International Journal of Communication 17(2023), 5830–5851 1932–8036/20230005

Why Netizens Report Harmful Content Online: A Moderated Mediation Model

XINZHOU XIE♦[1]
Peking University, China

LIN SHI
University of Chinese Academy of Social Sciences, China

YAOYING ZHU
Tsinghua University, China

Internet content governance appeals to the systematic coordination of multiple social actors. In addition to government regulation and platform governance, people's initiatives and participation have become increasingly critical. Based on cognitive dissonance theory, our article investigates the psychological mechanism of people's behavioral intentions to report harmful content online. A cross-sectional survey involving 3,000 Chinese netizens was conducted. The empirical results showed that the perceived necessity of Internet content governance was positively related to behavioral intention to report harmful content online through the mediation of aversion to harmful Internet content and that information-seeking motive negatively moderated the relationship between perceived necessity and aversion. This moderated mediation model provides theoretical and practical implications for understanding and playing the role of netizens in Internet content governance.

Keywords: Internet content governance, content reporting, harmful content, cognitive dissonance, usage motive

The Internet has produced and delivered rich and varied content in texts, photos, audio, and video forms, becoming crucial to knowing and understanding the world. On the Internet, people can acquire knowledge from news or online courses, seek pastime through movies or games, and achieve personal value through self-expression or self-presentation. However, under such an open architecture that promotes the

Xinzhou Xie: xzxie@pku.edu.cn
Lin Shi: shilin950116@163.com
Yaoying Zhu (corresponding author): zyy32@126.com
Date submitted: 2021-11-03

[1] Xinzhou Xie and Lin Shi contributed equally to this work and shared first authorship. The authors appreciate the feedback and suggestions from Larry Gross, Kady Bell-Garcia, and the anonymous reviewers. This research was supported by the National Social Science Fund of China (Grant No. 18ZDA317).

rapid dissemination of unverified information (Lazer et al., 2018), people may also encounter false information, fake news, pornographic content, violent content, and the like, the amount of which has been increasing with the ease of generating content online and the anonymity that social media provides (Giachanou & Rosso, 2020).

To purify the content environment for a better user experience, Internet companies and their platforms have recently tried to recognize and deal with harmful content. For example, many social network sites (e.g., Facebook, Twitter, Instagram) have incorporated the "flag" function to enable users to report offensive content online (Crawford & Gillespie, 2016). Furthermore, some online news sites (e.g., Reddit) now afford the moderators of forums the opportunity to maintain perceived quality (Squirrell, 2019) or encourage users to rate (or recommend) comments (or commenters) that exclude those of low quality from discussions (Singer, 2014). Apparently, the enormity of taking action against harmful content makes media platforms turn to ordinary users for help (Porten-Cheé, Kunst, & Emmer, 2020). Users' initiatives and participation through content reporting have become increasingly important to Internet content governance.

In China, Internet content governance has its official name: "Internet Information and Content Ecosystem Governance." According to relevant provisions promulgated by the Cyberspace Administration of China (CAC, 2019), Internet information and content ecosystem governance refer to activities carried out by the government, enterprises, society, and netizens to promote content that conveys "positive energy" (cf. Du, 2014; Peidong & Lijun, 2018) and deal with illegal and bad information. The aim of these activities is to maintain a clean environment on the Internet. The critical role of users is recognized. The CAC has set up a reporting center for illegal and bad information, where users have access to report content they deem harmful online, including categories such as politics, terrorism, scams, pornography, vulgarity, infringement, rumors, and more (more defined types of harmful content, see Appendix 1). The CAC also requires platforms to set up convenient entries for users' complaints and reports (CAC, 2019). Thus, content reporting has dual implications in China, which include not only reporting harmful content to platforms within their affordance scale (e.g., the "flag" function) but also reporting platforms or sites that shield harmful content to the government within the social system of Internet content governance.

Previous studies have raised and discussed the concept of "content moderation," which structures community participation in the overall Internet content governance system to facilitate cooperation and civility (Grimmelmann, 2015). For example, Myers West (2018) conceptualized the educational model for content moderation systems to cultivate the affective relationship between users and platforms, while Fang, Guo, and Zhou (2010) proposed a punitive or controlling model for harmful content in terms of information content security. From the users' perspective specifically, Porten-Cheé et al. (2020) put forward the concept of "online civic intervention" (OCI) to describe users' content reporting as a new form of user-based political participation online that has maintained an accessible and reasoned public debate (p. 515). Several studies have been conducted to investigate factors that might influence users' intentions to report harmful content online, such as gender, age, income, knowledge, morality, authoritarianism, etc. (e.g., Caplan, Hanson, & Donovan, 2018; Watson, Peng, & Lewis, 2019; Wilhelm & Joeckel, 2019). However, psychological insights still need further investigation, and users' participation in Internet content governance has not yet been fully studied in China.

Based on cognitive dissonance theory, the current study sets out to answer the question of why and under what circumstances users will report harmful content while surfing online. The contributions of the current study are mainly from three aspects: First, we revealed the importance and possibility of users' engagement in Internet content governance, which especially supplemented a user perspective for related research in China; second, we investigated the psychological mechanism of users' behavioral intention to report harmful content online, promoting our understanding toward the formulation of users' initiatives in such a public affair; third, the empirical results offered practical implications for encouraging user's participation in online content reporting.

Theoretical Background and Hypothesis Development

The cognitive dissonance theory proposed by Festinger (1957) is one of the most influential theories in social psychology. The theory assumes that inconsistent cognitions arouse dissonance, an aversive state that triggers a desire to mitigate the underlying inconsistency and maintain consistency. Subsequent studies conducted in various contexts have revealed a deep-seated need for cognitive consistency among humans (Kruglanski et al., 2018). It follows the ideas contained in self-consistency theory, which is believed to be the predecessor of cognitive dissonance theory and theories of cognitive consistency (Stevens, 1992). According to the self-consistency theory, each person is defined by a unique structure of ideas and has the superordinate motive of striving for unity that preserves and modifies the idea structure. To preserve this structure, people tend to resist environmental, interpersonal, or psychological events that contradict their self-referential ideas (Lecky, 1945).

In the context of Internet content governance in China, netizens play an important role in this government-led work that calls for social collaboration because of their large scale. Nevertheless, based on different ideas and experiences, perceptions of the necessity of Internet content governance can vary from user to user. Accordingly, we proposed a construct named "perceived necessity of Internet content governance" (perceived necessity) to describe the belief that content on the Internet needs governance for a clean online environment. Perceived necessity is the unique structure of ideas around the Internet content governance of each person. Whether strong or weak, people tend to avoid inconsistent events to keep them stable.

Potential (in)consistency occurs when people encounter harmful online content. The occurrence of harmful content means that there are still content producers disregarding Internet content governance and producing harmful content to destroy online cleanliness. For people with a strong perception of necessity, this notion and related behavior are completely inconsistent with their idea structures, eliciting cognitive dissonance. We proposed another construct called "aversion to harmful Internet content" (aversion), the degree to which people are disgusted with harmful content online, to depict this aversive state. Following the baseline of the relationship between cognition and attitude, evaluative judgments result from cognitive processes (Ajzen & Fishbein, 2000), and emotions could be conditioned to situations perceived as supportive of ideas (Stevens, 1992). Thus, we assumed that, when coming across harmful content, people with a high level of perceived necessity would suffer more from cognitive dissonance that took the explicit form of aversion (cf., Chiang & Su, 2012). In short, perceived necessity was supposed to elicit aversion.

H1: Perceived necessity is positively associated with aversion.

Attitude has been theorized and testified as a crucial predictor of behavioral intentions (Chaffee & Roser, 1986). The explanatory power of attitude on intention prediction has been widely acknowledged from daily behaviors (Sheeran & Orbell, 1999) to online participation, including Internet banking (Tan & Teo, 2000), Internet shopping (Ahn, Ryu, & Han, 2004), e-government usage (Bhattacherjee & Sanford, 2006), and online political participation (Gastil & Xenos, 2010). In the paradigm of cognitive dissonance, specifically to reduce dissonance and increase consonance, people have to make behavioral changes. In the present study, reporting can be the main and most effective way for netizens to eliminate harmful content and objects that cause dissonance, thus reducing dissonance. Behavioral intention to report harmful content online (behavioral intention) refers to the extent of users' intentions to report harmful content online when they encounter it. Behavioral intention was postulated to be promoted by aversion.

H2: Aversion is positively associated with behavioral intention.

Attitude acts as an important mediator between cognition and behavioral intention. As mentioned above, cognitive dissonance stems from inconsistency between perceived necessity and harmful content. As the affective output of cognitive dissonance, aversion represents discomfort when one's fixed structure of ideas is threatened. To leave this aversive state, people are likely to seek behavioral change. Following this influence path, aversion was presumed to mediate the relationship between perceived necessity and behavioral intention.

H3: Perceived necessity has an indirect effect on behavioral intention mediated by aversion.

Rather than (in)consistency itself, the affective consequences of (in)consistency are supposed to depend on motivation, which is the nature of people's goals. More specifically, consistency is likely to arouse positive feelings when it validates a desired belief, while inconsistency is likely to elicit negative feelings when it invalidates a desired belief (Gawronski, 2012). Thus, motivation determines individuals' affect in a situation (Higgins, 1987; Manstead, Frijda, & Fischer, 2004), with an extension to the context of an individual's affective reaction to (in)consistent information (Kruglanski et al., 2018). In this study, we also wondered about the potential effect of motives on cognitive dissonance. We contextualized the motive here as an information-seeking motive. For users with a strong motive for information seeking, harmful content can be seen as hindering the fulfillment of information seeking because harmful content reduces information quality, increasing the difficulty and cost of acquiring qualified information. As a result, dissonance invalidates motivation, making aversion even stronger. Therefore, it can be assumed that motive positively moderates the relationship between perceived necessity and aversion.

H4: Information-seeking motive positively moderates the relationship between perceived necessity and aversion, such that the indirect effect of perceived necessity on behavioral intention will be stronger for those who have high information-seeking motive.

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

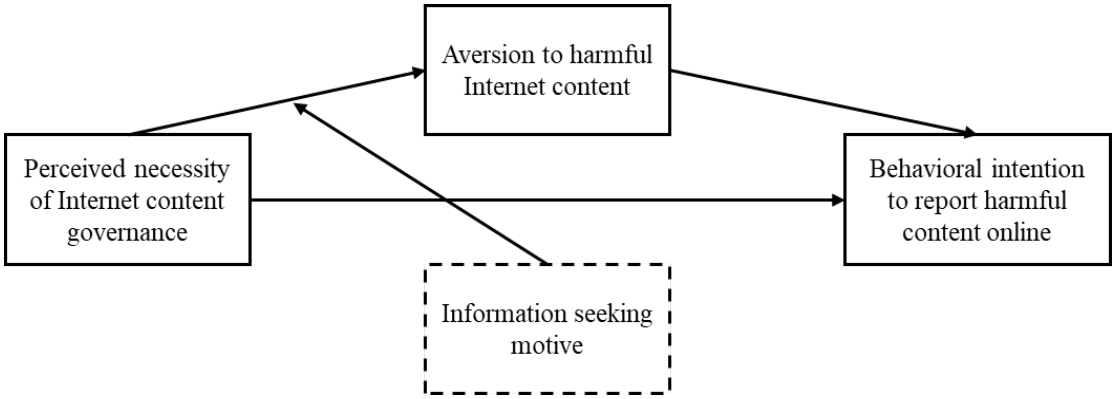

Figure 1. The research model depicting the relationship between the constructs and explaining the hypotheses above.

Method

Respondents and Procedures

Three thousand Internet users in China (Chinese netizens) were recruited via the IPSOS (China), a professional institute of market research guaranteeing diverse samples, in July 2020. Respondents were informed that they were anonymous, that their privacy was protected, that they were volunteering, and that they would receive rewards for participating in the survey.

The study had a final sample of 52.4% males and 47.6% females. Their mean age was 32.74 years (*SD* = 11.64); 15.0% of them were between 16–19 years, 28.1% between 20–29 years, 27.1% between 30–39 years, 19.8% between 40–49 years, and 10.0% more than 50 years. Fifteen percent of them were from a first-tier city, 35.0% from a second-tier city, and 50.0% from a third-tier city. Eighty-seven point six percent of them were from urban areas and 12.4% were from rural areas. About educational background, 10% were from junior high school and below, 48.7% from senior high school or technical school, 21.5% from junior college, and 19.8% from college and above.

Measures

The scales of perceived necessity, aversion, and behavioral intention were newly developed following a strict development process. First, we reviewed relevant literature, including academic articles, policy texts, and platform rules, and held a focus group discussion comprising six PhD candidates to generate three pools of initial items. Initial four items related to perceived necessity derived from the literature (Crawford & Gillespie, 2016; Myers West, 2018; Porten-Cheé et al., 2020; Watson et al., 2019) and the discussion on how to evaluate Internet content governance. For measurements of aversion and behavioral intention, 16 types of harmful content were selected from definitions of harmful content demonstrated in policy texts and platform rules (see Appendix 1) and from the discussion on common harmful content online. Next, the initial items were evaluated by experienced experts (*N* = 5) to guarantee that each item was understandable, precise, and coherent with the construct. Four items related to perceived necessity, 16

items related to aversion, and 16 items related to behavioral intention were modified and retained. All items were rated on a five-point scale.

To verify the validity of the newly developed scales, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were performed successively. For EFA, a presurvey with a random sample (n = 436, female 47.9%, M_{age} = 31.44) originating from IPSOS's online-interviewee panel was conducted. Before EFA, Bartlett's test yielded a significant result (χ^2 = 9,801.92, p < .001), and the KMO test revealed a value of 0.95, well above the suggested value of 0.80, which met the prerequisites for conducting EFA. Principal component analysis (PCA) with varimax rotation was applied to determine the factor structure, standards of which included eigenvalue (q > 1.0), variance explained by potential factors, cross loadings of each item ($q \geq 0.50$), and reasonableness of factor meaning. EFA run on SPSS version 27 extracted 3 factors from a total of 36 items, which accounted for 54.98% of the total variance. The item loadings were all above 0.50 (see Table 1), so a total of 36 items remained. The meaning of the extracted factors corresponded with the constructs proposed above.

Table 1. Exploratory Factor Analysis Results (n = 436).

Items	1	2	3
AV12. I am disgusted with personal attacks online.	**0.763**	0.198	0.158
AV15. I am disgusted with invasion of privacy.	**0.754**	0.172	0.132
AV6. I am disgusted with illegal software.	**0.732**	0.145	0.211
AV16. I am disgusted with crime information.	**0.732**	0.114	0.148
AV8. I am disgusted with Internet viruses.	**0.712**	0.122	0.239
AV7. I am disgusted with hacker attacks.	**0.689**	0.251	0.131
AV3. I am disgusted with Internet scamming.	**0.687**	0.097	0.298
AV11. I am disgusted with Internet rumors.	**0.686**	0.217	0.088
AV2. I am disgusted with fake information.	**0.685**	0.111	0.201
AV14. I am disgusted with copyright infringement.	**0.668**	0.280	0.076
AV13. I am disgusted with human flesh search online.	**0.663**	0.289	0.011
AV9. I am disgusted with violent content.	**0.645**	0.183	0.291
AV10. I am disgusted with reactionary content.	**0.637**	0.264	0.185
AV4. I am disgusted with false advertising.	**0.626**	0.176	0.090
AV1. I am disgusted with vulgar content.	**0.617**	0.251	0.155
AV5. I am disgusted with spam marketing.	**0.598**	0.225	0.026
BI11. When encounter Internet rumors, I will report it.	0.171	**0.757**	–0.055
BI6. When encounter illegal software, I will report it.	0.062	**0.749**	0.180
BI5. When encounter spam marketing, I will report it.	0.023	**0.748**	–0.042
BI14. When encounter copyright infringement, I will report it.	0.225	**0.735**	–0.016
BI4. When encounter false advertising, I will report it.	0.044	**0.733**	–0.056
BI13. When encounter human flesh search online, I will report it.	0.230	**0.704**	–0.029
BI15. When encounter invasion of privacy, I will report it.	0.326	**0.690**	0.099
BI2. When encounter fake information, I will report it.	0.210	**0.684**	0.238

BI12. When encounter personal attacks online, I will report it.	0.327	**0.678**	0.051
BI1. When encounter vulgar content, I will report it.	0.280	**0.652**	0.187
BI8. When encounter Internet viruses, I will report it.	0.204	**0.648**	0.211
BI7. When encounter hacker attacks, I will report it.	0.284	**0.622**	0.370
BI10. When encounter reactionary content, I will report it.	0.297	**0.613**	0.270
BI9. When encounter violent content, I will report it.	0.270	**0.596**	0.344
BI16. When encounter crime information, I will report it.	0.369	**0.549**	0.288
BI3. When encounter Internet scamming, I will report it.	0.365	**0.547**	0.370
PN4. Internet content governance needs to rely on technical means.	0.146	0.074	**0.771**
PN3. The Internet content ecosystem needs to develop a unified norm.	0.256	0.165	**0.718**
PN1. The Internet content needs governance.	0.219	0.133	**0.693**
PN2. Internet content governance needs the cooperation of all social parties.	0.227	0.061	**0.611**

Note. Table 1 shows the results of the rotating component matrix. The items were sorted from high to low according to the loading value.

For CFA, we conducted a second round of presurvey with another random sample (n = 461, female 48.8%, M_{age} = 32.79) offered by the IPSOS as well. CFA (run on Amos version 23) was performed to further validate the EFA. For the validity of the whole factor structure, most of the indices indicated a good fit for the model (see Table 2). For internal consistency, the factor loadings were acceptable and significant, and the values of composite reliability (*CR*) were all above 0.70, representing good internal consistency (see Table 3). Although *AVE*s of perceived necessity and behavioral intention were slightly below the suggested value of 0.50, a high level of *CR*s (q > 0.70) made *AVE*s greater than 0.40 acceptable to support adequate convergent validity (Fornell & Larcker, 1981; Lam, 2012). Moreover, the square root of the *AVE* of each variable was greater than the correlation coefficient between the variable and the others, respectively, indicating good discriminant validity (see Table 4).

Table 2. Model Fit Indices.

CMIN/DF	*RMSEA*	*GFI*	*CFI*	*IFI*	*TLI*
3.524	0.074	0.773	0.855	0.855	0.845

Table 3. Confirmatory Factor Analysis Results (n = 461).

Construct	Items	*Unstd.*	*S.E.*	*t-value*	*Std*	*CR*	*AVE*
Perceived necessity	PN1	1.000			0.715	0.766	0.453
	PN2	1.040	0.081	12.781***	0.757		
	PN3	0.805	0.073	11.079***	0.613		
	PN4	0.789	0.073	10.769***	0.593		
Aversion	AV1	0.988	0.070	14.074***	0.634	0.918	0.507
	AV2	1.096	0.064	17.099***	0.750		
	AV3	1.118	0.063	17.694***	0.772		
	AV4	0.936	0.065	14.503***	0.651		
	AV5	0.922	0.063	14.700***	0.659		

	AV6	1.058	0.059	17.874***	0.778		
	AV7	0.825	0.057	14.565***	0.653		
	AV8	1.06	0.061	17.494***	0.765		
	AV9	1.007	0.063	16.073***	0.712		
	AV10	1.057	0.064	16.415***	0.725		
	AV11	1.028	0.064	16.087***	0.712		
	AV12	1.115	0.062	17.957***	0.781		
	AV13	0.933	0.063	14.893***	0.666		
	AV14	0.959	0.066	14.501***	0.651		
	AV15	1.075	0.060	18.024***	0.784		
	AV16	1.000			0.766		
Behavioral intention	BI1	1.000			0.699	0.914	0.493
	BI2	1.026	0.071	14.377***	0.699		
	BI3	0.947	0.065	14.493***	0.705		
	BI4	1.028	0.083	12.374***	0.599		
	BI5	1.038	0.079	13.114***	0.636		
	BI6	1.188	0.075	15.936***	0.778		
	BI7	1.049	0.068	15.371***	0.749		
	BI8	1.102	0.073	15.092***	0.735		
	BI9	1.056	0.070	15.089***	0.735		
	BI10	0.938	0.065	14.530***	0.707		
	BI11	0.940	0.069	13.675***	0.664		
	BI12	1.070	0.069	15.527***	0.757		
	BI13	1.112	0.074	15.041***	0.732		
	BI14	1.136	0.074	15.323***	0.747		
	BI15	1.027	0.069	14.828***	0.722		
	BI16	0.924	0.063	14.717***	0.716		

***$p < .001$.

Table 4. The Square Root of AVE and the Correlation Coefficient of Perceived Necessity, Aversion and Behavioral Intention.

	Perceived necessity	Aversion	Behavioral intention
Perceived necessity	0.673[a]		
Aversion	0.534	0.712[a]	
Behavioral intention	0.386	0.687	0.702[a]

[a] The square root of *AVE*.

The scale of the information-seeking motive was adapted from an empirical study conducted in China to examine the applicability of media dependency theory in cyberspace (Xie, 2004). Items reflecting users' motives for information-seeking online were selected: "my purpose of Internet usage is to obtain news information," "my purpose of Internet usage is to obtain information related to work

or study," "my purpose of Internet usage is to improve the efficiency of information seeking," and "my purpose of Internet usage is to obtain government services." The items were rated on a five-point scale as well.

Considering the existing research results concerning users' intention to report content, we controlled demographic variables during model testing, including gender, age, income, and knowledge. For measurement, income was specified as average monthly income with assigned values of 1–7 from below CNY 2,000, CNY 2,000–3,999, CNY 4,000–5,999, CNY 6,000–7,999, CNY 8,000–9,999, CNY 10,000–11,999, to CNY 12,000 and above; knowledge was substituted by education with assigned values of 1–4 from junior high school and below to college and above.

Statistical Analyses

First, we calculated descriptive statistics and a correlation matrix. Second, we adopted PROESS macro to testify to the moderated mediation model as hypothesized. Model 4 was applied to test the mediation effect of aversion, and Model 7 was applied to test the moderation effect of motive. Moreover, we used a bootstrapping method (Hayes & Scharkow, 2013), which produced 95% bias-corrected confidence intervals from 5,000 resamples of the data, to investigate the significance of the effect indicated by the zero exclusion of the confidence.

Results

Preliminary Analyses

We first wondered how people perceived the necessity of Internet content governance and how they placed themselves in such a public affair. The results of the survey suggested that, as shown in Figure 2, most respondents recognized the necessity of Internet content governance. More than 80% of the respondents agreed with the statement that the Internet content needed governance, and 74.30% believed that the Internet content ecosystem needed to develop a unified norm. Moreover, it has been widely accepted that Internet content governance should be a systematic work that requires all social parties (81.20% agreement) and technical means (76.90% agreement) to work together.

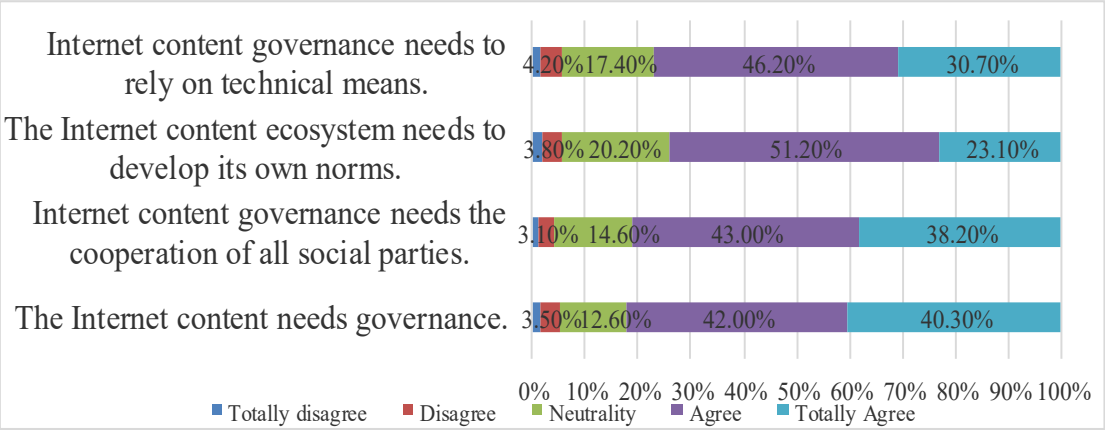

Figure 2. The specific results of the survey on respondents' perceived necessity.

Table 5 shows the descriptive statistics and correlations for the variables measured. The internal consistency alphas were all above 0.70. It can be observed that perceived necessity was positively related to aversion (r = 0.425, p < .01) and behavioral intention (r = 0.346, p < .01), and aversion was positively linked to behavioral intention (r = 0.620, p < .01). In addition, motive was positively associated with perceived necessity (r = 0.516, p < .01), aversion (r = 0.369, p < .01), and behavioral intention (r = 0.333, p < .01).

Table 5. Descriptive Statistics, Alpha Coefficients, and Correlations.

	M	*SD*	1	2	3	4	5	6	7	8
1.Gender	–	–	–							
2.Age	32.75	11.46	0.198**	–						
3.Income	3.40	1.72	0.055**	0.338**	–					
4.Education	2.51	0.92	−0.224**	−0.271**	0.320**	–				
5.Perceived necessity	4.05	0.67	−0.188**	−0.056**	0.082**	0.199**	(0.771)			
6.Aversion	4.38	0.67	−0.215**	−0.072**	−0.027	0.198**	0.425**	(0.946)		
7.Behavioral intention	4.24	0.68	−0.124**	−0.105**	–0.040*	0.076**	0.346**	0.620**	(0.940)	
8.Motive	3.95	0.63	−0.142**	0.109	0.142**	0.170**	0.516**	0.369**	0.333**	(0.720)

Note. N = 3,000. Internal reliabilities (alpha coefficients) for the constructs are given in parentheses on the diagonal. *p < .05. **p < .01. ***p < .001.

Testing for Mediation Effect

The results of the mediation effect analysis are reported in Table 6, which revealed that perceived necessity positively influenced aversion (B = 0.3844, SE = 0.0166, $p < .001$. see Model 1 of Table 6), and aversion positively affected behavioral intention (B = 0.5995, SE = 0.0162, $p < .001$. See Model 2 of Table 6). This supported Hypotheses 1 and 2. Bootstrapping analysis demonstrated the significance of the indirect effect of perceived necessity on behavioral intention (B = 0.2304, SE = 0.0179) indicated by a 95% confidence interval of [0.1958, 0.2664], which excluded zero. The mediation effect was, therefore, verified, supporting Hypothesis 3.

Table 6. Testing the Mediation Effect of Aversion.

Predictors	Model 1 (Aversion)			Model 2 (Behavioral intention)		
	B	*SE*	*t*	*B*	*SE*	*t*
Gender	−0.1512	0.0227	−6.67***	0.0265	0.0203	1.31
Age	0.0030	0.0011	2.70**	−0.0056	0.0010	−5.68***
Income	−.0453	0.0075	−6.05***	0.0110	0.0067	1.64
Education	0.1067	0.0140	7.61***	−0.0684	0.0126	−5.44***
Perceived necessity	0.3844	0.0166	23.17***	0.1106	0.0160	6.91***
Aversion				0.5995	0.0162	36.96***
R^2	0.2175			0.4031		
F	166.47***			336.84***		

Note. N = 3,000. Each column is a regression model that predicts the criterion at the top of the column.
*$p < .05$. **$p < .01$. ***$p < .001$.

Testing for Moderation Effect

The result of the moderation effect analysis is displayed in Table 7, where the interaction of perceived necessity with motive negatively predicted aversion (B = −0.0982, SE = 0.0202, $p < .001$. See Model 1 of Table 7).

Table 7. Testing the Moderated Mediation Effect of Motive on Behavioral Intention.

Predictors	Model 1 (Aversion)			Model 2 (Behavioral intention)		
	B	*SE*	*t*	*B*	*SE*	*T*
Gender	−.1349	0.0222	−6.07***	0.0265	0.0203	1.31
Age	0.0022	0.0011	2.02*	−0.0056	0.0010	−5.68***
Income	−0.0501	0.0074	−6.81***	0.0110	0.0067	1.64
Education	0.1004	0.0137	7.31***	−0.0684	0.0126	−5.44***
Perceived necessity	0.6480	0.0765	8.47***	0.1106	0.0160	6.91***
Aversion	–	–	–	0.5995	0.0162	36.96***
Motive	0.6021	0.0824	7.30***	–	–	–
Perceived necessity x Motive	−0.0982	0.0202	−4.87***	–	–	–
R^2	0.2521			0.4031		
F	144.09***			336.84***		

Note. *N* = 3,000. Each column is a regression model that predicts the criterion at the top of the column. *p < .05. **p < .01. ***p < .001.

We then plotted a simple slope that illustrated the relationship between perceived necessity and aversion for high and low levels of motive, respectively. As Figure 3 shows, the slope representing the relationship between perceived necessity and aversion with a high level of motive was weaker ($B_{\text{high information-seeking motive}}$ = 0.1985, *SE* = 0.0260, *t* = 7.6423, p < .001), while the slope was comparatively stronger when the motive was relatively weak ($B_{\text{low information-seeking motive}}$ = 0.3213, *SE* = 0.0199, *t* = 16.1185, p < .001).

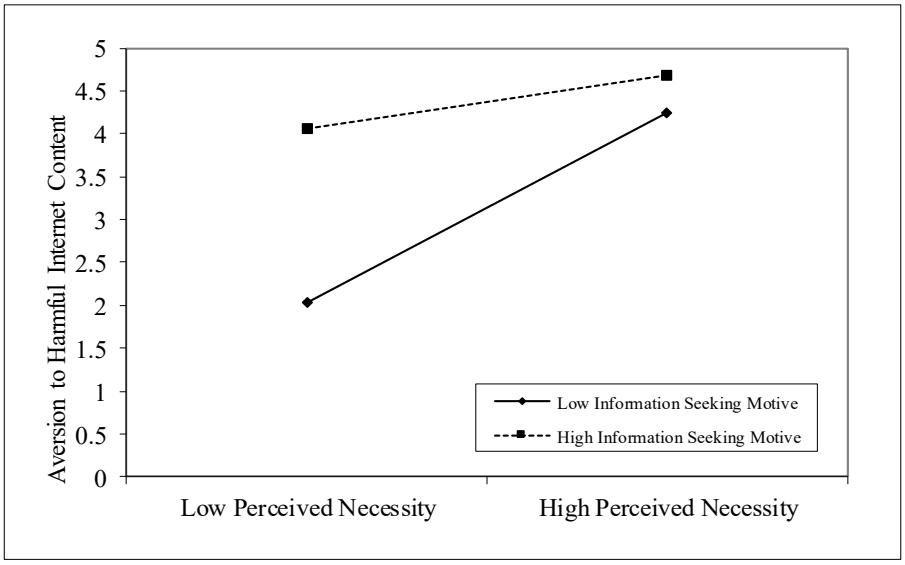

Figure 3. Interaction effect of perceived necessity and motive. High and low levels of perceived necessity and motive were distinguished by one standard deviation above and below the mean.

We tested the conditional indirect effects of perceived necessity on behavioral intention and found that for respondents with a high level of motive, perceived necessity had a weaker indirect effect on behavioral intention (*B* = 0.1190, *SE* = 0.0194, 95% *CI* = [0.0820, 0.1577]), compared with those with a low level of motive (*B* = 0.1926, *SE* = 0.0192, 95% *CI* = [0.1553, 0.2299]). The index of moderated mediation was reported as follows: *B* = −0.0589, *SE* = 0.0160, 95% *CI* = [−0.0887, −0.0263]. This indicates that, under the moderation effect of motive, the indirect effect of perceived necessity on behavioral intention was significantly mitigated. Thus, motive had a significant moderation effect on the relationship between perceived necessity and aversion and the indirect effect of perceived necessity on behavioral intention. However, the direction of the influence was found to differ from that presumed in Hypothesis 4.

Discussion and Implications

Supported by empirical evidence, the present study sought to understand the psychological mechanism of the behavioral intentions to report harmful content online among Chinese netizens. Most of the hypotheses raised above were verified: Perceived necessity was found to have an indirect effect on behavioral intention mediated by aversion; however, the positive relationship between perceived necessity and aversion was negatively moderated by motive, and then the whole indirect effect was significantly mitigated, contradicting the hypothesized influence direction. The empirical results provided some theoretical insights into social psychology and behavior research in the context of Internet content governance in China.

On the whole, based on cognitive dissonance theory, the current study built and verified a moderated mediation model concerning why and under what circumstances users would report harmful content online. Most notably, it enriched the theory of cognitive dissonance. The theoretical outline of "cognition, attitude, and behavior" and convincing empirical evidence helped clarify the working mechanism of cognitive dissonance orienting to behavioral change. Specific to the research context, the psychological mechanism of online content reporting was investigated. Previous studies have mainly focused on users' characteristics and personalities (e.g., Caplan et al., 2018; Watson et al., 2019; Wilhelm & Joeckel, 2019). We complemented them by revealing psychological processes involving factors such as cognition, attitude, and motivation.

To be more specific, the positive relationship between perceived necessity and aversion was examined in that people who perceived more necessity for Internet content governance tended to generate more aversion to harmful content. This result echoes previous research (e.g., Kalch & Naab, 2018; Wilhelm & Joeckel, 2019) that negative attitude (i.e., aversion) toward harmful content, not only online but also offline, such as through gossip (Wu, Birtch, Chiang, & Zhang, 2018), results from a process of cognitive dissonance (inconsistency between "what you believe in" versus "what you're exposed to"). Further, with a positive correlation between aversion and behavioral intention confirmed, content reporting proved to be an effective way to reduce dissonance. The reduction was laid in two paths: one was to reduce or even eliminate harmful content that caused inconsistency online, and the other was to seek confirmation or emotional backup. In an empirical work concerning misinformation diffusion around food safety, it was found that people with high trust in food safety experienced more dissonance, driving them to diffuse misinformation to relieve uncertainty or share the burden of anxiety (Wang, He, Xu, & Zhang, 2020). This could also be the

case in this study: people reported harmful content online to confirm that "harmful content is harmful indeed" and to maintain cognitive consistency of perceived necessity, as well as share the burden of discomfort and aversion.

The mediation effect of aversion and the indirect effect of perceived necessity on behavioral intention were verified. Dating back to the ideas of self-consistency again, people's motives for striving for unity take two forms: resistance and assimilation. Assimilation, which permits the simultaneous resolution of inconsistencies and the adaptive evolution of personality, is supposed to be preferable to resistance (Stevens, 1992). By placing aversion as the mediator, the present research, however, proves the existence and significance of resistance in the context of online content reporting. That is, to cope with the cognitive dissonance resulting from harmful online content, people are likely to take action of resistance. This resistance manifests as an aversion to emotion and content reporting in behavior.

Contrary to our hypothesis, the information-seeking motive negatively moderated the relationship between perceived necessity and aversion, such that the indirect effect of perceived necessity on behavioral intention became weaker for those who had a high information-seeking motive. Possible explanations are that, with a high information-seeking motive, people tend to promote more openness to various content, including harmful content, for an efficient and effective outcome of information acquisition. During this process, content quality may give way to informativeness. For example, Internet advertising has often been perceived as an intrusion during online activities (Li, Edwards, & Lee, 2002), while for users with a strong motive for information seeking, it can be acceptable if it delivers informativeness (Celebi, 2015). Similar situations also happen where privacy concern is overridden by hedonic motivations (Chen & Kim, 2013), fake news sharing during the COVID-19 pandemic wrapped in motivations of helping others (Duffy, Tandoc, & Ling, 2020; Plume & Slade, 2018).

Moreover, this study made a methodological contribution by developing new scales of perceived necessity, aversion, and behavioral intention. The scale of perceived necessity describes people's belief that Internet content needs governance and enables us to measure people's cognitive structures around Internet content governance on which people's views may have individual differences. As for the scales of aversion and behavioral intention, we expanded the vague and generalized context of content reporting to a more detailed one. By comprising 16 types of harmful content, the measurements became more comprehensive and contextualized. These newly developed scales showed good reliability in the current study, providing measurement tools for subsequent research on user engagement in content governance and content moderation.

For managerial implications, since the influence of perceived necessity was significant, it was necessary to raise and cultivate people's cognition of the necessity and their capacity for Internet content governance. Nowadays, netizen reporting has become a new approach to interaction between government authorities and citizens (Alhammad, Hajar, Alshathry, & Alqasabi, 2021). Communicating the great significance of users' participation as "You Matter" and appealing to the rhetoric of the online community and public sphere as "You Should" (Myers West, 2018) help nurture their self-discipline, including behavior regulation of content production and misuse correction of content reporting (Gillespie, 2018), and online citizenship. Reward incentive mechanisms for reasonable content reporting should also be built to ignite

people's initiatives. Finally, the verified mediation effect suggests that content reporting can act as a key to dissonance reduction. However, whether cognitive dissonance is relieved depends on the responses of the information receivers (Wang et al., 2020). If the receiver confirms the perception of necessity for Internet content governance and the definitions of harmful content, the sender with a high level of perceived necessity would stick to harmful content reporting, and the sender with a low level of perceived necessity would conversely reestablish cognitive consonance. Hence, the feedback produced by the government or the platforms after content reporting is of great significance. For instance, to inform the users of the notion as "You Can," platforms should increase the visibility of the digital affordance for content reporting, for example, explicitly convey information on usage policies and intervention options (Naab, Kalch, & Meitz, 2018), clarify the criteria and mechanism of content moderation in advance (Baker, Wade, & Walsh, 2020; Myers West, 2018), and deal with users' reports effectively with timely feedback and detailed processing results (Baker et al., 2020), fostering not only involvement but also trust in platforms.

Limitations and Future Research Directions

This research has several limitations that need to be addressed in future studies. With cultural and ideological differences taken into consideration, the concept of "harm" can be controversial (Baker et al., 2020). Because of the process of cognitive dissonance, people's aversion to harmful content will also be subject to the scale of what content they define as "harmful." Thus, there is a need for future research to further investigate the compatibility of consistency among cognition, attitude, and intention in online content reporting. We also welcome and encourage researchers to validate the measurements of the perceived necessity of Internet content governance, aversion to harmful Internet content, and behavioral intention to report harmful content online in future research. Moreover, to fully understand why and under what circumstances netizens are willing to report harmful content online, in addition to cognition, attitude, and motive, factors such as trust, ideology, social influence, and perceived efficacy should be investigated to further promote the explanatory power of the predictive model.

References

Ahn, T., Ryu, S., & Han, I. (2004). The impact of the online and offline features on the user acceptance of Internet shopping malls. *Electronic Commerce Research and Applications, 3*(4), 405–420. doi:10.1016/j.elerap.2004.05.001

Ajzen, I., & Fishbein, M. (2000). Attitudes and the attitude-behavior relation: Reasoned and automatic processes. *European Review of Social Psychology, 11*(1), 1–33. doi:10.1080/14792779943000116

Alhammad, M. M., Hajar, L., Alshathry, S., & Alqasabi, M. (2021). Motivational factors impacting the use of citizen reporting applications in Saudi Arabia: The case of Balagh application. *International Journal of Advanced Computer Science and Applications, 12*(6), 264–272. doi:10.14569/IJACSA.2021.0120629

Baker, S. A., Wade, M., & Walsh, M. J. (2020). The challenges of responding to misinformation during a pandemic: Content moderation and the limitations of the concept of harm. *Media International Australia, 177*(1), 103–107. doi:10.1177/1329878X20951301

Bhattacherjee, A., & Sanford, C. (2006). Influence processes for information technology acceptance: An elaboration likelihood model. *MIS Quarterly, 30*(4), 805–825. doi:10.2307/25148755

Caplan, R., Hanson, L., & Donovan, J. (2018). *Dead reckoning: Navigating content moderation after "fake news"*. Data & Society Research Institute. Retrieved from https://datasociety.net/pubs/oh/DataAndSociety_Dead_Reckoning_2018.pdf

Celebi, S. I. (2015). How do motives affect attitudes and behaviors toward internet advertising and Facebook advertising? *Computers in Human Behavior, 51*, 312–324. doi:10.1016/j.chb.2015.05.011

Chaffee, S. H., & Roser, C. (1986). Involvement and the consistency of knowledge, attitudes, and behaviors. *Communication Research, 13*(3), 373–399. doi:10.1177/009365086013003006

Chen, H.-T., & Kim, Y. (2013). Problematic use of social network sites: The interactive relationship between gratifications sought and privacy concerns. *Cyberpsychology, Behavior, and Social Networking, 16*(11), 806–812. doi:10.1089/cyber.2011.0608

Chiang, I.-P., & Su, Y.-H. (2012). Measuring and analyzing the causes of problematic Internet use. *Cyberpsychology, Behavior, and Social Networking, 15*(11), 591–596. doi:10.1089/cyber.2011.0099

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

Cyberspace Administration of China. (2019). *Provisions on information and content ecosystem governance*. Retrieved from http://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm

Du, S. (2014). Social media and the transformation of "Chinese nationalism": "Igniting positive energy" in China since the 2012 London Olympics. *Anthropology Today, 30*(1), 5–8. doi:10.1111/1467-8322.12083

Duffy, A., Tandoc, E., & Ling, R. (2020). Too good to be true, too good not to share: The social utility of fake news. *Information, Communication & Society, 23*(13), 1965–1979. doi:10.1080/1369118X.2019.1623904

Fang, B., Guo, Y., & Zhou, Y. (2010). Information content security on the Internet: The control model and its evaluation. *Science in China Series F: Information Sciences, 53*(1), 30–49. doi:10.1007/s11432-010-0014-z

Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Redwood City, CA: Stanford University Press.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. doi:10.1177/002224378101800104

Gastil, J., & Xenos, M. (2010). Of attitudes and engagement: Clarifying the reciprocal relationship between civic attitudes and political participation. *Journal of Communication, 60*(2), 318–343. doi:10.1111/j.1460-2466.2010.01484.x

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition, 30*(6), 652–668. doi:10.1521/soco.2012.30.6.652

Giachanou, A., & Rosso, P. (2020, October). The battle against online harmful information: The cases of fake news and hate speech. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 3503–3504). doi:10.1145/3340531.3412169

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

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law & Technology, 17*, 42–109. Retrieved from https://yjolt.org/sites/default/files/grimmelmann_the-virtues-of-moderation_0.pdf

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

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review, 94*(3), 319–340. doi:10.1037/0033-295X.94.3.319

Kalch, A., & Naab, T. K. (2018). 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

Kruglanski, A. W., Jasko, K., Milyavsky, M., Chernikova, M., Webber, D., Pierro, A., & Di Santo, D. (2018). Cognitive consistency theory in social psychology: A paradigm reconsidered. *Psychological Inquiry, 29*(2), 45–59. doi:10.1080/1047840X.2018.1480619

Lam, L. W. (2012). Impact of competitiveness on salespeople's commitment and performance. *Journal of Business Research, 65*(9), 1328–1334. doi:10.1016/j.jbusres.2011.10.026

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., . . . Rothschild, D. (2018). The science of fake news. *Science, 359*(6380), 1094–1096. doi:10.1126/science.aao2998

Lecky, P. (1945). *Self-consistency; A theory of personality.* Washington, DC: Island Press.

Li, H., Edwards, S. M., & Lee, J.-H. (2002). Measuring the intrusiveness of advertisements: Scale development and validation. *Journal of Advertising, 31*(2), 37–47. doi:10.1080/00913367.2002.10673665

Manstead, A. S., Frijda, N., & Fischer, A. (2004). *Feelings and emotions: The Amsterdam symposium.* Cambridge, UK: Cambridge University Press.

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

Peidong, Y., & Lijun, T. (2018). "Positive energy": Hegemonic intervention and online media discourse in China's Xi Jinping Era. *China: An International Journal, 16*(1), 1–22. doi:10.1353/chn.2018.0000

Plume, C. J., & Slade, E. L. (2018). Sharing of sponsored advertisements on social media: A uses and gratifications perspective. *Information Systems Frontiers, 20*(3), 471–483. doi:10.1007/s10796-017-9821-8

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

Sheeran, P., & Orbell, S. (1999). Augmenting the theory of planned behavior: Roles for anticipated regret and descriptive norms 1. *Journal of Applied Social Psychology, 29*(10), 2107–2142. doi:10.1111j.1559-1816.1999.tb02298.x

Singer, J. B. (2014). User-generated visibility: Secondary gatekeeping in a shared media space. *New Media & Society, 16*(1), 55–73. doi:10.1177/1461444813477833

Squirrell, T. (2019). Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society, 21*(9), 1910–1927. doi:10.1177/1461444819834317

Stevens, M. J. (1992). Prescott Lecky: Pioneer in consistency theory and cognitive therapy. *Journal of Clinical Psychology, 48*(6), 807–811. doi:10.1002/1097-4679(199211)48:6<807::AID-JCLP2270480615>3.0.CO;2-C

Tan, M., & Teo, T. S. (2000). Factors influencing the adoption of Internet banking. *Journal of the Association for information Systems, 1*(1). doi:10.17705/1jais.00005

Wang, R., He, Y., Xu, J., & Zhang, H. (2020). Fake news or bad news? Toward an emotion-driven cognitive dissonance model of misinformation diffusion. *Asian Journal of Communication, 30*(5), 317–342. doi:10.1080/01292986.2020.1811737

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. (2019). Gendered morality and backlash effects in online discussions: An experimental study on how users respond to hate speech comments against women and sexual minorities. *Sex Roles, 80*(7), 381–392. doi:10.1007/s11199-018-0941-5

Wu, L.-Z., Birtch, T. A., Chiang, F. F., & Zhang, H. (2018). Perceptions of negative workplace gossip: A self-consistency theory framework. *Journal of Management, 44*(5), 1873–1898. doi:10.1177/0149206316632057

Xie, X. (2004). "媒介依赖"理论在互联网环境下的实证研究 [The empirical study on the media dependency theory in cyberspace]. *Journal of Shijiazhuang University of Economics, 27*(2), 218–224. doi:10.13937/j.cnki.sjzjjxyxb.2004.02.027

Appendix 1. Types of Harmful Content Defined by Governance Subjects.

Governance subject	Defined types of harmful content
Cyberspace Administration of China	**Illegal information:** information that (1) opposes the basic principles established by the Constitution; (2) endangers national security, divulges state secrets, subverts state power, or undermines national unity; (3) damages to national honor and interests; (4) distorts, vilifies blasphemes or denies the deeds and spirits of heroes and martyrs, and infringe upon the names, portraits, reputation and honor of heroes and martyrs by insulting, slandering or other means; (5) advocates terrorism, extremism or inciting terrorist or extremist activities; (6) incites ethnic hatred, ethnic discrimination, or undermining ethnic unity; (7) undermines the state's religious policy and promoting cults and feudal superstitions; (8) spreads rumors and disrupts economic and social order; (9) spreads obscenity, pornography, gambling, violence, murder, terror or instigates crimes; 10) insults or slanders others, infringes on their reputation, privacy and other legitimate rights and interests; 11) contains other content prohibited by laws and administrative regulations. **Bad information:** information that (1) uses exaggerated titles, and the content is seriously inconsistent with the title; (2) hypes up scandals, bad deeds, etc.; (3) improperly comments on natural disasters, major accidents, and other disasters; (4) has sexual implication, sexual provocation, etc. that are likely to cause people to have sexual associations; (5) shows blood, horror, cruelty, etc. that cause physical and mental discomfort; (6) incites crowd discrimination, regional discrimination, etc.; (7) promotes vulgar and kitsch content; (8) may cause minors to imitate unsafe behaviors and behaviors that violate social morality, induces minors to have bad habits, etc.; (9) contains other content that has a negative impact on the Internet ecology.
Illegal and Bad Information Reporting Center	**Illegal and bad information:** information that (1) endangers national security, honor, and interests; (2) incites to subvert state power and overthrow the socialist system; (3) incites to split the country and undermine national unity; (4) advocates terrorism and extremism; (5) advocates ethnic hatred and ethnic discrimination; (6) spreads violent, obscene and pornographic information; (7) fabricates or spreads false information to disrupt economic and social order; (8) infringes upon the legitimate rights and interests of others' reputation, privacy, etc.; (9) contains other content prohibited by Internet-related laws and regulations. **Reporting access:** (1) politics; (2) terrorism; (3) scams; (4) pornography; (5) vulgarity; (6) infringement; (7) rumors; (8) historical nihilism related; (9) COVID-19 pandemic prevention related; 10) cyberbullying related; 11) online cultural products related; 12) minors related; and 13) others.

Weibo	**Irregularities:** a) post harmful information; b) publish false information; c) user dispute. **Harmful information:** (1) information that endangers national and social security (in accordance with the illegal information defined by the CAC); (2) spam; and (3) obscene information. **False information:** information that (1) is totally not true; (2) fabricates details; (3) contains inconsistency between pictures and texts; (4) exaggerates facts; (5) has expired; (6) is incomplete; and (7) is out of context. **User dispute:** (1) privacy leak; (2) personal attacks; (3) impersonating others; (4) plagiarism; (5) harassment; (6) fake user identity.
WeChat	**Violating content:** (1) content prohibited by laws and regulations (basically in accordance with the illegal information defined by CAC); (2) false information; (3) pornographic content; (4) gambling content; (5) content that contains violence and crimes; (6) content related to gangsterism and terrorism; (7) infringing content; (8) fraud information; (9) content that refers to illegal goods; 10) content that compromises platform security; 11) bad information (e.g., content that entices users to share); 12) others.
Douyin	**Forbidden information:** in addition to the illegal information defined by CAC, also includes information that (1) endangers the Internet security, uses the Internet to endanger national security, honor and interests; (2) intimidates and threatens others with violence and conducts human flesh searches; (3) involves the privacy, personal information or data of others; (4) spreads foul language and damages social order and good customs; (5) violates other's privacy, reputation, portrait, intellectual property and other legitimate rights and interests; (6) distributes commercial advertisements, or similar commercial solicitations, excessive marketing, and spam; (7) contains comments in languages other than those commonly used on this website; (8) has nothing to do with the information reviewed; (9) is meaningless, or uses a combination of characters deliberately to evade technical review; 10) infringes upon the legitimate rights and interests of minors or damages the physical and mental health of minors; 11) secretly photographs or records others without permission, infringes upon the legal rights of others; 12) contains horrific, violent, bloody, high-risk content that endangers the physical and mental health of the performer or others.