, 1956). The concept of herding behavior also illustrates how individuals tend to "observe others and make the same decisions or choices that the others have done" (Sun, 2013: 115). Conformity and herding behavior can be observed in various contexts. For example, consumers are influenced by previous reviews on platforms like TripAdvisor (Michael and Otterbacher, 2014), and the tone of individuals' comments is often shaped by the remarks of others in online discussions (Matook et al., 2022). Additionally, popular search terms (e.g. Twitter Trending Topics) or videos (e.g. YouTube Trending) highlighted on platforms can significantly shape individuals' perceptions of social and personal agendas and influence their motives to take actions (Ng, 2023; Zhang and Ng, 2023).

Online media platforms have transformed audiences from passive receivers to active interpreters of information by introducing interactive elements such as like and dislike buttons, as well as commenting (Boczkowski and Mitchelstein, 2011). Nevertheless, it is crucial to consider the distinction in audience conformity behaviors in relation to these modes of engagement, as they inherently exhibit different levels of cognitive commitment. Specifically, the act of clicking the "like" or "dislike" button represents more straightforward, nonverbal communication that involves less cognitive engagement compared to commenting (Hayes et al., 2018). This process embodies the concept of paralinguistic digital affordances (Hayes et al., 2018) and is characterized as shallow engagement (Ji et al., 2017). In contrast, commenting signifies a strong emotional response with a higher degree of cognitive involvement (Alhabash et al., 2015). Consequently, we treat liking/disliking and commenting as two distinct forms of audience engagement and investigate how different cues influence them.

Dislikes as negative cues

While most of the previous studies often focus on the effects of positive or approval cues (likes/upvotes), negative cues (dislikes/downvotes) remain understudied. Dislikes serve as direct, unfiltered indicators of negative reactions that social media users allocate to content (Lee et al., 2022). The single-click act of disliking is instantaneously reflected in the visible dislike count on the platform's interface. This feature, while ostensibly a tool for expressing disapproval, can be strategically manipulated. For instance, dislikes are often employed in online trolling through coordinated mass downvoting or vote brigading (Cheng et al., 2017). These tactics are particularly concerning as being frequently used to propagate toxicity and target specific groups.

In online interactions, negative cues often exert a stronger influence than positive ones, a phenomenon known as the negativity effect. This effect illustrates the human tendency to prioritize negative information over positive. Previous studies have found the negativity effect in different settings such as the weaker influence of positive reviews compared to negative ones on purchase decisions (Chevalier and Mayzlin, 2006). Highlighted in Baumeister et al.'s (2001) seminal work, *Bad is Stronger than Good*, this bias toward negativity is characterized by a greater impact and more in-depth processing of negative information. Indeed, the positive-negative asymmetry effect has been explored and confirmed by many other research studies (e.g. Chen and Ng, 2017; Peeters and Czapinski, 1990).

There are two possible explanations for the negativity bias. The first is a norm-violating perspective, which suggests that positive cues are more established and prevalent, making negative cues appear non-normative and resulting in the perceived informativeness of positive information being discounted (Hayes et al., 2016). People may perceive negative information to be more diagnostic and revealing of the true aspects of a subject due to the social risk of publishing it (Lee et al., 2022). The second explanation is from a frequency-as-information perspective, which posits that negative information is more informative because it is less common and signals a change from the more frequently experienced positive states (Peeters and Czapinski, 1990). The relative rarity of negative information increases its influence, making it more valuable.

While seemingly influential and informative for users in assessing information on social media platforms, negative cues are not uniformly implemented across platforms. A notable example of this is YouTube's decision to remove the visible count of dislikes. This action, while potentially mitigating targeted negativity, simultaneously deprives users of a crucial evaluative tool—the negative cue. The absence of such cues on a platform raises significant questions about user behavior and content assessment processes. We therefore propose the first research question:

> *Research Question 1 (RQ1).* How does the removal of negative cues influence users' engagement with media content, as measured by changes in the number of likes/dislikes, commenting activity, and the sentiment, subjectivity, and toxicity of comments?

The coexistence of multiple cues

While we have emphasized the significance of negative cues in shaping user engagements, it is crucial to recognize that these negative cues do not exist in isolation. Social media platforms provide diverse social cues that coexist and potentially interplay with each other. Previous studies have explored the influence of multiple concurrent cues on users' perceptions of message persuasiveness. However, most research has focused primarily on positive cues, often overlooking the significant impact of negative feedback on user behavior. Given that social media platforms typically display likes and dislikes together, our research aims to understand the nuanced dynamic and combined effects, specifically examining how the ratio of likes to dislikes influences user engagement.

As previously mentioned, social cues manifest not only as quantifiable metrics (number of likes/dislikes) but also in more nuanced forms, such as the valence of pre-existing comments. This exploration is guided by cue consistency theory (Maheswaran and Chaiken, 1991), complemented by information integration approaches (Anderson, 2013). According to cue consistency theory, multiple sources of information are more informative when they are corroborative rather than contradictory. When cues are consistent, integration typically follows straightforward models, such as linear averaging (Anderson, 2013). However, complexity arises when cues are inconsistent, leading to ambiguity about which cues are more influential. Our research, therefore, focuses on the intertwined effect of numeric cues (likes/dislikes) and content cues (tone of comments). We propose the following research question:

> *Research Question 2 (RQ2).* How do aggregated ratings (the ratio of dislikes to likes) and the valence of pre-existing comments affect user engagement with media content, both independently and in conjunction?

Users' ideological beliefs

Alongside heuristic cues, users' pre-existing ideological beliefs can significantly influence their engagement with media content. Confirmation bias, where individuals seek information that supports their existing beliefs, plays a significant role in this context (Kappes et al., 2019). The cognitive dissonance theory (Festinger, 1957) further supports this notion, suggesting that individuals are more likely to pay attention to pro-attitudinal messages to avoid cognitive dissonance. However, there are also competing arguments suggesting that users tend to engage with opposing viewpoints. For example, Bright et al. (2022) found that users engage with opposing views to undermine their soundness. Regardless, ideological alignment plays a crucial role in user engagement.

Given this understanding, users may be more inclined to dislike media content with opposing views to delegitimize the arguments (Bakshy et al., 2015) and to comment on such content to engage in debates. However, they may also avoid commenting on contradictory content to sidestep confrontation. This leads to the following research question:

> *Research Question 3 (RQ3).* How does ideological alignment between viewers' beliefs and media content impact their engagement?

Methods

This study used a triangulated research design, including an analysis of real-world YouTube data and a 3 × 3 factorial online experiment design to examine the effects of social cues on users' engagement. Analyzing real-world YouTube data offered insights into the effects of removing negative cues (addressing RQ1) by studying user behaviors in a natural setting. On the other hand, the experimental design enabled us to manipulate key features on the interface and establish causal relationships between multiple cues and user engagement, effectively addressing the remaining research questions (RQ2 and RQ3).

Study 1: a natural experiment on YouTube

YouTube's removal of the dislike count feature for videos presents a natural experiment. Although audiences can still dislike videos, the total number of dislikes is no longer visible. Content creators still have backstage access to the number of dislikes; therefore, the impact of the removal of the dislike count is solely on the audience side. Such a change offers an opportunity to study the effect of removing negative cues, answering RQ1.

Data and design

Since real-time data on the dislike number is only accessible to channel owners, we focused on the (a) number of likes as indicators of user reactions and (b) the content of comments to gauge people's opinion expression. We selected media content from news channels that often feature social or political news videos that generate opinion-based comments.

To ensure research validity, we included news outlets with diverse ideological positions. We selected 46 news outlets[1] based on media bias ratings by AllSides (2022) and collected video statistics of these accounts using YouTube's API.

Design A. To explore the effect of dislike removal on users' behavioral reactions (number of likes and comments), we gathered data over a consecutive 90-day period. This timeframe included 30 days before the policy change and 60 days after, yielding 6148 videos from 45 accounts between 10 October 2021 and 10 January 2022. Both like counts and comment counts were treated as features of each video, transforming each day's aggregated data into a distinct time point within ITS analysis. We aimed to determine whether the interruption resulted in any significant difference in the number of likes and comments of videos from these news channels. Design A is illustrated in Figure A1 in the Supplementary Materials.

Design B. In the subsequent design, to test the effect of YouTube's policy change on people's opinion expression, our focus shifted to individual comments beneath each video. These comments were treated as unique time points within the time series. In causal identification terms, the removal of dislike counts serves as an external "shock" or "intervention" in this quasi-experimental design. Specifically, we aimed to determine whether the intervention impacted subsequent comments. We collected all videos posted by these YouTube channels two weeks before YouTube's policy change. This specific timeframe was chosen for two primary reasons: first, having pre-intervention data points is pivotal for the effectiveness of an ITS design; second, commenting behavior exhibits time sensitivity, especially for news-related videos. Setting our starting point too far back could risk having no subsequent comments post-intervention. In total, 1116 videos along with all comments posted within 30 days of the video's creation date were collected using YouTube Data API. We opted for a 30-day period to minimize potential confounding factors that could arise from long-term data collection, as time series data can often be affected by exogenous historical factors (Box-Steffensmeier, 2014). Design B is illustrated in Figure A2 in the Supplementary Materials.

Variables and analysis

For the two designs, we conducted ITS analyses to examine the impact of hiding dislikes (as an interruption) on (1) the number of likes and comments, and (2) comments' sentiment, subjectivity, and toxicity. Following the ITS design principles in Shadish et al. (2002), we only included time series data with at least four observations both before and after the interruption. For Design A, the number of videos stays the same. For Design B, only 308 valid videos with 329,214 comments remain.

The ITS design utilizes pre-intervention observations as the baseline to estimate counterfactual observations for the post-intervention to infer the treatment effect (Ramsay et al., 2003). ITS identifies two distinct intervention outcomes: the level change and the trend change (Ramsay et al., 2003). The level change represents the immediate impact, and the trend change assesses the intervention's gradual effects over time.

Independent variables. To encapsulate these effects, we formulated three temporal variables: (1) T, denoting the number of days since the beginning of our observations; (2) I, a dichotomous variable indicating post-intervention or pre-intervention; and (3) T_2, which measures the days following the intervention. Variable I captures the level change, while T_2 captures the trend change. These two variables were our independent variables.

Dependent variables. In Design A, the daily average like counts and comment counts were the outcomes of interest. In Design B, we focused on the comment content, specifically the average sentiment, subjectivity, and toxicity scores. We measured sentiment[2] and subjectivity[3] using the TextBlob (2019) package, and toxicity[4] using *Perspective API* (Google Jigsaw, 2017).

Model specification. Our unit of analysis was daily-aggregated averages of the number of likes and comments, as well as the toxicity, subjectivity, and sentiment of the comments. These daily aggregates manifested a nested structure of the data. For Design A, the number of comments and likes also depended on the accounts, creating a multi-level effect structure. For Design B, the comments were nested within videos, which were embedded within accounts. Therefore, we employed multi-level mixed-effect models with segmented regressions to account for both video-level and account-level variances.

Control variables. Since audience reactions are influenced by both the foundational attributes of the channel and the qualitative aspects of the video content, we incorporated measurements of toxicity, subjectivity, and sentiment—evaluated on the video's title and description—as video-level predictors. Additionally, the channel's subscriber count served as an account-level control variable. To investigate the potential influence of media accounts' political leaning on user reactions, we also included the media leaning variable in the model. We also tested potential interaction effects between media leaning and the intervention. Media leaning refers to the perceived political orientation or ideological bias of a media outlet (Ribeiro et al., 2018). We categorized each account as left, right, or center using media bias ratings from *Allsides*.

Analysis. Time series data often exhibit inherent biases, including autocorrelation, where observations are related to prior observations or lags, and seasonality, representing recurring patterns within specific intervals. In our ITS design, we employed segmented regressions to test the intervention effect. This method has been utilized in previous research to evaluate the impact of new platform policies (Liang et al., 2022). While segmented regressions do not directly address autocorrelation and seasonality, we conducted time series diagnostics and robustness checks using AutoRegressive Integrated Moving Average (ARIMA) models, which account for autocorrelations, and results are listed in the Supplementary Materials.

To determine whether the effect varied across diverse media accounts, we compared our mixed-effect model (random intercepts) with a fixed-effect ordinary least squares (OLS) regression model. We then conducted analysis of variance (ANOVA) tests to determine the differences between the two models. For Design A, the results indicated that the random intercept models demonstrated a better fit for both the number of likes as DV ($\chi^2 = 2205.4$, $p < .001$) and the number of comments as DV ($\chi^2 = 3049.7$, $p < .001$). The variances in comments (intraclass correlation coefficient [ICC] = .73) and likes (ICC = .78) were predominantly driven by differences at the account level. For Design B, the ANOVA results also showed the multi-level effect model with random intercepts provided a better fit for comments' toxicity ($\chi^2 = 766.08$, $p < .001$), subjectivity ($\chi^2 = 152.61$, $p < .001$), and sentiment ($\chi^2 = 295.83$, $p < .001$).

Results

The effect of removing dislike count on user reaction

Table 1 shows the results from the segmented regression models with mixed effect, with each model predicting the daily average of the number of likes and comments, respectively.

According to the results from Model 1, the removal of the dislike count by YouTube did not immediately affect, nor did it exhibit a gradual influence on, the daily average number of likes. Moreover, media leaning was not associated with any significant effect on the likes. One interesting pattern we observed was that the toxicity of video content had a positive relationship with the daily average number of likes ($b = 0.11$, $p < .001$).

The results in Model 2 show that although there was a small declining trend in the daily number of comments before the intervention, the intervention itself did not introduce any immediate change in the level nor the gradual trend change, indicating that YouTube's removal of dislike counts did not significantly impact people's commenting behavior. Similarly, media leaning did not impact the daily average number of comments either. We also found that the video toxicity positively related to the daily average number of comments ($b = 0.16$, $p < .001$).

Table 1. Linear mixed-effects segment regression models predicting number of likes and comments.

	Model 1 DV: Daily average like count	Model 2 DV: Daily average comment count
Fixed Effects		
Intercept	6.19*** (0.28)	4.98*** (0.28)
Time (T)	0.00 (0.00)	−0.01* (0.00)
Level change (I)	0.07 (0.08)	0.16 (0.09)
Trend change (T2)	0.00 (0.00)	0.01 (0.00)
Video		
Toxicity	0.11*** (0.02)	0.16*** (0.03)
Subjectivity	0.04 (0.02)	0.04 (0.03)
Sentiment	−0.02 (0.03)	−0.05 (0.03)
Channel		
Subscriber Count	1.45*** (0.31)	1.55*** (0.29)
Leaning	−0.09 (0.20)	−0.08 (0.19)
Time × Leaning	0.00 (0.00)	0.00 (0.00)
Level × Leaning	0.04 (0.06)	0.06 (0.07)
Trend × Leaning	0.00 (0.00)	0.00 (0.00)
Variance of random effects		
Level 2	0.87	1.07
Level 1	3.10	2.85
N (Level 1 units)	2295	2295
N (Level 2 units)	45	45
AIC	6497.54	6450.55
BIC	6577.88	6529.80

Note: Coefficients are standardized. AIC = Akaike Information Criterion. BIC = Bayesian Information Criterion.
*$p < .05$, **$p < .01$, ***$p < .001$.

The effect of removing dislike count on opinion expression

Table 2 presents the results from three mixed-effect segmented regression models, each predicting a different dependent variable: toxicity, subjectivity, and sentiment.

In predicting the daily average of comment toxicity (Model 3), while a video's toxicity positively predicted comment toxicity ($b = 0.07$, $p < .001$), our analysis showed no significant immediate impact or gradual effect post-intervention regarding the treatment effect. We did observe an interesting pattern: a negative relationship between media leaning and the daily average of comment toxicity ($b = -0.13$, $p < .05$). This indicated that compared to right-leaning media, left-leaning media triggered more toxic comments. However, we also found that media leaning did not significantly moderate the treatment effect.

For subjectivity (Model 2), only video content subjectivity showed a significant relationship with comment subjectivity ($b = 0.04$, $p < .05$). Channel-level media leaning did not exert a significant influence on comment subjectivity. The treatment effect analysis revealed no significant impact.

Table 2. Linear mixed-effects segment regression models predicting content of comments.

	Model 3 DV: Toxicity	Model 4 DV: Subjectivity	Model 5 DV: Sentiment
Fixed Effects			
Intercept	−0.22*** (0.07)	0.00 (0.05)	0.12 (0.07)
Time (T)	0.00 (0.01)	0.00 (0.01)	−0.01 (0.01)
Level change (I)	0.05 (0.05)	0.07 (0.05)	0.05 (0.05)
Trend change (T2)	0.00 (0.01)	0.00 (0.01)	0.01 (0.01)
Video			
Toxicity	0.07*** (0.02)		
Subjectivity		0.04* (0.01)	
Sentiment			0.02 (0.02)
Channel			
Leaning	−0.13* (0.06)	0.02 (0.04)	0.07 (0.06)
Time × Leaning	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Level × Leaning	0.01 (0.04)	0.05 (0.04)	0.02 (0.04)
Trend × Leaning	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Variance of random effects			
Level 3	0.03	0.01	0.03
Level 2	0.07	0.02	0.03
Level 1	0.87	0.97	0.96
N (Level 1 units)	13661	13661	13661
N (Level 2 units)	308	308	308
N (Level 3 units)	22	22	22
AIC	36526.91	37762.66	37638.15
BIC	36616.89	37852.64	37728.12

*$p < .05$, **$p < .01$, ***$p < .001$.

Regarding sentiment (Model 3), our findings indicated media leaning at the channel level did not significantly influence comment sentiment. The treatment effect analysis showed no significant immediate or gradual impact.

Study 2: online experiment

To answer RQ2 and RQ3, we built a website that resembled the YouTube interface and customized two social feedback cues (shown in Figures F1 and F2 in the Supplementary Materials): (1) dislike counts and (2) pre-existing comment valence. Participants were asked to watch videos on gun control, choose to like or dislike the video, and leave comments in the comment section. We chose the topic of gun control since it is a highly debated and polarizing issue (Wozniak, 2016). News addressing sensitive or controversial topics tends to generate increased engagement (Toepfl and Piwoni, 2017). Therefore, the topic of gun control provides an ideal context for studying how social cues might influence people's reactions. The study was approved by the university's Institutional Review Board.

Participants

We recruited 700 participants with diverse attitudes[5] toward gun control from Amazon Mechanical Turk in January 2023. To ensure data quality, individuals who answered either of the two attention questions incorrectly were excluded from the analysis. The final data set consisted of 606 participants. Their demographic information is shown in the Supplementary Materials.

Experimental design

Stimulus 1: dislike-to-like ratio. In each of the three experimental conditions, we set the like count to 1000. We then varied the number of dislikes to create different scenarios: a dislike-to-like ratio of 1:1 with 1000 dislikes, a ratio of 10:1 with 10,000 dislikes, and a condition where dislikes were hidden. Our primary focus was on examining the impact of removing dislike counts. As such, we concentrated on videos that were presented as either controversial (showing an approximately equal number of likes and dislikes) or highly disapproved (with dislikes outnumbering likes by a factor of ten). We displayed like and dislike counts in units of "Ks" (thousands) to avoid suspicion about the exact numbers, so that participants only saw these counts as 1k or 10k on the platform. Furthermore, preliminary data from Study 1 indicated that the average number of likes for news media videos was a few thousand, further validating that our experimental settings realistically reflected typical user experiences on YouTube.

Stimulus 2: pre-existing comment valence. We collected real-world user comments posted under videos related to gun control. To maintain consistency, we intentionally selected comments of comparable length. In addition, we specifically chose comments with a strong emotional valence to ensure that the sentiment of each comment could be easily discerned without requiring extensive interpretation. In each treatment, we showed four pre-existing comments beneath each video. In an all-supporting comment condition, all four comments agreed with the video. In an all-opposing comment condition, all four comments disagreed. In a mixed condition, there were two supporting and two opposing comments (see Table F1 in the Supplementary Materials).

Procedures. After consenting to participate in the study, participants were asked to indicate their pre-existing attitudes about gun control. Participants were then randomly assigned to one of the nine experiment conditions. In each condition, participants were instructed to view two videos that presented conflicting perspectives on the topic of gun control: one video, sourced from MSNBC,[6] advocated for tighter gun control, while the other, obtained from *Fox News*,[7] opposed tighter policy. We chose these two videos of similar duration, format, and presentation style to ensure no influence from extraneous factors.[8] Furthermore, to minimize potential biases and ensure a fair comparison, the order in which the two videos were presented was randomized. Each video has the same settings for the two stimuli.

After viewing each video, participants were asked to express their approval/disapproval by clicking either the like or dislike button and had the option to comment. Given the autonomy to choose if and how to interact with the content, participants could engage according to their natural inclinations and preferences, mirroring a realistic social media environment. They were not required to like, dislike, or comment, allowing for an authentic representation of user behavior. Their interactions with the webpage, including clicks and comments, were recorded in the backend database. The webpage was seamlessly integrated into Qualtrics as a full-screen display for an immersive experience. Subsequently, participants were directed to complete a post-exposure questionnaire within Qualtrics.

Dependent variables. The key dependent variables were (1) participants' actions—clicking dislike/like buttons and posting comments. For users who left comments on YouTube videos, this study analyzed (2) the sentiment, subjectivity, and toxicity of their comments using the same measurements in Study 1. Web-log data were collected to track the behaviors of each participant and the content of their comments. Figure E1 in Supplementary Materials illustrates the overall research design.

Independent variables

Ideological alignment. Ideological alignment refers to the consistency between a participant's pre-exposure attitude toward gun control and the opinions expressed in the video they watched. To assess participants' pre-exposure attitudes toward gun control, we used a seven-point scale ranging from "1 = strongly oppose" to "7 = strongly favor." This measurement was based on three statements: (1) assault-style weapons should be allowed; (2) people should be allowed to carry concealed guns without a permit; and (3) waiting periods for people who want to buy guns legally should be shortened. If a participant's average response to these three statements was above four, they were labeled as gun supporters, and vice versa. An ideological alignment score of 1 was assigned when a participant's stance aligned with the video's perspective, and 0 when it did not. Since we used two videos that presented contrasting viewpoints on gun control, each participant was exposed to one condition featuring ideologically aligned content and another with ideologically misaligned content. This approach allowed us to examine participant responses to both congruent and divergent ideological perspectives.

Analytical approaches. We conducted analysis using chi-square tests, *t*-tests, and ANOVAs to analyze group differences. Subsequently, we used multinomial logistic regressions to investigate how social cues and ideological alignment impacted the participants' clicks, and logistic regressions to explore the effect on people's commenting behavior. Finally, we used multiple linear regressions to investigate the effect on the sentiment, subjectivity, and toxicity of the comments.

Results

The effect of social cues on user behavior

To answer RQ2, we conducted Chi-square tests of independence to examine the effect of the dislike-to-like ratio on clicking and commenting behavior, as well as the effect of pre-existing comments on clicking likes or dislikes. The results, shown in Table G1 in the Supplementary Materials, indicate that neither the dislike-to-like ratio nor the valence of pre-existing comments had a significant effect on clicks. We used multinomial logistic regressions (Table 3) and linear regressions (Table 4) to examine the relationship between aggregated ratings and pre-existing comments, and their interaction effects. Results in Table 3 showed no significant interaction effect between the dislike-to-like ratio and pre-existing comments on clicking behavior.

The effect of social cues on opinion expression

To answer the RQ2, Chi-square tests indicated a significant influence of the dislike-to-like ratio on commenting behavior, $\chi^2(2) = 8.53$, $p < .05$. Further analysis revealed that individuals commented significantly less when the dislike count was hidden compared to when the dislike-to-like ratio was 1:1. We also tested the relationship between dislike-to-like ratio and the sentiment, subjectivity, and toxicity of the comments. Results in Table 4 show that the ratio did not impact the sentiment or subjectivity. However, compared to the condition where the dislike count was hidden, when the dislike-to-like ratio was 1:1, the toxicity of comments was significantly lower ($b = -0.05$, $p < .01$). This finding indicates that when a video's content is controversial in nature (receiving an equal number of likes and dislikes), hiding the dislikes leads to more toxic comments.

For the influence of pre-existing comments on opinion expression, we found that the valence of pre-existing comments had a significant impact on commenting, $\chi^2(2) = 44.05$, $p < .001$. Post hoc analysis indicated the difference was primarily driven by the group exposed to opposing comments. ANOVAs indicated that there was a significant effect of pre-existing comments on toxicity, $F(2, 916) = 3.12, p < .05$. Tukey's Honestly Significant Difference (HSD) test indicated that people were more likely to post toxic comments when exposed to disapproving comments compared to pre-existing supportive comments.

In terms of the interaction effect of dislike-to-like ratio and pre-existing comments on opinion expression, logistic regression analyses in Table 4 indicated a significant interaction effect between these two types of social cues on commenting behavior. The interaction effect is visualized in Figure G1 in the Supplementary Materials. After incorporating the interaction terms into the model, the main effects of the dislike-to-like ratio and pre-existing comments on commenting behavior became non-significant, suggesting that the influence of the dislike-to-like ratio on commenting behavior was dependent on the valence of pre-existing comments. Specifically, when there were pre-existing opposing comments, the effect of the dislike-to-like ratio on commenting behavior was significantly higher compared to other conditions. However, among participants who commented, there was no significant interaction effect on the comments' toxicity, sentiment, or subjectivity (Table 4).

The effect of ideological alignment

To answer RQ3, Chi-square tests and *t*-tests were conducted. Results in Table G1 in the Supplementary Materials showed that ideological alignment significantly influenced

Table 3. Multinomial logistic regressions predicting clicking behavior.

	DV: click (reference category: no click)							
	Clicking dislike button ($n = 249$) (1)				Clicking like button ($n = 424$) (2)			
	B	(SE)	OR	95% CI	B	(SE)	OR	95% CI
Intercept (constant)	−1.41***	0.26	0.25	(0.15, 0.41)	−0.14	0.21	0.87	(0.58, 1.31)
Main Effect								
Dislike/like ratio (reference: dislike number hidden)								
1 (1k dislikes vs 1k likes)	−0.07	0.33	0.94	(0.49, 1.79)	0.28	0.27	1.33	(0.78, 2.27)
10 (10k dislikes vs 1k likes)	−0.04	0.34	0.96	(0.49, 1.86)	0.13	0.29	1.14	(0.65, 2.00)
Display comments (reference: mixed opinion)								
Support	−0.11	0.35	0.89	(0.45, 1.75)	0.13	0.29	1.14	(0.61, 1.80)
Oppose	−0.02	0.32	0.98	(0.52, 1.84)	0.04	0.28	1.05	(0.65, 2.00)
Ideological alignment (reference: aligned)								
Nonalignment	0.91***	0.17	2.48	(1.79, 3.43)	−0.62***	0.13	0.54	(0.41, 0.70)
Interaction Effect								
Ratio 1 * Oppose	0.38	0.47	1.47	(0.59, 3.66)	0.28	0.39	1.32	(0.62, 2.82)
Ratio 10 * Oppose	0.65	0.46	1.92	(0.78, 4.70)	0.08	0.40	1.08	(0.50, 2.35)
Ratio 1 * Support	−0.21	0.49	0.81	(0.31, 2.10)	−0.47	0.39	0.63	(0.29, 1.35)
Ratio 10 * Support	0.46	0.48	1.58	(0.61, 4.05)	−0.14	0.41	0.87	(0.39, 1.94)
Nagelkerke R^2	0.10							

*$p < .05$ **$p < .01$ ***$p < .001$.

The size n indicates the number of observations ($n = 1212$) instead of participants as each participant responded to two different videos.

B regression coefficient, *SE* standard error, *OR* odd ratio, *CI* confidence interval for *OR*.

Table 4. Logistic regression and linear regressions predicting comments.

	Dependent variable: comment				
	Whether comment (*n* = 1212) logistic (1)		Sentiment (*n* = 919) OLS (2)	Subjectivity (*n* = 919) OLS (3)	Toxicity (*n* = 919) OLS (4)
	B (SE)	OR (95% CI)	B (SE)	B (SE)	B (SE)
Intercept	0.85*** (0.20)	2.33 (1.58, 3.50)	0.13*** (0.03)	0.45*** (0.03)	0.15*** (0.02)
Main effect					
Dislike-to-like ratio (reference: dislike number hidden)					
1	−0.12 (0.26)	0.89 (0.53, 1.47)	0.07 (0.04)	−0.03 (0.04)	−0.05** (0.03)
10	0.29 (0.28)	1.33 (0.77, 2.32)	0.06 (0.04)	0.02 (0.05)	−0.00 (0.03)
Display comments (reference: mixed opinion)					
Support	0.29 (0.28)	1.33 (0.77, 2.33)	0.07 (0.04)	−0.08* (0.05)	−0.05* (0.03)
Oppose	0.07 (0.26)	1.08 (0.64, 1.81)	0.05 (0.04)	−0.01 (0.04)	−0.00 (0.03)
Ideological alignment (reference: aligned)					
Nonalignment	−0.03 (0.14)	0.97 (0.74, 1.28)	−0.07 *** (0.02)	0.02 (0.02)	0.04*** (0.01)
Interaction Effect					
Ratio 1 * Oppose	2.11 *** (0.50)	8.25 (3.23, 23.45)	−0.07 (0.06)	0.05 (0.06)	0.04 (0.04)
Ratio 10 * Oppose	1.93 *** (0.54)	6.87 (2.53, 21.12)	−0.04 (0.06)	−0.07 (0.06)	−0.01 (0.04)
Ratio 1 * Support	−0.44 (0.38)	0.65 (0.31, 1.35)	−0.05 (0.06)	0.10 (0.06)	0.05 (0.04)
Ratio 10 * Support	−0.61 (0.40)	0.55 (0.25, 1.19)	−0.08 (0.06)	−0.02 (0.06)	0.01 (0.04)
R^2/AIC	AIC: 1260.03		$R^2 = .01$	$R^2 = .01$	$R^2 = .02$

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$.

individuals' clicking behavior, $\chi^2(2) = 81.90$, $p < .001$. Specifically, participants tended to like the content when exposed to videos that aligned with their ideology. However, ideological alignment did not significantly impact whether people commented or not. Among those who commented, *t*-tests revealed that ideological alignment significantly improved the sentiment, $t(908.95) = 3.75$, $p < .001$, and reduced the toxicity, $t(905.67) = -3.13$, $p < .001$, of comments (Table 3).

Discussion

External influences of social cues play a significant role in shaping opinion formation, often leading to phenomena like "herding behavior" (Muchnik et al., 2013; Zhang and Ng, 2023). Our study focused on YouTube's policy of removing dislike counts and assessed its impacts on user engagement. While YouTube's intention was to protect content creators from dislike mobs, our findings cast doubt on its effectiveness.

Our observational study on YouTube revealed that the intervention had no significant immediate or gradual impact on the number of likes or comments. A detailed analysis on comment content also showed no significant effect on sentiment, subjectivity, or toxicity across various media leanings. This finding suggests that the removal of dislike counts may not have achieved the intended effect of improving the positivity, at least in terms of observable engagement metrics. This outcome challenges the assumptions of previous theories, which posit that social cues, such as aggregated ratings, can trigger the "bandwagon heuristic" and influence users' actions (Muchnik et al., 2013). Furthermore, it questions the applicability of the negativity bias effect in this context (Rozin and Royzman, 2001). If "bad is stronger than good," as the theory suggests, removing dislike counts should have nudged audiences in a more positive direction. However, our observational study results did not support this expectation.

In our online experiment, we found that neither the dislike count nor pre-existing comments significantly influenced the decision to like or dislike a video. Instead, this decision was solely driven by the ideological alignment between individuals' beliefs and the video content. Participants were more inclined to upvote and less likely to downvote when exposed to videos that aligned with their ideology. This finding aligns with the theories of confirmation bias (Kappes et al., 2019) and cognitive dissonance (Festinger, 1957), which emphasize the role of individuals' pre-existing beliefs in shaping their engagement with media content. These results suggest that the impact of social cues, such as audience conformity (Cialdini and Goldstein, 2004) and the negativity bias effect (Baumeister et al., 2001), may be less prominent than initially anticipated when ideological alignment is taken into account.

Our findings on comments underscore the importance of the coexistence of multiple cues and their collective influence on user engagement. Cue consistency theory (Maheswaran and Chaiken, 1991) and information integration approaches (Anderson, 2013) suggest that the consistency between different cues plays a crucial role in determining users' behavior. We found that users were more likely to engage, as evidenced by commenting, when cues were aligned—for instance, when a high dislike-to-like ratio was complemented by opposing comments. In contrast, inconsistent cues—such as hiding dislikes while showing opposing comments—resulted in reduced commenting activity. This finding supports the idea that cue incongruity can introduce uncertainty and thereby affect user behavior (Maheswaran and Chaiken, 1991).

Additionally, users' comments tended to be more toxic when the dislike count was hidden, compared to scenarios where the dislike-to-like ratio was balanced (1:1). This suggests that for media content perceived as controversial (more likely to garner similar numbers of dislikes and likes), removing the dislike count can foster a more toxic environment for the expression of opinions. One possible explanation is that when individuals can't gauge the extent of disapproval toward media content through dislike counts, they may feel compelled to express discontent through comments. Despite the ability to dislike content, the inability to assess communal sentiment may lead to feelings of reduced agency or decreased solidarity with like-minded others.

While our study offers valuable insights into the effect of social cues on user engagement, there are several limitations. First, this study focused on news-related content within YouTube in the United States, limiting generalizability to other content types, platforms, and global demographics. Second, participants in our second study might have been aware that they were being observed, which may have influenced their behavior and responses. Third, although the study's focus was primarily on the audience side, YouTube's policy change might also affect content creators or the platform more broadly—areas the study did not explore. Future research could address these aspects to provide a more comprehensive understanding of the policy change's impact.

Our study is among the first to empirically examine YouTube's policy change, offering a comprehensive view of how various social cues interact and affect user engagement. The implications of our findings are multifaceted: for platforms like YouTube, simply removing dislike counts may not adequately promote positivity or reduce toxicity. Instead, user behavior is driven by deeper factors such as ideological alignment and individual predispositions. Digital platforms must consider the complex interplay between platform-provided social cues and user predispositions when creating policies to enhance online discourse. Future research should investigate the reasons behind the limited influence of aggregated like and dislike numbers, which could provide valuable insights for refining online rating systems.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Maggie Mengqing Zhang https://orcid.org/0000-0001-6771-3820
Yee Man Margaret Ng https://orcid.org/0000-0001-5043-9159

Supplemental material

Supplemental material for this article is available online.

Notes

1. The complete list of these media accounts is included in Table B1 of the Supplementary Materials.
2. Sentiment results were on a scale of −1 (negative sentiment) to 1 (positive sentiment).
3. Subjectivity results were on a scale of 0 (objective information) to 1 (subjective opinion).
4. Toxicity results were on a scale of 0 to 1 (highly toxic).
5. The final data set contains 362 pro-gun-control participants and 244 anti-gun-control participants.
6. Video: Matthews: We Need to Stand Up for Gun Control Hardball MSNBC by MSNBC. Link: https://www.youtube.com/watch?v=kl-phE8KWoo
7. Video: Common arguments for gun control, shot down by Fox News. Link: https://www.youtube.com/watch?v=_RDu0YDeqNk
8. See Table F2 in the Supplementary Materials for detailed information of the videos used in the experiment.

References

Alhabash S, Baek J, Cunningham C, et al. (2015) To comment or not to comment? How virality, arousal level, and commenting behavior on YouTube videos affect civic behavioral intentions. *Computers in Human Behavior* 51: 520–531.

AllSides (2022) *Allsides Media Bias Ratings*. AllSides. Available at: https://www.allsides.com/media-bias/ratings

Anderson NH (2013) Unified psychology based on three laws of information integration. *Review of General Psychology* 17(2): 125–132.

Asch SE (1956) Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological Monographs: General and Applied* 70(9): 1–70.

Bakshy E, Messing S and Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science* 348: 6239.

Baumeister RF, Bratslavsky E, Finkenauer C, et al. (2001) Bad is stronger than good. *Review of General Psychology* 5(4): 323–370.

Boczkowski PJ and Mitchelstein E (2011) How users take advantage of different forms of interactivity on online news sites: clicking, E-Mailing, and commenting. *Human Communication Research* 38(1): 1–22.

Bonabeau E (2009) Decisions 2.0: the power of collective intelligence. *MIT Sloan Management Review* 50(2): 45.

Box-Steffensmeier JM (2014) *Time Series Analysis for the Social Sciences*. Cambridge: Cambridge University Press.

Bright J, Marchal N, Ganesh B, et al. (2022) How do individuals in a radical echo chamber react to opposing views? Evidence from a content analysis of Stormfront. *Human Communication Research* 48(1): 116–145.

Chen GM and Ng YMM (2017) Nasty online comments anger you more than me, but nice ones make me as happy as you. *Computers in Human Behavior* 71: 181–188.

Cheng J, Bernstein M, Danescu-Niculescu-Mizil C, et al. (2017) Anyone can become a troll: causes of trolling behavior in online discussions. In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, Portland, OR, 25 February–1 March, pp. 1217–1230.

Chevalier JA and Mayzlin D (2006) The effect of word of mouth on sales: online book reviews. *Journal of Marketing Research* 43(3): 345–354.

Cialdini RB and Goldstein NJ (2004) Social influence: compliance and conformity. *Annual Review of Psychology* 55(1): 591–621.

Festinger L (1957) *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Google Jigsaw (2017) *Perspective Developers*. Perspectiveapi.com. Available at: https://developers.perspectiveapi.com/s/about-the-api

Hayes RA, Carr CT and Wohn DY (2016) One click, many meanings: interpreting paralinguistic digital affordances in social media. *Journal of Broadcasting & Electronic Media* 60(1): 171–187.

Hayes RA, Wesselmann ED and Carr CT (2018) When nobody "Likes" you: perceived ostracism through paralinguistic digital affordances within social media. *Social Media + Society* 4(3): 205630511880030.

Ji YG, Li C, North M, et al. (2017) Staking reputation on stakeholders: how does stakeholders' Facebook engagement help or ruin a company's reputation? *Public Relations Review* 43(1): 201–210.

Kappes A, Harvey AH, Lohrenz T, et al. (2019) Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience* 23(1): 130–137.

Landemore H (2017) *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton, NJ: Oxford Princeton University Press.

Lee SM, Thomer AK and Lampe C (2022) The use of negative interface cues to change perceptions of online retributive harassment. *Proceedings of the ACM on Human-computer Interaction 6(CSCW2):* 334.

Leimeister JM (2010) Collective intelligence. *Business & Information Systems Engineering* 2: 245–248.

Liang F, Zhu Q and Li GM (2022) The effects of flagging propaganda sources on news sharing: quasi-experimental evidence from Twitter. *The International Journal of Press/Politics* 28(4): 909–928.

Maheswaran D and Chaiken S (1991) Promoting systematic processing in low-motivation settings: effect of incongruent information on processing and judgment. *Journal of Personality and Social Psychology* 61(1): 13.

Matamoros-Fernández A (2017) Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society* 20(6): 930–946.

Matook S, Dennis AR and Wang YM (2022) User comments in social media firestorms: a mixed-method study of purpose, tone, and motivation. *Journal of Management Information Systems* 39(3): 673–705.

Michael L and Otterbacher J (2014) Write like I write: herding in the language of online reviews. *Proceedings of the International AAAI Conference on Web and Social Media* 8(1): 356–365.

Modgil S, Singh RK, Gupta S, et al. (2021) A confirmation bias view on social media induced polarisation during Covid-19. *Information Systems Frontiers* 26: 417–441.

Muchnik L, Aral S and Taylor SJ (2013) Social influence bias: a randomized experiment. *Science* 341(6146): 647–651.

Ng YMM (2023) A cross-national study of fear appeal messages in YouTube trending videos about COVID-19. *American Behavioral Scientist*. Epub ahead of print Feb 21. DOI: 10.1177/00027642231155363

Peeters G and Czapinski J (1990) Positive-negative asymmetry in evaluations: the distinction between affective and informational negativity effects. *European Review of Social Psychology* 1(1): 33–60.

Rainie H and Wellman B (2012) *Networked: the New Social Operating System*. Cambridge, MA: MIT Press.

Ramsay CR, Matowe L, Grilli R, et al. (2003) Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. *International Journal of Technology Assessment in Health Care* 19: 613–623.

Ribeiro F, Henrique L, Benevenuto F, et al. (2018) Media bias monitor: quantifying biases of social media news outlets at large-scale. In: *Proceedings of the international AAAI conference on web and social media*, Buffalo, NY, 3–6 June.

Rozin P and Royzman EB (2001) Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review* 5(4): 296–320.

Shadish WR, Cook TD and Campbell DT (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Wadsworth, OH: Cengage Learning.

Sun H (2013) A longitudinal study of herd behavior in the adoption and continued use of technology. *MIS Quarterly* 37(4): 1013–1041.

Sundar SS (2008) *The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility* (pp. 73-100). Cambridge, MA: MacArthur Foundation Digital Media and Learning Initiative.

Textblob (2019) *Textblob: Simplified Text Processing — Textblob 0.16.0 Documentation*. Textblob.readthedocs.io. Available at: https://textblob.readthedocs.io/en/dev

Toepfl F and Piwoni E (2018) Targeting dominant publics: how counterpublic commenters align their efforts with mainstream news. *New Media & Society* 20: 5.

Wozniak KH (2016) Public opinion about gun control post–Sandy Hook. *Criminal Justice Policy Review* 28(3): 255–278.

YouTube (2021) An Update to Dislikes on Youtube. Blog.youtube; Youtube Official Blog, 10 November. Available at: https://blog.youtube/news-and-events/update-to-youtube/

Zhang MM and Ng YMM (2023) #TrendingNow: How Twitter Trends impact social and personal agendas? *International Journal of Communication* 17: 2048–2067.

Author biography

Maggie Mengqing Zhang is a Ph.D student in the Institute of Communications Research at the University of Illinois at Urbana-Champaign's College of Media. She is interested in studying generative artificial intelligence, social media and human computer interaction.

Yee Man Margaret Ng (Ph.D., Journalism, University of Texas at Austin) is an Associate Professor in the Department of Journalism at the University of Illinois at Urbana-Champaign's College of Media, with an affiliate appointment in the Department of Computer Science. Her research examines technology use, social media, and platform migration using computational methods.