Online Journal of Communication and Media Technologies, 2025, **15**(4), e202542
e-ISSN: 1986-3497

OPEN ACCESS

Research Article

Users' perception of media content and content moderation: Exploring antecedents of reporting harmful comments from a dual perspective

Yaoying Zhu [1]
0000-0003-2464-7071

Zhuo Song [2*]
0000-0002-2323-0023

[1] Teaching Center for Writing and Communication, School of Humanities, Tsinghua University, Beijing, CHINA
[2] School of Journalism and Communication, Nanjing Normal University, Nanjing, CHINA
[*] Corresponding author: pkuviola@163.com

Citation: Zhu, Y., & Song, Z. (2025). Users' perception of media content and content moderation: Exploring antecedents of reporting harmful comments from a dual perspective. *Online Journal of Communication and Media Technologies, 15*(4), e202542. https://doi.org/10.30935/ojcmt/17619

ARTICLE INFO

Received: 11 Jul 2025
Accepted: 19 Nov 2025

ABSTRACT

Extensive participation by users is essential for the effectiveness of content moderation. Thus, it is pivotal to understand what factors influence users' acceptance of reporting harmful comments to the social media platform. On the basis of existing literature on the third-person effect and human-machine interaction, in the current study, we explored the antecedents of reporting harmful comments to the platform in terms of perceptions surrounding media content and content moderation from a dual "content-moderation" perspective. Through a survey of Weibo users in China (N = 500), we examined how perceived media effects, perceived human agency, and perceived justice of the reporting mechanism influence behavioral responses. The results revealed that perceived adverse media effects on others, perceived fairness and perceived transparency increased users' engagement in content moderation. Moreover, the findings indicated that perceived human agency attenuated the relationship between perceived adverse media effects on others and reporting behavior. These insights contribute to the burgeoning field of research exploring how users perceive and interact with sociotechnical systems in the domain of user reporting. This study also innovatively integrates perceptions related to content and moderation, gaining more comprehensive understandings of reporting behavior. The current findings have practical implications for platform operators seeking to develop moderation tools for constructive discourse.

Keywords: content moderation, report, media effects, perceived justice, human-AI collaboration, "content-moderation" perspective

INTRODUCTION

Social media has developed into a vibrant platform for voicing opinions about public issues, serving as an arena of digital expression (Dahlberg, 2011). As social media platforms have progressed, managing harmful online communication, including hate speech, vulgarity, racism, and other antisocial communication practices has garnered significant attention. To curb the detrimental consequences of online harmful communication, social media platforms have implemented various means of moderation to improve governance, such as the use of "flag." Flagging is a sociotechnical apparatus by which users can report content that violates the norms of the platform, then the platform decides whether to block or delete it (Crawford & Gillespie, 2016). In China, "flag" is also referred to as "complaint" or "report". In addition to diligent self-regulation by the platform (Crawford & Gillespie, 2016), "flag" in China is compulsory under governmental provisions, and has been widely adopted (Xie et al., 2023).

Previous studies have made important initial attempts to investigate the predictors of reporting behaviors (Kalch & Naab, 2017; Kunst et al., 2021; Wilhelm et al., 2019; Xie et al., 2023), emphasizing factors related to content, moderation, context and individual characteristics (Meerson et al., 2025). However, existing studies generally adopted a single perspective and have neglected to consider that users operate within a set of sociotechnical assemblages afforded by the platform (Gillespie, 2018; van Dijck, 2018). Sociotechnical design decisions about how to perform moderation and users' perceptions about moderation may influence how they interpret, engage with and interact with the platform's processes (Bhandari et al., 2021; Helberger et al., 2018; Jhaver et al., 2019a). Thus, a more detailed and richer understanding of user reporting behavior is currently lacking.

Informed by the third-person effect (TPE) and human-machine interaction literature, the current study was designed to address by adopting a dual "content-moderation" perspective, exploring the antecedents of reporting harmful comments to the platform (RTP) by examining users' perceptions regarding media content and content moderation. The current study contributes in two significant ways. First, we position user reporting in the field of human-machine interaction. Reporting mechanisms, as part of content moderation, involves "visual interfaces, sociotechnical computational systems and communication practices" (Myers West, 2018, p. 4369). The ways in which users "interface" with reporting processes, which involves the setting of affordances or disaffordances, might have an impact on their behavior (Cover et al., 2025). The current study expands the existing literature by illustrating how users' attitudes toward sociotechnical systems shape their usage by applying agency locus and organizational justice. Agency locus touches on the critical concern of the tension between human and artificial intelligence (AI) agency in emerging human-AI collaborative moderation (Sundar, 2020), whereas organizational justice scrutinizes the reporting mechanism from the perspective of organizational decision-making in an unpredictable digital environment. Second, we refine media effects research by integrating it with critical nuances of human-AI interaction, enhancing its relevance for understanding user behavior in the AI era. Our research reveals the interaction effects of perceived media effects on others (PME3) and perceived human agency on adjusting users' behaviors. In previous studies, these factors have typically been considered as two separate characteristics accounting for bystander intervention against disruptive online behavior (Meerson et al., 2025). The current findings provide more nuanced and dynamic insights into the reporting mechanism, elucidating the coordination between the platform and users within complex power dynamics.

LITERATURE REVIEW

Flagging as a Method of Content Moderation

Moderation can be defined as "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" (Grimmelmann, 2015, p. 47). In content moderation, platform operators take on roles as creators of norms and rule enforcers (Gillespie, 2018; Schwarz, 2019). Social media platforms have developed intricate, complex and multi-layered content-moderation systems involving human-machine collaboration to process the sheer number of online content (Myers West, 2018). Most platforms employ a combination of two approaches in practice (Stockinger et al., 2025), rendering "automation-versus-human" a crucial issue in AI-powered moderation (Zhao & Zhang, 2024).

Individuals can contribute to moderation by engaging in coping behaviors, such as flagging. Flagging denotes a mechanism for users to report offensive content to a social media platform within the predetermined rubrics of a platform's community guidelines and then wait for a decision regarding the "acceptability" of the flagged behavior (Crawford & Gillespie, 2016). Flagging constitutes a collaboration between decentralized users, platforms, humans and algorithms, as well as social structures (Crawford & Gillespie, 2016; Young, 2022). Reporting directly to platforms is considered an effective approach for addressing ubiquitous problematic content across academic research and public discussions (Šori & Vehovar, 2022).

Because the extensive participation of users underlies the effectiveness of content moderation, it is important to identify the disparities between moderation practice and public understandings. A bystander perspective has been widely adopted that construes flagging as bystander intervention in online disruptive

behavior which is targeted on others (Aljasir, 2023; Bhandari et al., 2021; Obermaier, 2024). These studies have tended to focus on bystanders' psychological processes and features (Leonhard et al., 2018; Schmid et al., 2024; Ziegele et al., 2019). Porten-Cheé et al. (2020) considered flagging as a low-threshold online civic intervention to voluntarily restore public discourse and developed an explanatory model encompassing content, individual and contextual factors. Meerson et al. (2025) developed a research framework for elucidating bystander intervention in toxic communication by emphasizing content, moderation, context and individual characteristics. The current study seeks to complement this prior research by incorporating factors related to media content and the sociotechnical systems that configure the platform's moderation.

TPE, Influence of Presumed Influence, and PME3

Examining flagging as a reactive remedy after exposure to harmful media content may shed light on users' perceptions of such effects on specific groups. The TPE hypothesis was first proposed by Davison (1983), who hypothesized that individuals perceive others as more susceptible to media influence than themselves, often leading to behaviors based on this cognitive disparity. The TPE consists of two dimensions: the perceptual component (as known as third-person perception) and the behavioral component. The latter component suggests that this self-other perceptual gap can motivate individuals to take remedial actions, particularly when they perceive significant negative media effects (Davison, 1983). Drawing on the TPE, the influence of presumed influence (IPI) denotes a process in which "people perceive some influence of a message on others and then react to that perception of influence" (Gunther & Storey, 2003). Thus, the PME3 may inspire attitudinal and behavioral responses.

In terms of upholding censorship or moderation, the PME3 is reported to have stronger predictive power than the other-self gap in the perceived media effects (Chung & Moon, 2016; Chung & Wihbey, 2024; Jhaver & Zhang, 2025). Wang and Kim (2020) reported that exposure to uncivil comments may elevate presumed flaming intention of others, which consequently promotes individuals' inclination to report uncivil comments. On the basis of this line of research, the current study proposed that the linkage between PME3 and corrective behaviors still exists in the context of social media platforms. Users are entitled to be crucial participants by executing informal social control (Watson et al., 2019) on platforms. When individuals are confronted with harmful comments, the more intense influence on others they evaluate, they are more motivated to actively moderate and restrain harmful discourse in a collaborative manner. Building on prior studies, we proposed that:

H1: PME3 is positively related to willingness to report harmful comments to the platform.

Perceptions of the Reporting Mechanism

Scholars have initiated explorations into the complexities of content moderation and criticized it for its legitimacy recently. An increasing body of research has investigated the normative frameworks for supervising the ways in which platforms exert their suppressive power. Each of these frameworks relates to a set of salient public values, ranging from human rights-based approaches (Common, 2019; Dias Oliva, 2020) and constitutional principles represented by transparency and justice (Leerssen, 2023; Suzor, 2018), to fairness, accountability, and transparency (FAT) model which has been applied widely to audit algorithmic systems (Gorwa, 2018; Lepri et al., 2018). This line of research seeks to outline the deviation between the status quo and the ideal way in which platforms comply with their responsibilities.

Another area of research applying a user-centric perspective under the field of human-machine interaction has emerged. Content moderation relates to a range of punitive enforcement measures, including the removal, filtering and reduction (Gillespie, 2022; Goanta & Ortolani, 2022). The perceptions of the users who experience the execution of the "power" may impact the legitimacy of the subsequent operation of power (Meerson et al., 2025). Some researchers have examined how users perceive, assess, and deal with content moderation, exploring folk theories that have been developed to make sense of interactions with sociotechnical systems (Jhaver et al., 2019a; Lyu et al., 2024; Vaccaro et al., 2020). In the current study, we seek to extend this line of research by introducing agency locus and perceived justice as approaches for academic assessment of users' perceptions regarding reporting mechanisms.

Agency locus

There has long been controversy regarding whether machines have agency because of their lack of intentionality, which was traditionally considered to be the key criterion for defining agency (Dennett, 1988). However, during human-machine interaction, users may perceive AI systems as thinking and acting in a way that is analogous to human behavior to some degree and may perceive these systems as exhibiting some level of "apparent" agency (Liu, 2021). Given that modern AI systems increasingly rely on automated machine learning, which processes mass data to define or refine decision-making rules, some systems may generate algorithms and rules autonomously, independent of human control. Drawing on the concept of the "locus of control," which refers to a person's cause attribution (Duttweiler, 1984; Rotter, 1966), Liu (2021) introduced the concept of "agency locus" to distinguish between two types of AI systems: one reflects human agency, which refers to human-made rules, and the other reflects machine agency, which refers to machine-generated rules.

In contrast to this rather definite dichotomy, most social media platforms implement content moderation through human-AI collaboration. This is a common strategy that leverages the efficiency of AI to process data in bulk, while relying on human subjective judgment to address sensitive issues. Users may feel confused about whether content moderation decisions are made by humans or AI (Meerson et al., 2025), and the agency is shared between humans and AI.

Molina and Sundar (2022) employed a user-centric approach to develop this concept, assessing the degree to which a user attributes the rule-maker of an AI system to a human or a machine, and measuring it continuously via two variables: perceived AI agency and perceived human agency. Inspired by this research, agency locus indicates the degree to which a user thinks the decision-making system reflects AI or human autonomy. A tightly coupled collaboration between humans and machines may create a tension between human agency and AI agency. Although some users may embrace the efficiency of machines in performing tasks, many are worried that machine agency may overwhelm human agency and exclude humans from the decision-making loop (Kang & Lou, 2022; Laapotti & Raappana, 2022; Sundar, 2020). Such perceptions may affect how users interpret, rate and employ the AI system (Liu et al., 2022; Pan et al., 2025; Wang, 2021). Users' experiences are largely shaped by the dynamics between human agency and AI agency during actual interactions.

In the human-AI interaction model which was proposed by Sundar (2020), the "action" route is activated when users are promised user agency and in deeper involvement with AI systems by interacting with them in an anthropomorphic manner. If users perceive more human agency, they are likelier to develop stronger trust in AI and foster greater negotiating agency (Molina & Sundar, 2022; Sundar, 2020) because perceived human agency functions as an informational cue to enhance users' perceived control of knowledge about the decision-making process and simulate the "mental states" of the AI system (Liu, 2021; Ososky et al., 2013). Providing sufficient human agency underlies positive user's reactions among users (Kang & Lou, 2022). When users perceive that the reporting mechanism follows human-made rules and is mainly executed by humans, in other words, perceive a greater degree of human agency, they are likely to show stronger use intentions. Thus, we postulated the following hypothesis:

H2: Perceived human agency is positively related to reporting to the platform.

Perceived justice of the reporting mechanism

Content moderation cannot be assessed from a single facet and thus emphasizes different sets of normative principles. Suzor (2018) evaluated the legitimacy of content moderation, examining the extent to which private governance is consensual, transparent, procedurally just and fairly enforced. FAT model highlights fairness, accountability and transparency as critical goals that are central to the spirit of content moderation (Jhaver et al., 2019b; Juneja et al., 2020).

Among these principles, justice has long been regarded as an important criterion and salient value. Colquitt (2001) divided the concept of organizational justice into four main constructs: procedural justice, distributive justice, interactional justice and informational justice. The rationale of organizational justice has been applied to the examination of users' perceptions regarding algorithmic decision-making (Binns et al., 2018; Fussell et al., 2008), providing a new theoretical perspective for elucidating users' psychological

experiences and behavioral outcomes. Gonçalves et al. (2023) positioned their research in the context of content removal, employing five indicators as different constructs of organizational justice–outcome fairness, procedural fairness, transparency, legitimacy, and trust, to examine users' support for different types of algorithmic moderation.

Expanding on this previous research, the current study incorporated related literature on organizational justice and focused on fairness and transparency as representative constructs of justice. This allowed us to dialogue with prior regulation and literature emphasizing these two values (Banovic et al., 2023; Jhaver et al., 2019a).

Regarding the effect of perceived justice, stronger perceptions of justice and fairness may increase the future tendency to abide by community norms (Tyler et al., 2021) and voluntary bystander intervention (Naab et al., 2018). Additionally, perceived justice promotes organizational citizenship behavior according to the theory of organizational justice (Moorman, 1991). RTP can be regarded as a form of civic engagement practiced by users on the platform, which does not receive direct rewards and is not obviously related to users' personal interests, but is undertaken for the purpose of assisting others and sanitizing cyberspace. Perceived justice of the reporting mechanism may motivate users to report harmful comments. First, perceived fairness prompts users to believe that the platform will handle the reports impartially. Thus, users can understand that they are likely to obtain a reasonable and satisfactory outcome through this voluntary practice. Second, perceived transparency is not only just about articulating written community rules but also involves clearly and effectively communicating with users about how these rules are interpreted and enforced in specific contexts, thus reducing users' uncertainty about the outcome. Therefore, we postulated that:

H3: Perceived fairness of the reporting mechanism is positively related to reporting to the platform.

H4: Perceived transparency of the reporting mechanism is positively related to reporting to the platform.

A Dual Perspective: The Integration of TPE and Perceptions About the Reporting Mechanism

Above, we described two categories of factors influencing users' behavior: perceptions related to media content, and content moderation that is technically and institutionally designed by the platform. The two types of variables addressed in the current study correspond to content characteristics and moderation characteristics. A few previous studies have investigated the interaction effects between these characteristics (Wang & Kim, 2023). We seek to further this exploration of how perceptions related to media content interact with those related to content moderation to shape users' behavior. There are several reasons for this integration.

First, previous scholars have attempted to broaden media effects research by revealing a series of significant conditioned factors influencing the effect of media content, including voters' candidate preferences (Kim, 2016), belief in misinformation regarding COVID-19 vaccine (Lim et al., 2025), respondents' social media use (SMU) (Luo et al., 2024), and individual political views (Kim, 2025). These research findings, which span across various contexts and backgrounds, indicate that the influences of perceived media effects are intricately nuanced when combined with the effects of factors beyond the text. Integrating media effects research with perceptions pertaining to sociotechnical systems is also a response to erstwhile call for examining media effects through the lens of social interaction and context (Katz, 2001) in the AI era.

Second, Siles and Boczkowski (2012) argued for a texto-material perspective for theorizing user agency by examining users' appropriation of media texts and material features that concurrently underline this process. In this context, material features refer to material configurations composed of codes, standards, formats and infrastructures (Fuller, 2008; Manovich, 2002) of online platforms. Users enact their agency in a hybrid social dynamic during which they interpret media context and interact with the sociotechnical system, attributing meaning to both content and material configurations (Livingstone, 2003; Siles, 2011, 2012). This provides an articulation point for the theoretical integration. Users' decisions involve both the stimulation of harmful content and the incubation of favorable affordances presented by platforms. In addition to the perceived media effects of problematic content, perceptions related to a complex of sociotechnical assemblages that steers moderation, which we conceptualized as perceived human agency and perceived justice, jointly determine users' reporting behaviors.

On top of that, some previous studies have examined the perceived transparency or fairness of the automated system, which is more inclined to users' perceptions about platform technical configurations, as a kind of algorithmic affordance (Shin, 2020; Shin & Park, 2019; Shin et al., 2020), and explored how algorithmic affordance moderates users' behavior. In accordance with this research, we proposed the following research questions (RQs):

RQ1: What are the interaction effects between PME3 and perceived human agency on reporting to the platform?

RQ2: What are the interaction effects between PME3 and perceived fairness on reporting to the platform?

RQ3: What are the interaction effects between PME3 and perceived transparency on reporting to the platform?

METHODS

Procedures

We distributed an online questionnaire to 500 users of Weibo (the most widely used social media platform in China) in March 2022. The sample composition refers to the user profile of Chinese social media and is adjusted according to the actual supply capacity of the sample pool. Fifty percent of the respondents were female, and 50 percent were male. In terms of age, respondents ranged from 16 to 63 years old, with about 37.2% aged 16-25, 27.8% aged 26-35, and 35% aged above 35, with an average age of 31.26 years (standard deviation [SD] = 8.683). Regarding education level, more than 80% of respondents had a bachelor's degree. Respondents were informed that their participation was voluntary and anonymous.

Variables and Measures

PME3

Perceived media effects of online comments on others were measured using a Likert-type scale ranging from 1 (not at all) to 7 (very much), prompting respondents to rate PME3 for five types of harmful online comments. These comments included vulgarity, insulting, inciting violence, hate speech and rumors. The ratings were averaged into corresponding indices (mean [M] = 5.7312, SD = 1.1091, α = .890).

RTP

To measure RTP, three items on a 7-point Likert-type scale (1 = strongly disagree, 7 = strongly agree) were adapted from prior scales (Guo & Johnson, 2020; Wang & Kim, 2020) (M = 5.2267, SD = 1.2874, α = .840).

Perceived human agency

Perceived human agency was measured using a 7-point Likert-type scale that included three items adopted from a previous study (Molina & Sundar, 2022) (M = 4.8473, SD = 1.1331, α = .679).

Perceived fairness

Borrowing from the measurement of perceived fairness in the context of content removal in a previous study (Gonçalves et al., 2023), we developed an index appraising the degree to which users perceive the platform's decisions about reporting requests as fair. This variable was operationalized through a Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree), asking respondents to evaluate four statements (M = 5.3460, SD = 1.1432, α = .920). Confirmatory Factor Analysis (CFA) showed that RMSEA was 0.0778 (< 0.08), CFL was 0.996 (> 0.9), TLI was 0.988 (> 0.9).

Perceived transparency

On the basis of the measurement of perceived transparency in the context of content removal reported in a previous study (Gonçalves et al., 2023), we developed an index measuring the perceived transparency of the reporting mechanism, gauging how much information users have access to regarding the platform's decision about reporting requests. This variable was measured using a Likert-type scale ranging from 1

Table 1. Descriptive statistics and correlations

	1	2	3	4	5	6	7	8	9	10
1.PME3		.506**	.038	.339**	.299**	-.001	-.100*	-.022	.068	.102*
2.RTP			.283**	.574**	.601**	.069	-.161**	.037	.093*	.485**
3.Human agency				.405**	.404**	-.001	-.055	-.029	.019	.338**
4.Fairness					.783**	-.003	-.046	.017	.034	.295**
5.Transparency						.013	-.067	.100*	.032	.387**
6.Gender							.004	-.044	-.056	.021
7.Age								-.041	-.126**	-.161**
8.Education									.123**	.071
9.SMU										.246**
10.Past reporting frequency										

Note. N = 500; *p < .05; **p < .01; ***p < .001

Table 2. Hierarchical multiple regression analyses predicting reporting harmful comments to the platform

Predictors	Step 1		Step 2		Step 3	
	B	Standard error	B	Standard error	B	Standard error
Gender	.150	.101	.153	.086	.155	.076*
Age	-.013	.006*	-.007	.005	-.009	.004
Education	.016	.093	.056	.079	.000	.071
SMU	-.033	.039	-.051	.034f	-.022	.030
Past reporting frequency	.396	.034***	.365	.029***	.250	.028***
PME3			.534	.039***	.392	.037***
Human agency					-.005	.038
Fairness					.211	.055***
Transparency					.265	.056***
R^2	.246		.454		.575	
F	32.311***		68.271***		73.597***	
ΔR^2	.239		.447		.567	
Adjusted R^2			.208		.120	

Note. N = 500; *p < .05; **p < .01; ***p < .001; B: Unstandardized coefficients

(strongly disagree) to 7 (strongly agree), asking respondents to evaluate five statements (M = 5.3624, SD = 1.1409, α = .915). CFA showed that RMSEA was 0.0452 (< 0.05), CFL was 0.997 (> 0.9), TLI was 0.994 (> 0.9).

Control variables

We controlled for three relevant demographic variables: gender (1 = male), age (M = 32.52 SD =11.42), and education. In addition, we added the frequency of SMU and previous reporting frequency. The variable of education was coded as 1-6 points from primary school and below to a master's degree and above. The variable of daily frequency of Weibo use was coded as 1-7, with 1 representing less than 10 minutes, 2 representing 10-30 minutes, 3 representing 31-60 minutes, 4 representing 1-2 hours, 5 representing 2-3 hours, 6 representing more than 3 hours, and 7 representing more than 6 hours per day. The frequency of previous reporting allowed respondents to report how often they used the reporting tool on Weibo (1 = never used, 7 = frequently used). **Appendix A** shows variables and measures.

RESULTS

SPSS 26 and PROCESS plug-ins were employed for data analysis. The statistical analyses were divided into four parts. First, descriptive statistics and correlations were calculated. We then conducted hierarchical linear regression to test our hypotheses **H2** to **H4**. We created a model in which the participants' reporting to the platform served as the dependent variable. In step 1 of the regression model, we added the control variables age, gender, education, SMU, and previous reporting frequency. In step 2, we introduced the independent variable PME3 for the "content" layer. In step 3, we added the independent variable perceived human agency, perceived fairness, and perceived transparency for the "moderation" layer.

Table 1 shows the significant correlations between variables. **Table 2** shows the results of hierarchical multiple regression analyses. Regarding **H1**, the regression model revealed a significant influence of participants' PME3 on their reporting behaviors (B = .392, p < .001). Thus, **H1** was supported.

Table 3. Testing the moderating effect of PME3 and perceived human agency/fairness/transparency

Predictors	Reporting to the platform		Reporting to the platform		Reporting to the platform	
	B	Standard error	B	Standard error	B	Standard error
Constant	-2.9979***	.8857	-.4234	.6922	-.3498	.7384
PME3	1.0979***	.1407	.4168***	.1117	.4585***	.1197
Human agency	.7995***	.1607				
Fairness			.4117***	.1146		
Transparency					.4670***	.1241
PME3*human agency	**-.1126*** **	**.0270**				
PME3*transparency					-.0079	.0215
PME3*fairness			-.0025	.0203		
Gender	.1380	.0835	.1620*	.0776	.1467	.0772
Age	-.0055	.0049	-.0086	.0046	-.0080	.0045
Education	.0080	.0781	.0449	.0720	-.0280	.0719
SMU	-.0318	.0329	-.0323	.0305	-.0191	.0304
Past reporting frequency	.3160***	.0300	.2829***	.0274	.2477***	.0281
R^2	.4873	.5556	.5621			
F	58.3341***	76.7196***	78.7677***			

Note. N = 500; Each column is a regression model that predicts the variable at the top of the column; *p < .05; **p < .01; ***p < .001; B: unstandardized coefficients

Perceived fairness and perceived transparency were significantly positively related to RTP (B = .211, p < .001; B = .265, p < .001). Thus, **H3** and **H4** were supported. The positive correlation between perceived human agency and reporting to the platform was not significant (B = -.005, p > .05), providing no support for **H2**.

Finally, to answer **RQ1**, **RQ2**, and **RQ3**, we executed the PROCESS macro (model 1) developed by Hayes (2017) to test the moderating effect. **Table 3** shows the regression analysis results after incorporating interaction terms. Human agencies moderate the relationship between PME3 and reporting to the platform (B = -.1126, p < .001). The interaction effects between PME3 and fairness, as well as between PME3 and transparency, were not significant (p > .05).

To decompose the interaction effects, we plotted the slope, which illustrates the direct relationships for different levels of human agency. As shown in **Figure 1**, when human agency was lower, the correlation between PME3 and reporting to the platform was higher (B = 0.680, t = 15.307, p < 0.001), whereas the slope with higher human agency was lower (B = 0.425, t = 17.617, p < 0.001).

DISCUSSION & IMPLICATIONS

The current study demonstrated that PME3, perceived fairness, and perceived transparency influenced users' tendency to report harmful comments to the platform. Also, we found that perceived human agency played a moderating role in the effects of PME3 on users' reporting of harmful comments to the platform.

Based on the IPI model, our research examined the relationship between PME3 and behavioral consequences, focusing on users' reporting of harmful comments to the platform. The findings of the current study indicated that PME3 is a robust and significant predictor of corrective actions, in line with previous appeals for more investment in the behavioral domain and corrective reactions (Lim & Golan, 2011; Sun et al., 2022). The current findings indicate that when users perceive the detrimental effects of media content on the public, they desire platforms to take on the responsibility to protect vulnerable others and are willing to share the expense of detecting problematic content. The significant influences of transparency and fairness can be inspected from two perspectives:

1. From the instrumental view, substantive and valuable information allows users to experience a sense of control and predictability (Lind & Tyler, 2013) regarding decisions, without requiring a high level of cognitive effort for processing (ter Hoeven et al., 2021). Useful information helps align users' perception with social norms (Jhaver et al., 2019b) and improve efficacy for attaining desired outcomes.
2. From a moral perspective, if users believe that the reply given by the platform is fair, they are likely to feel that their moral values regarding the judgment of "inappropriateness" have been respected and recognized by the platform. Users' willingness to engage in content moderation is likely to be promoted after their morality-driven actions result in favorable returns (Blau, 2017).

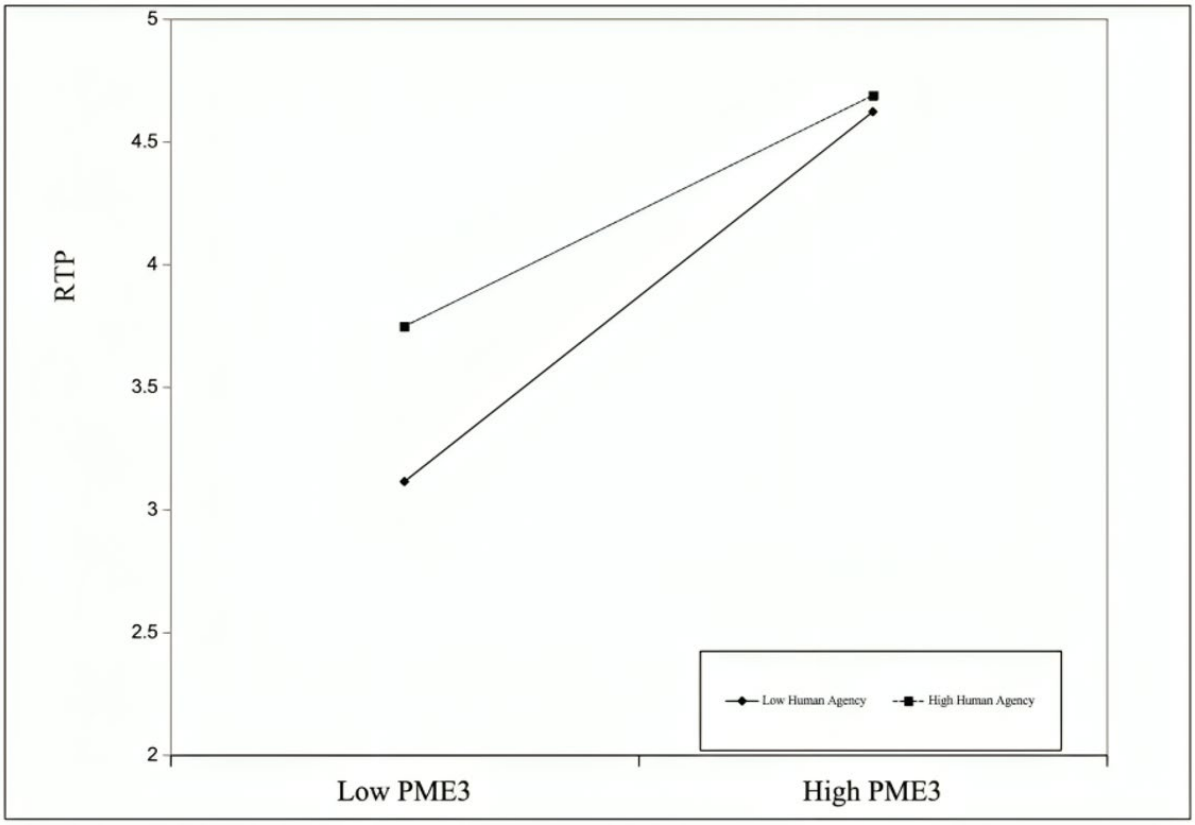

Figure 1. Interaction effect of PME3 and perceived human agency in reporting harmful comments to the platform (high and low levels of PME3 and human agency represent one SD above and below the M, respectively) (Source: Authors)

The triggering role of transparency and fairness indicates a discrepancy between users and platform in deciphering the reporting mechanism, in accord with previous calls for efforts to identify gaps between moderation practices and public demands (Riedl et al., 2021). For platforms, flagging and reporting are indispensable for achieving governance efficiency and justifying censorial actions (Crawford & Gillespie, 2016). But users do not identify themselves as distributed labor employed by the platform. Instead, reporting serves as a mechanism for users to exert autonomy and give input into governance (Chipidza & Yan, 2022; Flynn et al., 2025). Zhang et al. (2023) proposed the metaphor of a civilian picking up trash to clean the street out of moral motivations to represent users' engagement in reporting behavior. Public values, such as fairness and transparency, which ensure that users are indeed involved in the process, are more endorsed by users and may diverge from the platform's goals pursuing for efficiency (Shim & Jhaver, 2024; Zhang et al., 2023). The introduction of organizational justice theory in this study advances existing literature in the following ways:

1. Although organizational justice has been applied in the context of content removal (Gonçalves et al., 2023), we extended the research scope of this theory to the context of reporting mechanisms. The process of users reporting to the platform and awaiting an adjudication is analogous to a decision-making process within an organization, thus rendering the theory applicable. Given the emergence of privatized bureaucracy in platform governance (Balkin, 2018), future research can leverage theories from organizational theory to examine platform-user interactions.
2. Prior research about justice has mainly been conducted under punitive paradigms in which users are "offenders" of community norms (Jhaver et al., 2019b; Juneja et al., 2020). The current study goes further by demonstrating that justice also promotes users' voluntary self-moderation for common good orientation (Friess et al., 2021), prompting them to take on nurturing and supportive roles (Seering et al., 2022).

A noteworthy fact is that although perceived human agency shows non-significant relationship with reporting to the platform when accounting for fairness and transparency, a significant interaction effect of human agency and PME3 on reporting to the platform was verified. After controlling for fairness and

transparency, the relationship between human agency and reporting is non-significant. This may be because that human agency exerts its influence on reporting through the mechanisms of fairness and transparency, in alignment with Liu (2021)'s finding that agency locus influences users' judgments of transparency.

The significant interaction effect demonstrated that perceived human agency inhibited the catalyzing effect of PME3, revealing that the relationship between PME3 and reporting to the government was weaker ($B = 0.425$, $p < 0.001$) among individuals with comparatively high levels of perceived human agency and stronger ($B = 0.680$, $p < 0.001$) among those who perceived lower human agency. Perceived human agency acted as an antecedent leading to human behavioral outcome during human-machine interaction in the previous study (Molina & Sundar, 2022). We tentatively examined the role of perceived human agency in the interaction with media effects. Previous research demonstrated that the influence of presumed media effects may be amplified or attenuated by a person's beliefs in other factors (Kim, 2025; Lim et al., 2025). The current study enriches the theoretical landscape of HMI and media effects research by revealing that content and moderation characteristics do not function in isolation but operated in tandem. Individuals exhibiting a higher level of perceived human agency show greater confidence in the thinking and acting capabilities of the reporting mechanism, reducing dependence on PME3. In contrast, individuals who perceive a lower level of human agency are less likely to believe that the reporting mechanism possesses human judging capabilities and are skeptical about its effectiveness. Therefore, they require more intense motivation regarding deleterious content to trigger their actions. This similar attenuating effect has also been observed in other research. Ji and Kim (2020) found that in crisis communication when publics perceive a high level of government controllability or consumer collective efficacy, the impact of issue involvement on public reactions might be impaired because they rely more heavily on the incentives from confidence in governmental or collective capabilities. This finding also underscores the need for advancing conventional media effect research by considering environmental factors.

The effects of the agency of the moderator on users' reactions have not been sufficiently addressed in content moderation research (Meerson et al., 2025). Because collaborative human-AI moderation has become a prevalent strategy receiving insufficient research attention, we focused on the concept of agency locus to investigate users' preferences. Previous studies have referred to the notion of attribution of agency in human-machine interaction (Molina & Sundar, 2022; Pan et al., 2025) and related it to users' engagement. However, a consensus regarding whether human agency is positively associated with users' behavioral outcomes has not yet been reached. A meta-analysis has suggested that future research should furnish empirical insights into human-machine interaction in heterogeneous contexts considering the different roles of AI communicators (Huang & Wang, 2023). The current study explored the specific effects of perceived human agency being localized in an understudied context. The results indicate that when AI acts as a curator deciding the appropriateness of content, perceived human agency of the reporting mechanism serves as an important cognitive cue for inducing reporting, even exceeding the influence of PME3.

Previous studies have largely regarded reporting as an individual's independent coping behavior from a single perspective. Our study represents a theoretical advancement in adopting a dual perspective considering perceptions surrounding media content and content moderation. Jhaver and Zhang (2025) distinguished two types of content moderation based on platform-enacted decisions or individual control. Reporting incorporates the characteristics of both. On one hand, the reporting tool offered by platforms allows users to align content with their tastes. On the other hand, the identification and removal of problematic content is subject to the platform's decisions. User behavior is influenced not only by perceptions and preferences regarding media content, but also perceptions of how platforms formulate and enforce rules. Researchers have argued that users make calculations regarding the threat of online misbehavior and the effectiveness of investing in making a report (Wong et al., 2016). This finding resonated with our view that the stimulation of harmful content, perceived just and humanized platform-led moderation are dynamically shaping user behavior. Reporting implies a collaboration connecting personal moderation enacted by decentralized end-users and coercive measures enforced by platforms.

It should be noted that although our study was conducted solely on Weibo, the findings may be generalized to other platforms. First, the core argument in this study that "perceptions related to content and moderation jointly influence users' reporting behavior" reveals a universal psychological mechanism for individuals. The abstraction and theorization of the argument makes it independent of the specific platform

interface. Moreover, because flagging is required by governmental policies in China (Xie et al., 2023), the reporting mechanisms used by different platforms follow a largely similar process, with differences only in the rubrics and icons. Second, Weibo remains one of the major avenues for discussing public issues in China and being subject to a substantial amount of toxic communication (Li, 2023). The current study provides a valuable reference for understanding how to shape civil public discourse.

The findings presented here also have practical implications for the platforms governing their sites. We found that individuals tend to overestimate media effects on others and engage themselves in moderation based on their perceptions of the degree to which others are affected. Thus, it is necessary to assess the actual and perceived media effects of different types of harmful comments. Our results highlight the need to establish a transparent and fair reporting mechanism to encourage participation. To achieve this goal, platform operators should provide explicit policies and standardized procedures for facilitating user reporting. Comprehensive explanations about a platform's decisions are imperative for users to become well-informed regarding platform terms. The emphasis users place on fairness suggests that platforms can regularly collect users' appraisal of the reporting results for integrating public values into interface architecture, features, and policies (Chen et al., 2025). Because perceived human agency has an impact on users' motivations, it is important for platform designers to convey that human judgment and thinking abilities are involved in the decision-making processes. For instance, platforms can specify in their norms that professional human moderators will intervene when certain sensitive situations are encountered. When users report severe online harms, the platform's provision of care and support can then be perceived as humanized (Cover et al., 2025). There is an urgent need to align reporting frameworks with contemporary human values.

CONCLUSION

Through a survey in China, the current study developed a dual "content-moderation" framework exploring the antecedents of RTP in light of perceptions of content and moderation. The current study represents two main significant advancements. First, our results deepen human-machine interaction research on how user perspectives shape their behavior by weaving in critical aspects–perceived justice and perceived human agency of the reporting mechanism. Second, our findings expand on previous literature that has predominantly regarded reporting as an individual coping behavior by incorporating PME3 and perceptions related to the reporting mechanism. This approach not only broadens the theoretical base of TPE but also reveals crucial insights into the "platform-user interaction" essence of the reporting mechanism. Additionally, the current findings provide practical implications for platform operators to foster user participation.

Despite these contributions, the current study involved several limitations that should be considered. First, our study was conducted using a single platform, Weibo. Although this platform has a relatively high level of representativeness in China's cyberspace, further research should examine other prominent platforms, such as short video-based platforms, for comparison. Second, we used self-report methods, which may introduce a degree of subjectivity and lead to bias. In the future, more objective methods could be employed to observe user behavior. Finally, the current study did not focus on variations in different types of harmful comments. However, different types of harmful content may not elicit identical responses (Jhaver & Zhang, 2025). Further studies should focus on differentiating among various types of harmful comments. Overall, we are looking forward to more nuanced and robust research that expands on the current findings.

Author contributions: YZ: conceptualization, formal analysis, methodology, software, data curation, writing – original draft, writing – review & editing, supervision, validation, investigation; **ZS:** conceptualization, formal analysis, methodology, software, data curation, investigation, writing – original draft, writing – review & editing, visualization, funding acquisition, resources, project administration. Both authors approved the final version of the article.

Funding: This article was supported by the Youth Foundation of Humanities and Social Sciences Research of the Ministry of Education of China (grant number: 23YJC860022).

Ethics declaration: This study was approved by the Ethical Committee of School of New Media at Peking University on 1 March 2022 with approval code 2022030101. All participants were informed about the purposes of the research and written consent was obtained from them. Confidentiality was guaranteed through the anonymization of personal data. We ensured that the data we collected was used exclusively for academic research.

Declaration of interest: The authors declared no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Aljasir, S. (2023). Effect of online civic intervention and online disinhibition on online hate speech among digital media users. *Online Journal of Communication and Media Technologies, 13*(4), Article e202344. https://doi.org/10.30935/ojcmt/13478

Balkin, J. M. (2018). Free speech is a triangle. *Columbia Law Review, 118*(7), 2011-2056.

Banovic, N., Yang, Z., Ramesh, A., & Liu, A. (2023). Being trustworthy is not enough: How untrustworthy artificial intelligence (AI) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1-17. https://doi.org/10.1145/3579460

Bhandari, A., Ozanne, M., Bazarova, N. N., & DiFranzo, D. (2021). Do you care who flagged this post? Effects of moderator visibility on bystander behavior. *Journal of Computer-Mediated Communication, 26*(5), 284-300. https://doi.org/10.1093/jcmc/zmab007

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). *'It's reducing a human being to a percentage' Perceptions of justice in algorithmic decisions* [Paper presentation]. The 2018 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3173574.3173951

Blau, P. (2017). *Exchange and power in social life*. Routledge. https://doi.org/10.4324/9780203792643

Chen, X., Guan, T., & Yang, Y. (2025). Allocating content governance responsibility in China: Heterogeneous public attitudes toward multistakeholder involvement strategies. *Policy & Internet, 17*(2), Article e432. https://doi.org/10.1002/poi3.432

Chipidza, W., & Yan, J. (2022). The effectiveness of flagging content belonging to prominent individuals: The case of Donald Trump on Twitter. *Journal of the Association for Information Science and Technology, 73*(11), 1641-1658. https://doi.org/10.1002/asi.24705

Chung, M., & Wihbey, J. (2024). Social media regulation, third-person effect, and public views: A comparative study of the United States, the United Kingdom, South Korea, and Mexico. *New Media & Society, 26*(8), 4534-4553. https://doi.org/10.1177/14614448221122996

Chung, S., & Moon, S.-I. (2016). Is the third-person effect real? A critical examination of rationales, testing methods, and previous findings of the third-person effect on censorship attitudes. *Human Communication Research, 42*(2), 312-337. https://doi.org/10.1111/hcre.12078

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology, 86*(3), 386-400. https://doi.org/10.1037//0021-9010.86.3.386

Common, M. (2019). *The importance of appeals systems on social media platforms*. SSRN. https://doi.org/10.2139/ssrn.3462770

Cover, R., Beckett, J., Brevini, B., Lumby, C., Simcock, R., & Thompson, J. D. (2025). Reporting online abuse to platforms: Factors, interfaces and the potential for care. *Convergence: The International Journal of Research into New Media Technologies*. https://doi.org/10.1177/13548565251324508

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society, 18*(3), 410-428. https://doi.org/10.1177/1461444814543163

Dahlberg, L. (2011). Re-constructing digital democracy: An outline of four 'positions'. *New Media & Society, 13*(6), 855-872. https://doi.org/10.1177/1461444810389569

Davison, W. P. (1983). The third-person effect in communication. *The Public Opinion Quarterly, 47*(1), 1-15. https://doi.org/10.1086/268763

Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and Brain Sciences, 11*(3), 495-505. https://doi.org/10.1017/S0140525X00058611

Dias Oliva, T. (2020). Content moderation technologies: Applying human rights standards to protect freedom of expression. *Human Rights Law Review, 20*(4), 607-640. https://doi.org/10.1093/hrlr/ngaa032

Duttweiler, P. C. (1984). The internal control index: A newly developed measure of locus of control. *Educational and Psychological Measurement, 44*(2), 209-221. https://doi.org/10.1177/0013164484442004

Flynn, A., Vakhitova, Z., Wheildon, L., Harris, B., & Robards, B. (2025). Content moderation and community standards: The disconnect between policy and user experiences reporting harmful and offensive content on social media. *Policy & Internet, 17*(3), Article e70006. https://doi.org/10.1002/poi3.70006

Fuller, M. (2008). *Software studies: A lexicon*. MIT Press. https://doi.org/10.7551/mitpress/9780262062749.001.0001

Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008). *How people anthropomorphize robots* [Paper presentation]. The 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction. https://doi.org/10.1145/1349822.1349842

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. https://doi.org/10.12987/9780300235029

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society, 8*(3). https://doi.org/10.1177/20563051221117552

Goanta, C., & Ortolani, P. (2022). *Unpacking content moderation: The rise of social media platforms as online civil courts*. SSRN. https://doi.org/10.2139/ssrn.3969360

Gonçalves, J., Weber, I., Masullo, G. M., Torres da Silva, M., & Hofhuis, J. (2023). Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. *New Media & Society, 25*(10), 2595-2617. https://doi.org/10.1177/14614448211032310

Gorwa, R. (2018). Towards fairness, accountability, and transparency in platform governance. *AoIR Selected Papers of Internet Research, 2018*. https://doi.org/10.5210/spir.v2018i0.10483

Grimmelmann, J. (2015). The virtues of moderation. *The Yale Journal of Law & Technology, 17*(42), 42-109.

Gunther, A. C., & Storey, J. D. (2003). The influence of presumed influence. *Journal of Communication, 53*(2), 199-215. https://doi.org/10.1111/j.1460-2466.2003.tb02586.x

Guo, L., & Johnson, B. G. (2020). Third-person effect and hate speech censorship on Facebook. *Social Media + Society, 6*(2). https://doi.org/10.1177/2056305120923003

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis, second edition: A regression-based approach*. Guilford Publications.

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society, 34*(1), 1-14. https://doi.org/10.1080/01972243.2017.1391913

Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. *Journal of Communication, 73*(6), 552-562. https://doi.org/10.1093/joc/jqad024

Jhaver, S., & Zhang, A. X. (2025). Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society, 27*(5), 2930-2950. https://doi.org/10.1177/14614448231217993

Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019a). "Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1-33. https://doi.org/10.1145/3359294

Jhaver, S., Bruckman, A., & Gilbert, E. (2019b). Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1-27. https://doi.org/10.1145/3359252

Ji, Y., & Kim, S. (2020). Crisis-induced public demand for regulatory intervention in the social media era: Examining the moderating roles of perceived government controllability and consumer collective efficacy. *New Media & Society, 22*(6), 959-983. https://doi.org/10.1177/1461444819874473

Juneja, P., Rama Subramanian, D., & Mitra, T. (2020). Through the looking glass: Study of transparency in Reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction, 4*(GROUP), 1-35. https://doi.org/10.1145/3375197

Kalch, A., & Naab, T. K. (2017). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *SCM Studies in Communication and Media, 6*(4), 395-419. https://doi.org/10.5771/2192-4007-2017-4-395

Kang, H., & Lou, C. (2022). AI agency vs. human agency: Understanding human-AI interactions on TikTok and their implications for user engagement. *Journal of Computer-Mediated Communication, 27*(5), Article zmac014. https://doi.org/10.1093/jcmc/zmac014

Katz, E. (2001). Lazarsfeld's map of media effects. *International Journal of Public Opinion Research, 13*(3), 270-279. https://doi.org/10.1093/ijpor/13.3.270

Kim, H. (2016). The role of emotions and culture in the third-person effect process of news coverage of election poll results. *Communication Research, 43*(1), 109-130. https://doi.org/10.1177/0093650214558252

Kim, M. (2025). A direct and indirect effect of third-person perception of COVID-19 fake news on support for fake news regulations on social media: Investigating the role of negative emotions and political views. *Mass Communication and Society, 28*(2), 229-252. https://doi.org/10.1080/15205436.2023.2227601

Kunst, M., Porten-Cheé, P., Emmer, M., & Eilders, C. (2021). Do "good citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics, 18*(3), 258-273. https://doi.org/10.1080/19331681.2020.1871149

Laapotti, T., & Raappana, M. (2022). Algorithms and organizing. *Human Communication Research, 48*(3), 491-515. https://doi.org/10.1093/hcr/hqac013

Leerssen, P. (2023). An end to shadow banning? Transparency rights in the digital services act between content moderation and curation. *Computer Law & Security Review, 48*, Article 105790. https://doi.org/10.1016/j.clsr.2023.105790

Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *SCM Studies in Communication and Media, 7*(4), 555-579. https://doi.org/10.5771/2192-4007-2018-4-555

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology, 31*(4), 611-627. https://doi.org/10.1007/s13347-017-0279-x

Li, M. (2023). Promote diligently and censor politely: How Sina Weibo intervenes in online activism in China. *Information, Communication & Society, 26*(4), 730-745. https://doi.org/10.1080/1369118X.2021.1983001

Lim, J. S., & Golan, G. J. (2011). Social media activism in response to the influence of political parody videos on YouTube. *Communication Research, 38*(5), 710-727. https://doi.org/10.1177/0093650211405649

Lim, J. S., Lee, C., Kim, J., & Zhang, J. (2025). Influence of COVID-19 vaccine misinformation beliefs on the third-person effect: Implications for social media content moderation and corrective action. *Online Information Review, 49*(3), 497-516. https://doi.org/10.1108/OIR-04-2024-0220

Lind, E. A., & Tyler, T. R. (2013). *The social psychology of procedural justice*. Springer.

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human-AI interaction. *Journal of Computer-Mediated Communication, 26*(6), 384-402. https://doi.org/10.1093/jcmc/zmab013

Liu, Y., Mittal, A., Yang, D., & Bruckman, A. (2022). *Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing* [Paper presentation]. The 2022 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.3517731

Livingstone, S. (2003). The changing nature of audiences: From the mass audience to the interactive media user. In A. N. Valdivia (Ed.), *A companion to media studies* (pp. 337-359). Blackwell. https://doi.org/10.1002/9780470999066.ch17

Luo, C., Zhu, Y., & Chen, A. (2024). What motivates people to counter misinformation on social media? Unpacking the roles of perceived consequences, third-person perception and social media use. *Online Information Review, 48*(1), 105-122. https://doi.org/10.1108/OIR-09-2022-0507

Lyu, Y., Cai, J., Callis, A., Cotter, K., & Carroll, J. M. (2024). *"I got flagged for supposed bullying, even though it was in response to someone harassing me about my disability": A study of blind TikTokers' content moderation experiences* [Paper presentation]. The 2024 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3613904.3642148

Manovich, L. (2002). *The language of new media*. University of Toronto Press. https://doi.org/10.22230/cjc.2002v27n1a1280

Meerson, R., Koban, K., & Matthes, J. (2025). Platform-led content moderation through the bystander lens: A systematic scoping review. *Information, Communication & Society*. https://doi.org/10.1080/1369118X.2025.2483836

Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication, 27*(4), Article zmac010. https://doi.org/10.1093/jcmc/zmac010

Moorman, R. H. (1991). Relationship between organizational justice and organizational citizenship behaviors: Do fairness perceptions influence employee citizenship? *Journal of Applied Psychology, 76*(6), 845-855. https://doi.org/10.1037/0021-9010.76.6.845

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society, 20*(11), 4366-4383. https://doi.org/10.1177/1461444818773059

Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society, 20*(2), 777-795. https://doi.org/10.1177/1461444816670923

Obermaier, M. (2024). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media & Society, 26*(8), 4785-4807. https://doi.org/10.1177/14614448221125417

Ososky, S., Philips, E., Schuster, D., & Jentsch, F. (2013). *A picture is worth a thousand mental models: Evaluating human understanding of robot teammates* [Paper presentation]. The Human Factors and Ergonomics Society Annual Meeting. https://doi.org/10.1177/1541931213571287

Pan, W., Liu, D., Meng, J., & Liu, H. (2025). Human-AI communication in initial encounters: How AI agency affects trust, liking, and chat quality evaluation. *New Media & Society, 27*(10), 5822-5847. https://doi.org/10.1177/14614448241259149

Porten-Cheé, P., Kunst, M., & Emmer, M. (2020). Online civic intervention: A new form of political participation under conditions of a disruptive online discourse. *International Journal of Communication, 14*, 514-534.

Riedl, M. J., Naab, T. K., Masullo, G. M., Jost, P., & Ziegele, M. (2021). Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. *Policy & Internet, 13*(3), 433-451. https://doi.org/10.1002/poi3.257

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological monographs: *General and Applied, 80*(1), 1-28. https://doi.org/10.1037/h0092976

Schmid, U. K., Obermaier, M., & Rieger, D. (2024). Who cares? How personal political characteristics are related to online counteractions against hate speech. *Human Communication Research, 50*(3), 393-403. https://doi.org/10.1093/hcr/hqae004

Schwarz, O. (2019). Facebook rules: Structures of governance in digital capitalism and the control of generalized social capital. *Theory, Culture & Society, 36*(4), 117-141. https://doi.org/10.1177/0263276419826249

Seering, J., Kaufman, G., & Chancellor, S. (2022). Metaphors in moderation. *New Media & Society, 24*(3), 621-640. https://doi.org/10.1177/1461444820964968

Shim, Y., & Jhaver, S. (2024). *Incorporating procedural fairness in flag submissions on social media platforms*. arXiv. https://doi.org/10.48550/arXiv.2409.08498

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media, 64*(4), 541-565. https://doi.org/10.1080/08838151.2020.1843357

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*(9), 277-284. https://doi.org/10.1016/j.chb.2019.04.019

Shin, D., Zhong, B., & Biocca, F. A. (2020). Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management, 52*(6), Article 102061. https://doi.org/10.1016/j.ijinfomgt.2019.102061

Siles, I. (2011). From online filter to web format: Articulating materiality and meaning in the early history of blogs. *Social Studies of Science, 41*(5), 737-758. https://doi.org/10.1177/0306312711420190

Siles, I. (2012). Web technologies of the self: The arising of the "blogger" identity. *Journal of Computer-Mediated Communication, 17*(4), 408-421. https://doi.org/10.1111/j.1083-6101.2012.01581.x

Siles, I., & Boczkowski, P. (2012). At the intersection of content and materiality: A texto-material perspective on the use of media technologies. *Communication Theory, 22*(3), 227-249. https://doi.org/10.1111/j.1468-2885.2012.01408.x

Šori, I., & Vehovar, V. (2022). Reported user-generated online hate speech: The 'ecosystem', frames, and ideologies. *Social Sciences, 11*(8), 375-419. https://doi.org/10.3390/socsci11080375

Stockinger, A., Schäfer, S., & Lecheler, S. (2025). Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools. *New Media & Society, 27*(3), 1215-1234. https://doi.org/10.1177/14614448231190901

Sun, Y., Oktavianus, J., Wang, S., & Lu, F. (2022). The role of influence of presumed influence and anticipated guilt in evoking social correction of COVID-19 misinformation. *Health Communication, 37*(11), 1368-1377. https://doi.org/10.1080/10410236.2021.1888452

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAII). *Journal of Computer-Mediated Communication, 25*(1), 74-88. https://doi.org/10.1093/jcmc/zmz026

Suzor, N. P. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media + Society, 4*(3). https://doi.org/10.1177/2056305118787812

ter Hoeven, C. L., Stohl, C., Leonardi, P., & Stohl, M. (2021). Assessing organizational information visibility: Development and validation of the information visibility scale. *Communication Research, 48*(6), 895-927. https://doi.org/10.1177/0093650219877093

Tyler, T., Katsaros, M., Meares, T., & Venkatesh, S. (2021). Social media governance: Can social media companies motivate voluntary rule following behavior among their users? *Journal of Experimental Criminology, 17*(1), 109-127. https://doi.org/10.1007/s11292-019-09392-z

Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). "At the end of the day Facebook does what it wants" How users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), 1-22. https://doi.org/10.1145/3415238

van Dijck, J. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Wang, S. (2021). Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital Journalism, 9*(1), 64-83. https://doi.org/10.1080/21670811.2020.1851279

Wang, S., & Kim, K. J. (2020). Restrictive and corrective responses to uncivil user comments on news websites: The influence of presumed influence. *Journal of Broadcasting & Electronic Media, 64*(2), 173-192. https://doi.org/10.1080/08838151.2020.1757368

Wang, S., & Kim, K. J. (2023). Content moderation on social media: Does it matter who and why moderates hate speech? *Cyberpsychology, Behavior, and Social Networking, 26*(7), 527-534. https://doi.org/10.1089/cyber.2022.0158

Watson, B. R., Peng, Z., & Lewis, S. C. (2019). Who will intervene to save news comments? Deviance and social control in communities of news commenters. *New Media & Society, 21*(8), 1840-1858. https://doi.org/10.1177/1461444819828328

Wilhelm, C., Joeckel, S., & Ziegler, I. (2019). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research, 47*(6), 921-944. https://doi.org/10.1177/0093650219855330

Wong, R. Y., Cheung, C. M., & Xiao, B. (2016). *Combating online abuse: What drives people to use online reporting functions on social networking sites* [Paper presentation]. The 2016 49th Hawaii International Conference on System Sciences. https://doi.org/10.1109/HICSS.2016.58

Xie, X., Shi, L., & Zhu, Y. (2023). Why netizens report harmful content online: A moderated mediation model. *International Journal of Communication, 17*, 5830-5851.

Young, G. K. (2022). How much is too much: The difficulties of social media content moderation. *Information & Communications Technology Law, 31*(1), 1-16. https://doi.org/10.1080/13600834.2021.1905593

Zhang, A. Q., Montague, K., & Jhaver, S. (2023). *Cleaning up the streets: Understanding motivations, mental models, and concerns of users flagging social media posts*. arXiv. https://doi.org/10.48550/arXiv.2309.06688

Zhao, L., & Zhang, R. (2024). Unpacking platform governance through meaningful human agency: How Chinese moderators make discretionary decisions in a dynamic network. *New Media & Society, 27*(12), 6472-6491. https://doi.org/10.1177/14614448241274457

Ziegele, M., Naab, T. K., & Jost, P. (2019). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society, 22*(5), 731-751. https://doi.org/10.1177/1461444819870130

APPENDIX A

1. **Perceived media effects on others**
   (a) Vulgarity
   (b) Insulting
   (c) Inciting violence
   (d) Hate speech and discrimination
   (e) Rumors
2. **Report harmful comments to the platform**
   (a) I would report the comment to the platform.
   (b) I would e-mail and ask the platform to remove the comment.
   (c) I would like to submit a complaint to the platform regarding the comment.
3. **Perceived human agency**
   (a) The reporting mechanism is controlled by humans.
   (b) The rules for the reporting mechanism are made by humans.
   (c) The rules for the reporting mechanism are enforced by humans.
4. **Perceived fairness**
   (a) The handling results of reports on Weibo are generally appropriate.
   (b) I think the way Weibo handles reports is generally fair.
   (c) Handling results of the reports on Weibo are generally reasonable.
   (d) Overall, I am satisfied with the handling of the reports on Weibo.
5. **Perceived transparency**
   (a) I know how to report a comment to the platform.
   (b) I know how the platform deals with reported comments.
   (c) I know which sensitive words are set on the platform, and what contents will be deleted or blocked.
   (d) I know what type of contents is easier to report successfully.
   (e) If my comments are deleted or blocked, I know why.