International Journal of Communication 19(2025), 2882–2905 1932–8036/20250005

Antecedents of Reporting Harmful Comments: Testing the Moderating Role of Perceived Transparency

XINZHOU XIE[1]
Peking University, China

ZHUO SONG
Nanjing Normal University, China

QIYU BAI
Peking University, China

This study examines the antecedents of reporting harmful comments on the platform from a user's perspective, focusing on perceived media effects. Although some research has examined the "report" or "flag" function as a sociotechnical apparatus for content moderation on social media platforms, little attention has been paid to users' perceptions and thoughts. Through a survey of Internet users in China (N = 3000), the present study builds a model to explicate how audience feelings and judgments may elicit attitudinal and behavioral responses. We propose two potential mediators—responsibility attribution to the platform and support for content moderation—and explore the sophisticated mechanisms through which users' perceptions influence their engagement in content moderation. Moreover, this study tests the moderating role of perceived transparency and disentangles the interplay between audience attitudes and platform operation. The results also provide practical implications for platform providers and the state to create a more constructive public space for online deliberation.

Keywords: content moderation, platform governance, flag, media effect, perceived transparency

Social media has evolved into a vibrant platform for free expression and opinion exchange about current issues (Wang & Kim, 2020), even the birthplace of public events and the battlefield of diverse

Xinzhou Xie: xzxie@pku.edu.cn
Zhuo Song: pkuviola@163.com
Qiyu Bai (corresponding author): baiqiyu_pku@163.com
Date submitted: 2024-02-26

[1] The authors appreciate the feedback and suggestions from Silvio Waisbord, Kady Bell-Garcia, and the anonymous reviewers. This research was supported by the Youth Foundation of Humanities and Social Sciences Research of the Ministry of Education of China (Grant No. 23YJC860022).

opinions. Along with this, however, harmful comments also appear frequently on social platforms. Harmful comments are the most easily produced problematic content compared to other types, such as deep fake, and the damage to the conversation atmosphere is more immediate because the most important public opinion platforms in China today, such as Weibo, still exist in the form of text. The regulation of harmful comments is under public and scholarly debate.

To cope with toxic voices in comment sections, some platforms have implemented various means of moderation, such as reporting or "flagging" such comments. Flag is a sociotechnical apparatus on social media platforms, referring to the process by which users report comments that violate social norms, after which algorithms and human moderators decide whether to block or delete the relevant comment and penalize the poster (Crawford & Gillespie, 2016).

In China, "flag" also means "complaint" or "report." It is not only seen as diligent self-regulation by the platform or the rhetoric of justification (Crawford & Gillespie, 2016), but is also required by government provisions (Xie, Shi, & Zhu, 2023). This practice has a profound social foundation. For example, although Douyin rarely discloses the number of user reports directly, it shows its governance achievements by announcing a decline in user reports.

Since user participation is a precondition for the effectiveness of this mechanism, it is essential to empirically investigate the factors influencing Chinese users' reporting behavior and identify gaps between platform interventions and public understandings. Current research has paid attention to a critical review of flag (Crawford & Gillespie, 2016), users' folk theories of flag (Myers West, 2018; Suzor, Myers West, Quodling, & York, 2019), and factors predicting reporting behaviors (Kalch & Naab, 2017; Kunst, Porten-Cheé, Emmer, & Eilders, 2021; Wilhelm, Joeckel, & Ziegler, 2019; Xie et al., 2023). The last stream of research has some shortcomings. First, little is known about critical questions regarding what users perceive and think as constituents, not to mention the underlying psychological mechanisms involved (Riedl, Whipple, & Wallace, 2022); Second, extant studies often ignore the sociotechnical context that is institutionalized by the platform (Gillespie, 2018; van Dijck, 2018). Users communicate with each other within the sociotechnical architecture afforded by online platforms. Design choices and users' perceptions about these sociotechnical systems may steer how users interpret, engage, and interact (Helberger, Pierson, & Poell, 2018; Jhaver, Appling, Gilbert, & Bruckman, 2019). For example, Bhandari, Ozanne, Bazarova, and DiFranzo (2021) figured out the complex effects of moderator visibility, which is technically designed by the platform on bystanders' subsequent flagging behaviors.

To address this understudied aspect, our study aims to explore the antecedents of users reporting harmful comments to platforms, responding to calls for more academic assessments of users' attitudes (Einwiller & Kim, 2020). Integrating the related theories, we propose two potential mediators: responsibility attribution to the platform and support for platform content moderation. In addition, this study is among the first to test the moderating role of perceived transparency—one of the most salient values shaping users' perceptions of the sociotechnical system—and its effectiveness is still inconclusive. The findings explicate how users' perceptions enable them to engage in content moderation and disentangle the interplay between audience attitudes and platform operation. The results also provide

practical implications for platform providers and the government to create a more constructive public space for online deliberation.

Literature Review & Hypotheses

Content Moderation in China

In the realm of online platforms, content moderation indicates a dynamic in which the operators of platforms serve as setters of norms, interpreters of laws, arbiters of taste, adjudicators of disputes, and enforcers of rules they establish (Gillespie, 2018). Social platforms have created intricate, complex, and multi-layered content-moderation systems involving human-machine collaboration to process the exponential growth of online content (Myers West, 2018), including Chinese prevailing platforms (Chin, Park, & Li, 2022; Qiu & Dwyer, 2023).

On one hand, China retains an intrusive role in the platform society (De Kloet, Poell, Guohua, & Yiu Fai, 2019). The government imposes strict liability—referred to as "main-body responsibility" (Liu & Yang, 2022)—on Internet intermediaries to proactively prevent the circulation of illicit or unlawful content (MacKinnon, Hickok, Bar, & Lim, 2015; Yu, 2018). On the other hand, the Chinese state aspires to profit from platformization and transform into a supporter of Chinese digital companies. Although the parameters over sensitive issues remain indubitable, nascent space for concrete meaning-making incorporating multiple stakeholders has emerged (Cai & Dai, 2021; Schneider, 2018). To ensure its centrality is maintained, the Chinese government seeks to reinvent a state-led panoramic system encompassing multiple subjects via collaborative measures (Cai & Wang, 2022). Zhou and Liu (2024) defined this dual governance strategy that controls platforms via powerful regulation while simultaneously instrumentalizing them through benefit inducements as a "stick-carrot" approach.

Individuals contribute to moderation through two types of coping behaviors: restrictive and behavioral. Restrictive behaviors involve seeking protection from authorities (Lim, 2017), namely support for content moderation. Corrective actions involve individuals actively rectifying or challenging content, such as flagging a comment (Naab, Kalch, & Meitz, 2018).

Presumed Media Influence on Self and Others

By examining "flag" as a reactive response to harmful media content, we highlight users' perceptions of its effects on specific groups. It remains inconclusive which perceived target group of media effects is the most powerful predictor of user behavior. The third-person effect (TPE) and the influence of presumed influence (IPI) are the most applied and tested models (McLeod, Wise, & Perryman, 2017).

TPE was first proposed by Davison (1983), supposing that individuals tend to believe others are more affected by mediated messages than themselves—a phenomenon known as third-person perception (TPP)—and behave according to the self-other perceptual gap, often taking remedial actions to mitigate toxic effects Davison (1983). Building on TPE, the IPI describes the process by which people perceive media influence on others and then react to that perceived influence (Gunther & Storey, 2003). In other words,

presumed media influence on others (PMI3) may inspire attitudinal and behavioral responses, and presumed media influence on self (PMI1) may not necessarily generate consequences.

Because the above two indicators may miss part of the effect, another strand of research suggests that the additive index of perceived effects on self and others (PMI1 + PMI3), termed the diamond method or second-person perception, better captures the overall perceived media influences and serves as a powerful predictor of behavioral consequences (Neuwirth & Frederick, 2002; Riedl et al., 2022; Schmierbach, Boyle, Xu, & McLeod, 2011). In a highly interconnected platform environment, it becomes difficult to disentangle PMI1 from PMI3, since we survive in an interactive digital space where the boundary between publicity and privacy is blurry. The diamond method denotes mutual shared influence on the notion of "us" (influence on both themselves and others), implying common interest and the potential for social action (Neuwirth & Frederick, 2002). Though the diamond method was criticized for the simplification of reality (Baek, Kang, & Kim, 2019), our study's pinpoint was not to distinguish between third-person, first-person, and second-person effects. Instead, we attempt to probe into the underlying psychological mechanism ushering user perceptions toward behavioral responses. Hence, we employed the summative term of (PMI1 + PMI3) as an indicator of presumed media influence on both self and others, constructing a toxic conversational atmosphere as a shared concern.

Conventional media effect studies believe that the negative perception of media effect will stimulate individuals to tune-up behaviors aligned with these presumptions and mitigate baneful consequences (Gunther, Bolt, Borzekowski, Liebhart, & Dillard, 2006; Tal-Or, Tsfati, & Gunther, 2009). Users are entitled to be crucial participants in executing informal social control (Watson, Peng, & Lewis, 2019) on platforms. When individuals encounter harmful content, the more strongly they perceive its influence, the more motivated they are to either demand platform content moderation or flag the content in a decentralized manner to maintain a harmonious cyberspace.

Scholars have confirmed that presumed media influence is positively related to a range of restrictive and corrective behaviors in the domain of content moderation (Cheng & Chen, 2020; Riedl et al., 2022; Sun, Chia, Lu, & Oktavianus, 2022; Sun, Oktavianus, Wang, & Lu, 2022). We argue that when users show high presumed media influence on self and others, they are more likely to support content moderation and flag a comment to actively sanitize pernicious discourses. Based on prior studies, we propose the following:

H1: Presumed media influence on self and others (PMI1 + PMI3) is positively related to willingness to report harmful comments to the platform.

H2: Presumed media influence on self and others (PMI1 + PMI3) is positively related to support for content moderation.

Responsibility Attribution

When it comes to platform governance, one of the most intractable problems is "the problem of many hands" and how to allocate responsibility appropriately (Gorwa, 2019; Helberger et al., 2018). The platform undeniably holds significant accountability because it provides extensive techno-commercial

infrastructures that enable user participation (Gerlitz & Helmond, 2013), thereby shaping how users engage in content moderation and fulfill their responsibilities (Helberger et al., 2018).

Under the "main-body responsibility" (Liu & Yang, 2022) framework, the platform has found a solid moral and legitimate basis for privatized governance geared toward policing users. The platform administrators steer users on the frontline, and most violations regarding malicious content are handled internally by the platform. Since our study focuses on the antecedents of reporting harmful comments to the platform, we explore responsibility attribution to the platform.

Cheng, Wei, and Ge (2017) have found that individuals' perceptions of the adverse effects of city smog were positively related to responsibility attribution to the industrial enterprise. When individuals highly estimate the influence of undesirable content, it spurs an urgent expectation for others to shoulder responsibility, making them more likely to report harmful comments. Therefore, this study proposes that the greater the perceived media influence, the more users attribute responsibility.

H3: Presumed media influence on self and others is positively related to responsibility attribution to the platform.

Responsibility attribution has been documented as a mediator linking various antecedents, such as crisis severity and crisis type, to the ultimate responses of publics (Kim, 2014). Through an experiment, Jeong, Yum, and Hwang (2018) found that assigning more responsibility to the industry leads to stronger support for governmental punishment of the industry in the issue of smartphone addiction. Who should intervene and to what extent reveals users' attitudes toward content moderation and reflects their demands (Riedl, Naab, Masullo, Jost, & Ziegele, 2021), consequently shaping subsequent behaviors.

When individuals believe that the platform should account for negative consequences, they may demand that the company develop technical competence to curb the spread of harmful comments. Furthermore, they may be motivated to collaborate with the platform in detecting offensive content. Therefore, it is proposed that:

H4: Responsibility attribution is positively related to support for content moderation.

H5: Responsibility attribution is positively related to willingness to report harmful comments to the platform.

Restrictive and Corrective Actions

Extant research has often considered restrictive and corrective actions as two independent variables (Cheng & Chen, 2020; Golan & Lim, 2016; Naab, Naab, & Brandmeier, 2019). However, Lim (2017) demonstrated that individuals' support for regulating online advertising of cosmetic surgery (OACS) may increase their willingness to engage in corrective actions. "Flag" empowers peripheral users to fight for their values or social norms, and users' reporting behaviors have become a segment of content moderation.

Hence, we postulate that when an individual supports platform content moderation, he or she will be more active in reporting harmful comments to the platform.

H6: Support for platform content moderation is positively related to willingness to report harmful comments to the platform.

In an experiment conducted by Naab et al. (2018), responsibility attribution to professional moderators mediated the effect of response direction on flagging uncivil user comments. TPE or IPI has usually been tested as a direct effect in the literature (Baek et al., 2019; Cheng & Chen, 2020; Lee, 2021; Riedl et al., 2022), not recognizing that it is a complicated underlying mechanism. In this study, we consider the whole process and test the indirect influence of (PMI1+PMI3) on reporting responses via responsibility attribution and support for content moderation. Hence, we propose the following:

H7: The relationship between (PMI1+PMI3) and willingness to report harmful comments to the platform is mediated by responsibility attribution and support for content moderation.

Perceived Transparency of Flag Mechanism

Transparency has been a heated controversy in recent scholarly debates (Balkin, 2018; Klonick, 2017; Suzor, 2018). Normative policy research applauds the value of transparency as a vital component within a system for accountability, efficiency, and democracy (De Gregorio, 2020; Suzor, Dragiewicz, et al., 2019; Suzor, Van Geelen, & Myers West, 2018). Suzor (2018) argued that the transparency deficit in private governance stems from the obscure reasons released upon which moderation teams make and review decisions. Suzor, Myers West, et al. (2019) appealed for greater and meaningful transparency. Another stream of research reflects on the limitations of transparency against the backdrop of the "black-box" algorithm. Ananny and Crawford (2018) interrogated the undue optimism around the ideal of transparency and pointed out its inadequacy in understanding algorithmic systems.

In the conventional context, transparency is the degree to which an organization provides information about its decisions and working procedures to those not directly involved (Florini, 2007; Rawlins, 2008). In algorithmic systems, transparency can be defined as the extent to which the inner process of generating results is disclosed and explained. In complex technical environments, transparency involves subjective dimensions in the eye of the beholder (Shin & Park, 2019). The information provided by the platform may not be available or interpretable for users, in concert with a shift in focus from sender to receiver (Rasmussen, 1991; Williams, 2005).

Following the user-centric strand, this study contextualizes transparency from the stakeholder's perspective, termed perceived transparency, which describes the direct perception of an algorithm's utility and properties through its performance (Shin & Park, 2019). Perceived transparency of the flag mechanism refers to the extent to which users feel they understand the criteria and processes involved and why their requests are reviewed.

Scholars have related perceived transparency to the use of intention. In light of uncertainty reduction (Liu, 2021), high evaluation of controllability (Rader, Cotter, & Cho, 2018), trust and satisfaction enhancement (Lee, 2018; Shin, 2020; Sundar, 2020), users who understand how algorithm works are more likely to utilize the system properly (Lee & Boynton, 2017; Shin, Zhong, & Biocca, 2020).

The flag mechanism has been criticized for its opaqueness, which undermines its legitimacy and decreases users' volunteers (Common, 2020; Crawford & Gillespie, 2016). Users are more likely to flag uncivil comments when they are provided ample information about the meaning and strategy of the flagging button (Naab et al., 2018). Hence, we postulate that higher perceived transparency leads to a stronger intention to use the flag mechanism.

H8: Perceived transparency is positively associated with willingness to report harmful comments to the platform.

Previous research has attempted to examine the moderating role of affordance in media practices. "Affordance" here refers to the actual and perceived properties embedded in a technology that enable people to use it, as well as the actual practices through which people interact with it (Ellison & Vitak, 2015). The interaction between the platform and users can be seen as a social dynamic through which the agency of users exerts a reacting force on the technological structure. Some studies have found that platform affordances affect the forms of political participation (Halpern, Valenzuela, & Katz, 2017) or political campaign strategy adopted (Bossetta, 2018).

Although very few studies consider transparency, which is more inclined to value evaluation, as a kind of algorithmic affordance (Shin & Park, 2019; Shin et al., 2020), we applied this concept to explore how user perceptions related to platform technical configurations moderate their behavior. In line with this strand of research, we propose that perceived transparency magnifies the effect of antecedent variables on users' willingness to report.

H9a: Perceived transparency positively moderates the relationship between (PMI1 + PMI3) and willingness to report harmful comments to the platform.

H9b: Perceived transparency positively moderates the relationship between responsibility attribution and willingness to report harmful comments to the platform.

H9c: Perceived transparency positively moderates the relationship between support for content moderation and willingness to report harmful comments to the platform.

Figure 1 delineates the research model developed in the current study.

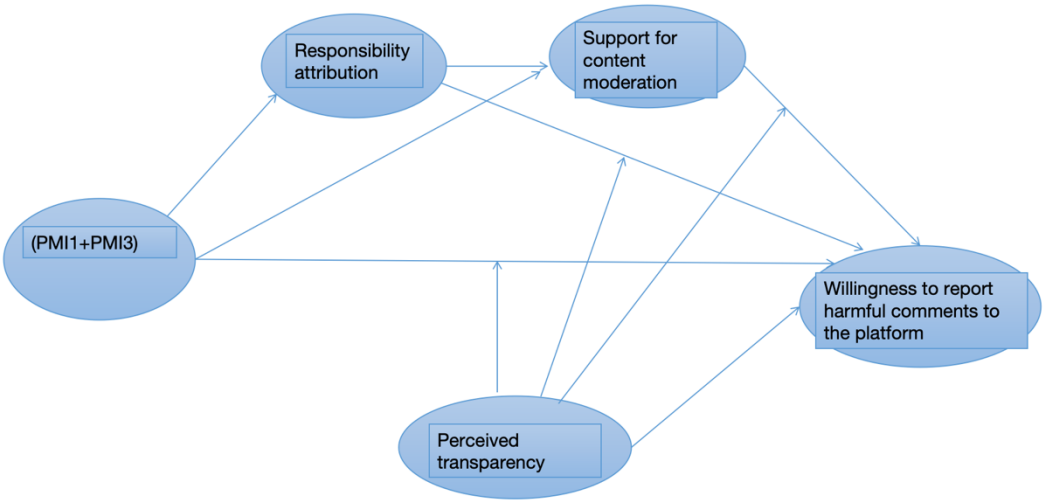

Figure 1. Illustrates the model we proposed in this study.

Methods

Procedures

In March 2021, we used the global market research and public opinion specialist, IPSOS (China), to collect data from 3,000 individuals in China. To ensure an accurate representation of Chinese netizens, the target population of participants tallied with the structure of netizens in the official report released by the China Internet Network Information Centre (CNNIC, 2020) through quota sampling. The sample included 49% females, and the mean age of the interviewees was 32.52 years (*SD* = 11.42). The difference between the target and actual proportions was controlled at 1.5%. The sample size was 3,000 (*N* = 3000). Anonymity, privacy, and the voluntary nature of the survey were guaranteed.

Variables and Measures

Presumed Media Influence on Self and Others (PMI1+PMI3)

Presumed media influence on self and others was operationalized using a Likert-type scale ranging from 1 (not at all) to 7 (very much), prompting respondents to rate presumed media influence on the third people and themselves of five types of harmful online comments, then adding them together. The 7-point Likert scale applies to all variables below.

These comments included vulgarity, insulting, inciting violence, hate speech, and rumors. We used five specific examples instead of definitions, and we asked respondents to categorize these examples to

ensure that they could be identified as a specific type in the pre-survey. These were averaged into an index. (α = .934)

Responsibility Attribution (RA)

Following Riedl et al. (2021), respondents estimated the extent to which they agreed that the platform was responsible for taking action against harmful comments. This variable was operationalized using a single item.

Support for Content Moderation (SCM)

Support for content moderation was measured by using and adapting scales from prior scales (Guo & Johnson, 2020; Naab et al., 2019; Wang & Kim, 2020). Participants were asked to appraise their agreement. (α = .881)

Willingness to Report Harmful Comments to the Platform (RTP)

To measure willingness to report harmful comments to the platform, three items were adapted from prior scales (Guo & Johnson, 2020; Wang & Kim, 2020). (α = .829).

Perceived Transparency (PT)

Based on the definition and existing measures of perceived transparency (Gonçalves, Weber, Masullo, Torres da Silva, & Hofhuis, 2023), we developed an index measuring perceived transparency of reporting mechanism, gauging how much information that users are informed about the platform's decision about reporting requests.

This variable was operationalized on a Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree), asking respondents to evaluate five statements. (α = .870) Confirmatory Factor Analysis (CFA) showed that RMSEA was 0.0432 (< 0.05), CFL was 0.996 (> 0.9), and TLI was 0.992 (> 0.9).

Control Variables

We sought to control for three relevant demographic variables: gender, age, education, and frequency of social media use (SMU). 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 frequency of SMU was coded as 1–7, representing less than 10 minutes, 10–30 minutes, 31–60 minutes, 1–2 hours, 2–3 hours, more than 3 hours, and more than 6 hours per day, respectively.

Results

SPSS 26 and PROCESS plug-ins were used for data analysis. The statistical analyses were divided into three main parts. First, descriptive statistics and correlations were calculated by computers. Then, we

used PROCESS advanced by Hayes (2017), selecting Model 6 to test the mediation path. Third, to examine the moderating role of perceived transparency, we executed the PROCESS macro (Model 89) developed by Hayes (2017). In addition, to analyze the indirect relations, we adopted the bootstrapping method (Hayes & Scharkow, 2013), with 95% bias-corrected confidence intervals (CI) from 5,000 resamples of the data.

Table 1 shows that the variables were significantly correlated. Regarding H1, H2, H3, H4, and H5, Table 2 shows regression analysis results that examined direct relation. We first proposed that (PMI1+PMI3) was positively related to willingness to report (H1) and this hypothesis was supported (b = .1783, $p < .001$), was positively related to support for content moderation (H2) and this hypothesis was supported (b = .0804, $p < .001$), was positively related to responsibility attribution (H3) and this hypothesis was verified (b = .2043, $p < .001$). Concerning H4 and H5, it was demonstrated that responsibility attribution was positively related to support for content moderation (b = .3923, $p < .001$) and willingness to report harmful comments to the platform (b = .2205, $p < .001$). As individuals attribute more responsibility to the platform, their support for content moderation and inclination to report harmful comments are strengthened. For H6, the results demonstrated that there existed a positive association between support for content moderation and willingness to report harmful comments (b = .2181, $p < .001$).

Table 1. Descriptive Statistics and Correlations.

	M	*SD*	1	2	3	4	5	6	7	8	9
1. Gender	-	-		.135**	-.125**	-.105**	-.063**	-.054**	-.088**	.000	.047*
2. Age	32.52	11.422	.		-.220**	-.074**	.000	.005	-.096**	-.026	-.070**
3. Edu	3.502	.984				.155**	.061**	.203**	.274**	.037*	.040*
4. SMU	4.50	1.373					.064**	.138**	.113**	.095**	.046*
5. PMI1+PMI3	10.331	2.694						.409**	.413**	.517**	.372**
6. RA	5.52	1.404							.613**	.468**	.372**
7. SCM	5.694	1.097								.444**	.432**
8. RTP	4.982	1.349									.545**
9. PT	4.967	1.202									

Note. N = 3,000. PMI = presumed media influence, RA = responsibility attribution, SCM = support for content moderation, RTP = willingness to report comments to the platform, PT = perceived transparency, SMU = frequency of social media use. *p <. 05. **p< .01. ***p< .001.

Table 2. Model Summary of Regression Results Predicting Willingness to Report Harmful Comments to the Platform.

	M_1 RA		M_2 SCM		RTP	
Predictors	*Coefficient*	*SE*	*Coefficient*	*SE*	Coefficient	*SE*
PMI1+PMI3	.2043***	.0085	.0804***	.1089	.1783***	.0082
M_1 RA			.3923***	.0121	.2205***	.0182
M_2 SCM					.2181***	.0237
Gender	-.0124	.0465	-.0480	.0307	.1293**	.0399
Age	.0063**	.0021	-.0062***	.0014	-.0040*	.0018
Edu	.2516***	.0241	.1588***	.0162	-.1198***	.0214
SMU	.0906***	.0169	.0016	.0112	.0363*	.0146
Constant	1.9188***	.1613	2.3607***	.1089	1.0023***	.1520
R^2	0.2087		0.4364		0.3716	
F	157.8871***		386.1785***		252.7597***	

Note. N = 3,000. Each column is a regression model that predicts the variable at the top of the column. PMI = presumed media influence, RA = responsibility attribution, SCM = support for content moderation, RTP = willingness to report comments to the platform, PT = perceived transparency, SMU = frequency of social media use.
*p < .05. **p < .01. ***p < .001.

To test serial mediation through responsibility attribution and support for content moderation (H7), a series of mediation analyses were conducted, provided by Hayes (2017). According to the results illustrated in Table 3, the serial mediation between (PMI1+PMI3) and willingness to report through responsibility attribution and support for content moderation was significant (b = .0175, *SE* = .0027, CI = [0.0123, 0.0231]). This suggested that when individuals presumed more influence on self and others, they were more likely to attribute responsibility to the platform, which heightened support for content moderation and eventually promoted willingness to report harmful comments. Respectively, the mediating relations between (PMI1+PMI3) on willingness to report through responsibility attribution (b = .0450, *SE* = 0.0054, CI = [0.0345, 0.0556]) and support for content moderation (b = .0175, *SE* = .0030, CI = [0.0120, 0.0237]) were also significant.

Table 3. Indirect Relations Between (PMI1+PMI3) and Willingness to Report Harmful Comments to the Platform Through the Mediators.

Mediation path	95% CI			
	Coefficient	*SE*	Lower	Upper
(PMI1+PMI3)→RA→RTP	.0450	.0054	.0345	.0556
(PMI1+PMI3)→SCM→RTP	.0175	.0030	.0120	.0237
(PMI1+PM13)→RA→SCM→RTP	.0175	.0027	.0123	.0231

Note. N = 3,000. PMI= presumed media influence, RA = responsibility attribution, SCM = support for content moderation, RTP = willingness to report comments to the platform. *SE* = standard error; CI = confidence interval.

H8 examined the positive relations between perceived transparency and willingness to report harmful comments. Results demonstrated that perceived transparency was a positive and significant direct predictor of willingness to report harmful comments to the platform (b = .3792, $p < .001$). As the levels of perceived transparency increased, it was more probable that platform users reported harmful comments to the platform.

H9s postulated perceived transparency moderated the direct relations among (PMI1+PMI3) (H9a), responsibility attribution (H9b), and support for content moderation (H9c) on willingness to report harmful comments. Table 4 shows the results using the PROCESS macro model 89 proposed by Hayes (2017). Perceived transparency moderated the direct relations between responsibility attribution and willingness to report harmful comments, support for content moderation, and willingness to report harmful comments (H8b: b = .0724, *SE* = .0125, $p < .001$; H8c: b = −.0603, *SE* = .0162, $p < .001$), while the interaction effect of (PMI1+PMI3) and perceived transparency was not significant ($p = .3252$). To decompose the interaction effects, we plotted the slope, which illustrates the direct relationships for different levels of perceived transparency. As Figure 2 shows, the gradient showing the correlation between responsibility attribution and willingness to report harmful comments with low perceived transparency was weaker (B = 0.108, t =2.933, $p < 0.01$), while the slope with high perceived transparency was relatively strong (B = 0.282, t = 5.362, $p < 0.001$). As for the direct relationships between support for content moderation and willingness to report harmful comments, the direct relations were significant and positive when perceived transparency was lower (B = 0.154, t = 4.837, $p < 0.001$), while insignificant when perceived transparency was higher (B = 0.009, t = 0.188, $p = 0.851$).

Table 4. Testing the Moderated Mediation Effect of Perceived Transparency.

	RTP	
Predictors	*Coefficient*	*SE*
PMI1+PMI3	.1670***	.0278
M_1 RA	-.1642**	.0617
M_2 SCM	.3812***	.0762
PT	.3792***	.0782
(PMI1+PMI3)*PT	-.0053	.0054
M_1*PT	.0724***	.0125
M_2 *PT	-.0603***	.0162
Gender	.0465	.0372
Age	-.0008	.0017
Edu	-.0795***	.0199
SMU	.0391**	.0135
Constant	.2041	.3470
R^2	0.4603	
F	231.6424***	

Note. N = 3,000. Each column is a regression model that predicts the variable at the top of the column. PMI = presumed media influence, RA = responsibility attribution, SCM = support for content moderation, RTP = willingness to report comments to the platform, PT = perceived transparency, SMU = frequency of social media use.
*$p < .05$. **$p < .01$. ***$p < .001$.

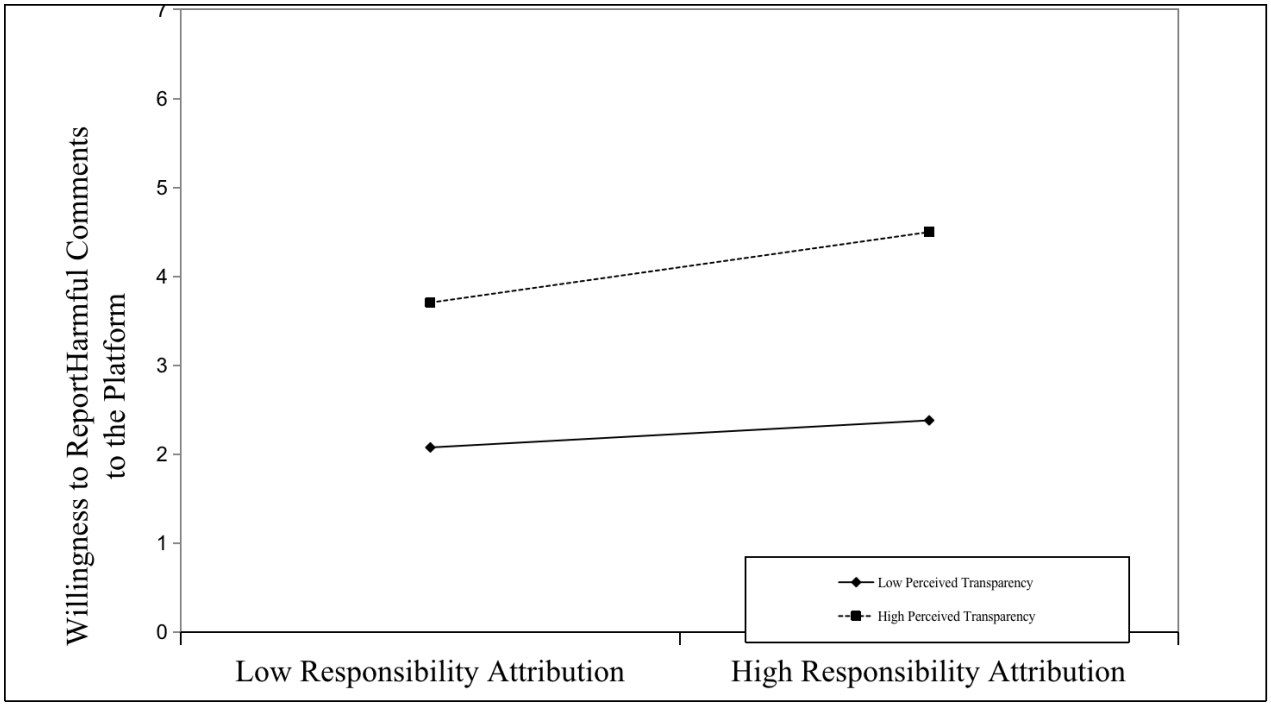

Figure 2. Interaction effect of responsibility attribution and perceived transparency in willingness to report harmful comments to the platform. High and low levels of responsibility attribution and perceived transparency represent one standard deviation above and below the mean, respectively.

Discussion and Implication

The study demonstrates that presumed media influence on self and others (PMI1+PMI3) induces responsibility attribution to the platform, which increases both support for content moderation (restrictive actions) and the likelihood of reporting harmful comments to the platform (corrective actions). Moreover, perceived transparency plays a moderating role in the process of (PMI1+PMI3), affecting responsibility attribution, support for content moderation, and reporting harmful comments to the platform.

First, the findings align with previous arguments that second-person perceptions (PMI1+PMI3) are strong predictors of support for content moderation (Riedl et al., 2022), and extend this relationship to include corrective actions, specifically reporting to the platform. There remains a paucity of research examining the causal mechanism by which individuals' presumptions about the impact of harmful comments align with their behaviors (Wang & Kim, 2020). To address this scholarly gap, we extend a comprehensive model that elucidates users' socio-psychological mechanism, driving public responses to harmful comments culturally and contextually.

Second, our study could be seen as the latest addition to the mediating role that responsibility attribution plays in motivating publics, advancing research about cooperative responsibility in platform

governance (Helberger et al., 2018). The Chinese government emphasizes that platform developers should regulate deviant comments to fulfill corporate social responsibility. The government subcontracts the regulation to the platforms as administrative governmentality, resulting in the platform becoming an intermediary between the government and users (Shao, Guan, Sun, Cole, & Liu, 2022). These results mirror official provisions and public appetite for platforms to protect vulnerable end users. Prior scholars have differentiated between restrictive and corrective behaviors as two parallel dependent variables (Naab et al., 2019). Nevertheless, our study implies that restrictive behaviors can be considered a catalyst for corrective behaviors, but not a prerequisite. We gain deeper insight into how users allocate responsibility between them and the platform. Compared to more expressive behaviors, flagging is less aggressive and exhausting, as the responsibility assigned to the platform and professional moderators immediately after users submit their claims (Porten-Cheé, Kunst, & Emmer, 2020). Users indeed downplay their engagement to limited effects, deeming that they only have an auxiliary obligation to alert the platform (Naab et al., 2018).

Third, our research examines the moderating effect of perceived transparency, providing evidence substantiating the positive role of transparency in content moderation. Substantive and valuable information allows users to grab a sense of control and predictability (Lind & Tyler, 1988) for decisions without requiring high cognitive efforts to process them (ter Hoeven, Stohl, Leonardi, & Stohl, 2021). Even if they fail to flag a comment temporarily, useful information helps align with social norms (Jhaver, Bruckman, & Gilbert, 2019), improve strategies for attaining valued outcomes. Based on social exchange theory, accurate and positive anticipation of outcomes reinforces reciprocity (Blau, 2017), ensuring users that their actions will result in favorable returns.

The incorporation of transparency advances existing literature in three ways: (1) Prior research about transparency has been mainly conducted under punitive paradigms in which users are "offender" of community norms. A series of findings foreground a similar contention that transparency alleviates users' confusion about sanctions and reduces future violations (Jhaver, Appling, et al., 2019; Juneja, Rama Subramanian, & Mitra, 2020; Seering, Kraut, & Dabbish, 2017). Our study takes a step further and proves that transparency also promotes users' voluntary self-moderating for common good orientation (Friess, Ziegele, & Heinbach, 2021), prompting them to become empowered online citizens. Considering the research trend of folk theories of sociotechnical systems in the HCI field (DeVito, Gergle, & Birnholtz, 2017; Jhaver, Appling, et al., 2019), we also add to this emerging literature by exploring how users' perspectives of "transparency" affect their interaction with the system in the domain of content moderation. (2) Studies exploring antecedents of reporting content predominantly focused on presumed effects of media content (Wang & Kim, 2020), or on users' personalities (Wilhelm et al., 2019), ignoring the fact that users interact with online content amid a set of sociotechnical assemblages configured by the platform (Gillespie, 2018). On one hand, the "flag" tool allows users to intervene in content moderation to align with their tastes and preferences. On the other hand, the identification and removal of inappropriate content hinges on platforms. "Flag" implies a collaboration synthesizing personal moderation enacted by decentralized end users and coercive measures enforced by platforms (Jhaver & Zhang, 2025). Given this, this article enriches academic perspective by considering perceptions surrounding media content and perceived transparency of governance mechanisms. (3) Albu and Flyverbom (2016) articulated two approaches to organizational transparency, conceptualized either as a matter of information disclosure or as a social process. The positive moderating role of transparency implies a potential path connecting the two streams of transparency.

Available, relevant, and understandable information conveyed to audiences is conducive to users' reporting behaviors, while increased reports function as "silent" expressions of dissatisfaction with particular content. Moderators' decisions and corresponding explanations are a means through which platforms can communicate to other stakeholders publicly, justify their arbitration, and integrate diverse interests. Transparency of the "flag" surfaces as a social-material circular process for tension and negotiation between platform and users over the boundaries of appropriateness, covering interaction between users, platforms, governments, and algorithms, as well as complexities of social dynamics (Crawford & Gillespie, 2016), echoing with appeals for more relational significance attached to transparency (Ananny & Crawford, 2018).

The current study has practical implications for platform operators and the government. Wang and Ge (2023) contended that the "report" mechanism in Chinese social platforms has descended into a fans' combating tool to silence rivals and intensify censorship. Inspiring healthy participation and decreasing malicious reporting have become pressing issues in platform governance. Because we identified responsibility attribution to the platform and support for content moderation as two mediating variables, it was necessary to highlight the responsibility of the platform in content governance. In addition to releasing periodical governance reports, major platforms can hold regular press conferences to announce achievements, rendering their duties more salient and contributing to public consciousness. Considering the inadequate governance resources grasped by the government, the Chinese government can put more effort into constructing benchmarks with which to supervise the way platforms comply with their obligations (Helberger et al., 2018), such as developing a meta-criterion for platform content moderation within multi-stakeholder frameworks. Walking through dominating social platforms in China, the feedback given to reporters is almost vague and generic, such as, "According to the relevant regulations, your report was not approved." Given the role of transparency, the platform should develop strategic designs to motivate the public to become involved in collaborative moderation and nurture online citizenship. The platform should explicitly inform the norms and implementation of the flag mechanism. Users also need specific explanations for why their reports are approved or declined, and why the content they deem harmful falls within or outside the scope defined by the platform. The platform's transparency reports can include more detailed data, such as the disposal rate of users' reports, the most reported content types, and the most dealt with content types. In addition, enterprises can initiate typical case databases to unify the standards of harmful comments.

Limitations and Future Research

There are several limitations that should be mentioned and discussed. First, our study did not consider how others' responses to harmful comments influence users' appraisal of media effects. Future research could examine whether refuting comments affect users' perceptions. Second, the control variables we employed were relatively limited. Later studies should consider other variables more comprehensively. Moreover, our study did not focus on a specific platform, so further research could select multiplatform for comparison to gain deeper insight. Finally, the significant correlations proved in this study cannot be understood as causality. In the future, we look forward to further developing the lines of investigation initiated in this research.

References

Albu, O. B., & Flyverbom, M. (2016). Organizational transparency: Conceptualizations, conditions, and consequences. *Business & Society, 58*(2), 268–297. doi:10.1177/0007650316659851

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989. doi:10.1177/1461444816676645

Baek, Y. M., Kang, H., & Kim, S. (2019). Fake news should be regulated because it influences both "others" and "me": How and why the influence of presumed influence model should be extended. *Mass Communication and Society, 22*(3), 301–323. doi:10.1080/15205436.2018.1562076

Balkin, J. M. (2018). Free speech is a triangle. *Columbia Law Review, 118*(7), 2011–2056. Retrieved from https://www.jstor.org/stable/10.2307/26524953

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. doi:10.1093/jcmc/zmab007

Blau, P. (2017). *Exchange and power in social life*. New York, NY: Routledge.

Bossetta, M. (2018). The digital architectures of social media: Comparing political campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. *Journalism & Mass Communication Quarterly, 95*(2), 471–496. doi:10.1177/1077699018763307

Cai, C., & Dai, L. (2021). Evolution of internet governance in China: Actors and paradigms. *China Quarterly of International Strategic Studies, 7*(1), 79–109. doi:10.1142/S2377740021500020

Cai, C., & Wang, T. (2022). Moving toward a "middle ground"?—The governance of platforms in the United States and China. *Policy & Internet, 14*(2), 243–262. doi:10.1002/poi3.303

Cheng, P., Wei, J., & Ge, Y. (2017). Who should be blamed? The attribution of responsibility for a city smog event in China. *Natural Hazards, 85*(2), 669–689. doi:10.1007/s11069-016-2597-1

Cheng, Y., & Chen, Z. F. (2020). The influence of presumed fake news influence: Examining public support for corporate corrective response, media literacy interventions, and governmental regulation. *Mass Communication and Society, 23*(5), 705–729. doi:10.1080/15205436.2020.1750656

Chin, Y. C., Park, A., & Li, K. (2022). A comparative study on false information governance in Chinese and American social media platforms. *Policy & Internet, 14*(2), 263–283. doi:10.1002/poi3.301

China Internet Network Information Centre. (2020). *46th statistical report on Internet development in China*. Retrieved from https://www.cnnic.net.cn/NMediaFile/old_attach/P020210205509651950014.pdf

Common, M. F. (2020). Fear the reaper: How content moderation rules are enforced on social media. *International Review of Law, Computers & Technology, 34*(2), 126–152. doi:10.1080/13600869.2020.1733762

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. doi:10.1177/1461444814543163

Davison, W. P. (1983). The third-person effect in communication. *The Public Opinion Quarterly, 47*(1), 1–15. Retrieved from http://www.jstor.org/stable/2748702

De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. *Computer Law & Security Review, 36*(4), 1–17. doi:10.1016/j.clsr.2019.105374

De Kloet, J., Poell, T., Guohua, Z., & Yiu Fai, C. (2019). The platformization of Chinese society: Infrastructure, governance, and practice. *Chinese Journal of Communication, 12*(3), 249–256. doi:10.1080/17544750.2019.1644008

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017). "Algorithms ruin everything" # RIPTwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3163–3174). Denver, CO: CHI.

Einwiller, S. A., & Kim, S. (2020). How online content providers moderate user-generated content to prevent harmful online communication: An analysis of policies and their implementation. *Policy & Internet, 12*(2), 184–206. doi:10.1002/poi3.239

Ellison, N. B., & Vitak, J. (2015). Social network site affordances and their relationship to social capital processes. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology* (pp. 203–227). Chichester, UK: John Wiley & Sons, Incorporated.

Florini, A. (2007). *The right to know: Transparency for an open world*. New York, NY: Columbia University Press.

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication, 38*(5), 624–646. doi:10.1080/10584609.2020.1830322

Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society, 15*(8), 1348–1365. doi:10.1177/1461444812472322

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, CT: Yale University Press.

Golan, G. J., & Lim, J. S. (2016). Third-person effect of ISIS's recruitment propaganda: Online political self-efficacy and social media activism. *International Journal of Communication, 10*, 4681–4701. Retrieved from https://ijoc.org/index.php/ijoc/article/view/5551/1792

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. doi:10.1177/14614448211032310

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society, 22*(6), 854–871. doi:10.1080/1369118X.2019.1573914

Gunther, A. C., Bolt, D., Borzekowski, D. L. G., Liebhart, J. L., & Dillard, J. P. (2006). Presumed influence on peer norms: How mass media indirectly affect adolescent smoking. *Journal of Communication, 56*(1), 52–68. doi:10.1111/j.1460-2466.2006.00002.x

Gunther, A. C., & Storey, J. D. (2003). The influence of presumed influence. *Journal of Communication, 53*(2), 199–215. doi: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), 1–12. doi:10.1177/2056305120923003

Halpern, D., Valenzuela, S., & Katz, J. E. (2017). We Face, I Tweet: How different social media influence political participation through collective and internal efficacy. *Journal of Computer-Mediated Communication, 22*(6), 320–336. doi:10.1111/jcc4.12198

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

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science, 24*(10), 1918–1927. doi:10.1177/0956797613480187

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

Jeong, S.-H., Yum, J., & Hwang, Y. (2018). Effects of media attributions on responsibility judgments and policy opinions. *Mass Communication and Society, 21*(1), 24–49. doi:10.1080/15205436.2017.1362002

Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019). "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. doi:10.1145/3359294

Jhaver, S., Bruckman, A., & Gilbert, E. (2019). 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. doi:10.1145/3359252

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. doi:10.1177/14614448231217993

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. doi: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. doi:10.5771/2192-4007-2017-4-395

Kim, S. (2014). What's worse in times of product-harm crisis? Negative corporate ability or negative CSR reputation? *Journal of Business Ethics, 123*(1), 157–170. doi:10.1007/s10551-013-1808-x

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review, 131*, 1598–1670. Retrieved from https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/

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. doi:10.1080/19331681.2020.1871149

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 1–16. doi:10.1177/2053951718756684

Lee, T. (2021). How people perceive influence of fake news and why it matters. *Communication Quarterly, 69*(4), 431–453. doi:10.1080/01463373.2021.1954677

Lee, T. H., & Boynton, L. A. (2017). Conceptualizing transparency: Propositions for the integration of situational factors and stakeholders' perspectives. *Public Relations Inquiry, 6*(3), 233–251. doi:10.1177/2046147X17694937

Lim, J. S. (2017). The third-person effect of online advertising of cosmetic surgery: A path model for predicting restrictive versus corrective actions. *Journalism & Mass Communication Quarterly, 94*(4), 972–993. doi:10.1177/1077699016687722

Lind, E. A., & Tyler, T. R. (1988). *The social psychology of procedural justice*. New York, NY: Springer US.

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. doi:10.1093/jcmc/zmab013

Liu, J., & Yang, L. (2022). "Dual-Track" platform governance on content: A comparative study between China and United States. *Policy & Internet, 14*(2), 304–323. doi:10.1002/poi3.307

MacKinnon, R., Hickok, E., Bar, A., & Lim, H.-i. (2015). *Fostering freedom online: The role of Internet intermediaries*. Reston, VA: UNESCO Publishing.

McLeod, D. M., Wise, D., & Perryman, M. (2017). Thinking about the media: A review of theory and research on media perceptions, media effects perceptions, and their consequences. *Review of Communication Research, 5*, 35–83. doi:10.12840/issn.2255-4165.2017.05.01.013

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society, 20*(11), 4366–4383. doi: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. doi:10.1177/1461444816670923

Naab, T. K., Naab, T., & Brandmeier, J. (2019). Uncivil user comments increase users' intention to engage in corrective actions and their support for authoritative restrictive actions. *Journalism & Mass Communication Quarterly, 98*(2), 566–588. doi:10.1177/1077699019886586

Neuwirth, K., & Frederick, E. (2002). Extending the framework of third-, first-, and second-person effects. *Mass Communication & Society, 5*(2), 113–140. doi:10.1207/S15327825MCS0502_2

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. Retrieved from https://ijoc.org/index.php/ijoc/article/view/10639/2934

Qiu, Y., & Dwyer, T. (2023). Regulating Zhibo in China: Exploring multiple levels of self-regulation and stakeholder dynamics. *Policy & Internet, 15*(2), 266–282. doi:10.1002/poi3.337

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Montreal, Canada: ACM.

Rasmussen, R. (1991). A communication model based on the conduit metaphor: What do we know and what do we take for granted? *Management Communication Quarterly, 4*(3), 363–374. doi:10.1177/0893318991004003005

Rawlins, B. (2008). Give the emperor a mirror: Toward developing a stakeholder measurement of organizational transparency. *Journal of Public Relations Research, 21*(1), 71–99. doi:10.1080/10627260802153421

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. doi:10.1002/poi3.257

Riedl, M. J., Whipple, K. N., & Wallace, R. (2022). Antecedents of support for social media content moderation and platform regulation: The role of presumed effects on self and others. *Information, Communication & Society, 25*(11), 1632–1649. doi:10.1080/1369118X.2021.1874040

Schmierbach, M., Boyle, M. P., Xu, Q., & McLeod, D. M. (2011). Exploring third-person differences between gamers and nongamers. *Journal of Communication, 61*(2), 307–327. doi:10.1111/j.1460-2466.2011.01541.x

Schneider, F. (2018). *China's digital nationalism*. Oxford, UK: Oxford University Press.

Seering, J., Kraut, R., & Dabbish, L. (2017). Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 111–125). New York, NY: ACM.

Shao, C. A., Guan, X., Sun, J., Cole, M., & Liu, G. (2022). Social media interactions between government and the public: A Chinese case study of government WeChat official accounts on information related to COVID-19. *Frontiers in Psychology, 13*(9), 1–23. doi:10.3389/fpsyg.2022.955376

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. doi: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*, 277–284. doi: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), 1–11. doi:10.1016/j.ijinfomgt.2019.102061

Sun, Y., Chia, S. C., Lu, F., & Oktavianus, J. (2022). The battle is on: Factors that motivate people to combat anti-vaccine misinformation. *Health Communication, 37*(3), 327–336. doi:10.1080/10410236.2020.1838108

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. doi: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. doi: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), 1–11. doi:10.1177/2056305118787812

Suzor, N. P., Dragiewicz, M., Harris, B., Gillett, R., Burgess, J., & Van Geelen, T. (2019). Human rights by design: The responsibilities of social media platforms to address gender-based violence online. *Policy & Internet, 11*(1), 84–103. doi:10.1002/poi3.185

Suzor, N. P., Myers West, S., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication, 13*, 1526–1543. Retrieved from https://ijoc.org/index.php/ijoc/article/view/9736/2610

Suzor, N. P., Van Geelen, T., & Myers West, S. (2018). Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette, 80*(4), 385–400. doi:10.1177/1748048518757142

Tal-Or, N., Tsfati, Y., & Gunther, A. C. (2009). The influence of presumed media influence: Origins and implications of the third-person perception. In R. L. Nabi & M. B. Oliver (Eds.), *The SAGE handbook of media processes and effects* (pp. 99–112). New Delhi, India: Sage Publications.

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. doi:10.1177/0093650219877093

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

Wang, E. N., & Ge, L. (2023). Fan conflicts and state power in China: Internalised heteronormativity, censorship sensibilities, and fandom police. *Asian Studies Review, 47*(2), 355–373. doi:10.1080/10357823.2022.2112655

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. doi:10.1080/08838151.2020.1757368

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. doi: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. doi:10.1177/0093650219855330

Williams, C. C. (2005). Trust diffusion: The effect of interpersonal trust on structure, function, and organizational transparency. *Business & Society, 44*(3), 357–368. doi:10.1177/0007650305275299

Xie, X., Shi, L., & Zhu, Y. (2023). Why netizens report harmful content online: A moderated mediation model. *International Journal of Communication, 17*, 5830–5851. Retrieved from https://ijoc.org/index.php/ijoc/article/view/18919/4327

Yu, W. (2018). Internet intermediaries' liability for online illegal hate speech. *Frontiers of Law in China, 13*(3), 342–356. doi:10.3868/s050-007-018-0026-5

Zhou, M., & Liu, S. D. (2024). Regulating *tuwei* culture and migrant youth through Kuaishou's platform governance. *Policy & Internet, 16*(1), 104–120. doi:10.1002/poi3.366

Appendix List of Items

Perceived Media Influence on Self and Others (PMI)

Rate perceived media influence on self and others of following online harmful comments. (1 = not at all, 7 = very much)

(a) vulgarity "A slut who wants to keep fame."
(b) insulting "You are an idiot."
(c) inciting violence "I'll have you stabbed to death."
(d) hate speech and discrimination "Muslims get out of China."
(e) rumors "Isatis root (a kind of Chinese medicine) can cure COVID-19."

Responsibility Attribution (RA)

How much do you think the platform is responsible for taking actions against harmful comments like you have read just now? (1= strongly disagree, 7 = strongly agree)

Support for Content Moderation (SCM)

How much do you agree with following statements? (1= strongly disagree, 7 = strongly agree)

(a) Harmful online comments should be prohibited by the platform.
(b) Online comments should be regulated by the platform.
(c) Publishers of harmful comments should be punished by the platform.
(d) I support the platform taking action to censor harmful comments.
(e) I support the platform closing the accounts of individuals or groups who post harmful comments.

Willingness to Report Harmful Comments to the Platform (RTP)

How much do you agree with the following statements? (1= strongly disagree, 7 = strongly agree)

(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 submit a complaint to the platform regarding this comment.

Perceived Transparency (PT)

How much do you agree with the following statements? (1= strongly disagree, 7 = strongly agree)

(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 content is easier to report successfully.
(e) If my comments are deleted or blocked, I know why.