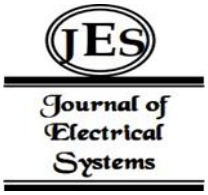
[1]Chengying Yu

[1]Yiwen Chen

Online Hate Speech Detection and Management from Bystander Intervention Perspective based on ETPB Model



Abstract: - The negative impact of hate speech spreading on social media, such as hatred towards targets' sexual orientation, ethnicity, refugee and gender etc., has received increasing attention in recent years. Encouraging users to intervene as bystanders, such as reporting or flagging the hatred speech and making counter-speech, has gradually become a new trend for Internet governance by SNS providers. This study is based on the extension of the theory of planned behavior (ETPB) to identify the predictive factors for bystander intervention intention from a cognitive sight. Research was conducted through 486 online social media user questionnaires; the conclusion is perceived behavioral control, moral norms, and behavioral attitudes can positively predict the behavior intention, while the effect of subjective normative is not significant. The results could be piloted and implemented by SNS providers to encourage more active intervention from users to improve the efficiency of online hate speech detection and management.

Keywords: Online hate speech; social media; Bystander; The theory of the planned behavior

1. Introduction

With the widespread use of social media in recent years and the frequent occurrence of various political events around the world, such as anti-Asian speech during Covid-19, racist and anti-refugee speech after refugee crisis and hatred speech related LGBT+ , has generated more concerns on social media in various countries. 2020 study in 6 countries including United Kingdom, United States, France, Spain etc. found that more than 70% of users had encountered hate speech on social media that created negative emotions such as anger and guilty for users (Reichelmann, et al., 2021); Online hate speech not only affects users' mental health, leading to psychological problems such as depression and frustration, but also creates a biased and intolerant online environment that fosters discrimination and hostility. In severe cases, it could exacerbate offline violence. Current research on online hate speech mainly focuses on the use of algorithms to build automatic identification models of online hate speech. However, hate speech is vaguely defined and difficult to be fully identified by the system, therefore SNS providers also rely on user reports and manual screening (Bian & Chen, 2021) The study focus on bystander intervention behavior to provide a differentiated hate speech management program to reduce online abusive words.

1.1 Online hate speech and bystander response

Online hate speech (OHS) is defined as abusive expression by individuals inciting violence, hatred or discrimination towards certain social groups (Hawdon, et al., 2017). Studies have demonstrated that OHS can have a long-term negative impact on the individuals and the society. It has a negative impact on the mental health of users, leading to the feelings of frustration, fear and anger (Masullo Chen & Lu, 2017), as well as psychological

[1] Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China; Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049 China.. Email:cy_yu66@126.com

stress and depression (Gelber & McNamara, 2016). Meanwhile, the dissemination of hate speech engenders an online milieu of prejudice and intolerance, which fosters discrimination, hostility, and in extreme cases, exacerbates offline violence such as terrorism (Gagliardone, et al., 2015). And finally, widespread hate speech also harms the social media user experience, Bullying language increasing social overload, including psychological distress and social network fatigue, ultimately leading to the discontinuation use of SNS. Therefore, hate speech detection is not only a legal obligation for SNS providers, but also in the interests of SNS company to reduce user churn.

A bystander is individuals who are present at an event but does not participate in it. OHS bystanders are users who witnesses the hate speech against certain social groups online and is not attributed to the group being attacked. Proactive interventions by users as bystanders to hate speech, such as flagging and reporting hate speech, providing counter-speech, and consoling the victims, not only assist SNS platform algorithms in identifying inappropriate speech expeditiously, but also effectively comfort the OHS victims (Hurd et al., 2022).

Most research on the OHS bystander intervention is based on Latane and Darley's 1970 Bystander Intervention Model (BIM), which examines the factors that influence bystander intervention intentions and behaviors in 5 steps. Personal factors, including empathy (Paterson, et al., 2019), personal responsibility (Naab, et al., 2018), peer norms (Henson, et al., 2020), and attitudes towards targeted hate groups (Weber, et al., 2020), are positively correlated with bystander intervention behaviors. Environmental factors also valued. The number of bystanders was negatively correlated with intervention behavior due to the diffusion of responsibility effect (Darley & Latané, 1986), while incident severity was positively associated with intervention behavior (Leonhard, et al., 2018).

While the findings of the studies have proven challenging to translate into practical guidance for SNS provider to encourage bystanders' intervention behaviors in the long term. For instance, empathy and self-efficacy have been demonstrated to be effective in short-term interventions, but facing challenges in long-term interventions due to users' personal trait (Soral, 2022). Furthermore, the efficacy of bystander intervention behaviors is constrained by the availability of resources such as SNS's support from technique perspective, which was rarely mentioned in previous studies.

1.2 The theory of the planned behavior

The Theory of Planned Behavior (TPB) was developed by Ajzen based on the Theory of Reasoned Action (TRA), which he and Fishbein proposed in 1980. The TPB explains the general decision-making process of an individual's behavior from the perspective of information processing, based on the expected value theory. The model has been widely employed to predict, explain, and intervene in social behaviors (Ajzen, 1991). Since TPB analysis a behavior intention from rational perspective, SNS can create a long-term program to encourage users' intervention via change users' perceptions on OHS.

The theory posits that behavioral intention (Intention, INT) is the intention to perform a particular behavior. It is proposed that the stronger the intention, the more likely the behavior is to occur. Behavioral intentions are determined by a combination of attitudes towards the behavior, subjective norms and perceived behavioral control. Model also can be extended with adding additional predictor variables to improve the explanatory level of behavior intention. (Ajzen, 2020), called extended theory of planned behavior (ETPB).

Attitude (AT) towards the behavior refers to the extent to which an individual values the behavior positively or negatively. Positive attitudes towards OHS intervention predict user's higher intervention intentions. In a study conducted by Hurd et al. (2022) on racist speech intervention among US college students, when White students were aware that positive interventions would comfort Black students and improve the campus atmosphere, students' attitudes towards interventions were more positive and significantly increased intervention intentions. Research related to cyberbullying has also found that positive outcome assessments, such as positive feelings following intervention also influence intervention attitudes. In contrast, negative outcome assessments, such as when an individual perceives the intervention to be ineffective or may experience peer disapproval, can negatively influence intervention attitudes (Desment, et al., 2014). This leads to the following hypothesis:

H1: Attitudes towards OHS intervention predict intervention intentions.

Subjective (SN) norm is defined as a person's perceived social pressure to perform a particular behavior. Parents, family members, peers, and the school environment are the most common significant others in cyberbullying intervention. (Santre, 2022; Desment, et al., 2014; Lazuras, et al., 2013). Additionally, research on environmentally relevant pro-social behavior frequently extends the concept of significant others to communities and governments (Lou, et al., 2020). Subjective norms predict bystanders' intentions to intervene in OHS events when they perceive the intervention is in line with people's expectations. This leads to the following hypothesis:

H2: Individuals' subjective norms predict OHS intervention intentions.

Perceived behavioral control (PBC) is a person's perception of how easy it is to perform the behavior may influence when they are willing to perform it. A lack of knowledge and skills to intervene as well as support from people around such as teachers or peers, are the difficulties being frequently mentioned in cyberbullying related studies (Desment, et al., 2014; Obermaier, 2022). Nevertheless, since indirect interventions such as reporting or flagging also require SNS platforms to provide substantial support, including the smooth user journey and instant feedback to users' report. When individuals perceive they have sufficient capability to intervene, they are more likely to engage in. This leads to the following hypothesis:

H3: The perceived behavioral control predicts user's intervention intentions in response to OHS.

Moral norms (MNs) are values that individuals develop during the socialization process (Conner & Armitage, 1998). Moral norms are activated when individuals are aware of the impact of their behaviors on others and agree that they have obligation to perform so (Rivis, et al., 2009). The TPB offers a rational approach to decision-making of the behavior, yet it fails to acknowledge the influence of morally relevant factors (Kaiser & Scheuthle, 2003). The incorporation of ethics as an additional variable in the TPB effectively improve the predictive model of bystanders' intervention in school bullying (Brehmer, 2023) and sexual assault (Branscum, et al., 2023) . Given that OHS intervention also concerns social morality and ethics, it was postulated that moral norms would positively predict intervention intentions, leading to:

H4: Individuals' perceptions of moral norms regarding to OHS interventions predict their intentions to intervene as bystanders.

ETPB model regarding to OHS refer to Figure 1.

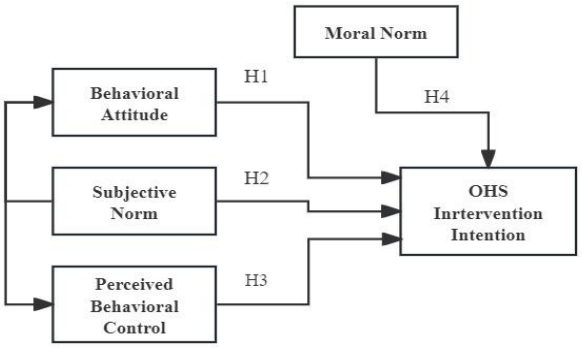


Figure 1 ETPB model of OHS intervention intention

2 METHODS

2.1 Participants

A convenience sampling method was employed to randomly recruit 539 SNS users to complete the questionnaires via SNS including Weibo, RED, Douban and WeChat during January and February 2024. Those who did not answer the questionnaires in a conscientious manner were excluded using the following methods: (1) incorrectly answered screening questions; and (2) answering the questionnaires in less than 100 seconds. 486 valid questionnaires were kept, representing a recovery rate of 90.17%. Among the valid responses, 216 (44.44%) were male and 270 (55.56%) were female. The majority of respondents, 232 (47.73%), were aged between 18 and 25, while 154 (31.69%) were aged between 26 and 30. Most of the respondents had attained a college education or higher (86.83%). The overall samples is consistent with the characteristics of Chinese social media users, exhibiting a balanced ratio of men and women and a predominantly young user base. The demographic characteristics of the subjects are presented in Table 1.

	Frequency	**%**
Gender		
Male	216	44.4
Female	270	55.6
Age		
Below 18	7	1.4
18-25	232	47.7
26-30	154	31.7
31-40	78	16.0
41 and above	15	3.1
Education		
Middle school or below	9	1.9
High school	55	11.3
College	134	27.6

Bachelor	240	49.4
Postgraduate or above	48	9.9

Table 1 Demographics of valid respondents

2.2 Measures

The independent and dependent variables in the TPB were adapted from well-established scales from relevant studies (Hayashi, 2021; Lou, et al., 2020), which measured behavioral attitudes with four items; subjective norms with three items; perceived behavioral control with four items; and willingness to intervene with three items. The extension variable, moral norm, was derived from the Brehmer (2023) and Branscum et al. (2023) studies on online and offline bullying interventions, which include four items. The variables were measured using a seven-point Likert scale, with scores ranging from one to seven.

The questionnaire also included demographics questions (gender, age, education, and experience of online hate speech victimization), which were employed as control variables. Following the design of the questionnaire, a preliminary survey was conducted to assess the reliability and validity of the scales. Based on the findings of this survey, the questionnaire was revised and subsequently adopted as the official version.

3 DATA ANALYSIS

After excluded invalid responses, data were processed using SPSS 21.0 for analyses of reliability, validity and correlation. Meanwhile, factor analysis and structural equation modelling were conducted using AMOS 24.0 sin order to validate the decision-making model of the intervention behavior.

The Harman one-factor test was employed, and the outcomes indicated the presence of eight factors with eigenvalues exceeding 1. The first factor accounted for 33.09% of the variance, which was below the critical threshold of 40%. Consequently, there is no significant issue of common method bias in this study.

To ensure the reliability and validity of the study, 486 valid samples were used to test the reliability and validity of the scale. The results are presented in Table 2, which shows that Cronbach's alpha coefficients are greater than 0.7. Furthermore, all factor loadings are greater than 0.7 and $p < 0.001$, indicating that the measures has been validated. The combined reliability (CR) is also greater than 0.7, which demonstrates that the questionnaire used has good reliability and construct validity. The average variance extracted (AVE) was found to be greater than 0.5, indicating that the latent variables exhibited good convergent validity. Finally, the correlation coefficients between the latent variables were less than 0.8, and the square root of the AVE was greater than the correlation coefficients between the latent variable and the other latent variables (see Table 4), which indicates that the discriminant validity of the scale used is satisfactory.

Measure	**Item**	**factor loading**	**Cronbach's α**	**CR**	**AVE**
Intervention Intention	INT1	0.80	0.866	0.869	0.688
	INT2	0.86			
	INT3	0.84			
Attitude	AT1	0.810	0.832	0.837	0.563

	AT2	0.687			
	AT3	0.726			
	AT4	0.773			
Subjective Norm	SN1	0.848	0.864	0.865	0.681
	SN2	0.795			
	SN3	0.831			
Perceived behavioral control	PBC1	0.799	0.858	0.860	0.606
	PBC2	0.760			
	PBC3	0.789			
	PBC4	0.764			
Moral Norm	MN1	0.715	0.821	0.821	0.535
	MN2	0.757			
	MN3	0.705			
	MN4	0.747			

Table 2 Reliability and convergent validity

	1	**2**	**3**	**4**	**5**
1 Attitude	0.750				
2 Subjective Norm	0.533***	0.825			
3 Perceived Behavioral Control	0.575***	0.552***	0.778		
4 Moral Norm	0.604***	0.504***	0.509***	0.731	
5 Intervention Intention	0.686***	0.570***	0.694***	0.648***	0.829

Table 3 Square root of AVE

4. Results

4.1 Descriptive statistics and correlation analysis

The means, standard deviations, and correlation coefficients of the variables are presented in Table 4. A significant positive correlation was observed between attitude, subjective norm, perceived behavioral control and moral norm, with the willingness to intervene respectively. Regarding to the demographic variables, gender was negatively correlated with subjective norms, while not with willingness to intervene. Age was negatively correlated with attitude and intervention intention. Educational level was positively correlated with moral norms and intervention intention. Finally, experience of cyberhate victimization was positively correlated with attitude, perceived behavioral control, and moral norms.

	M	**SD**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1 Gender			1								
2 Age	2.720	0.873	-0.040	1							
3 Education	3.541	0.886	0.084	0.174*							

4 Victimization	0.665	0.473	0.049	-0.028	-0.267***	1					
1 AT	5.857	0.814	0.041	-0.097*	-0.126**	0.100*	1				
2 SN	5.354	1.075	-0.140**	-0.055	0.079	-0.012	0.533***	1			
3 PBC	5.395	1.003	-0.078	-0.061	0.082	0.097*	0.575***	0.552***	1		
4 MN	5.719	0.857	0.025	-0.044	0.117**	0.127**	0.604***	0.504***	0.509***	1	
5 Intervention Intention	5.702	0.896	0.006	-0.112*	0.124**	0.105*	0.686***	0.570***	0.694***	0.648***	1

Table 4 Correlation Matrix

4.2 Model fit test

The structural equation model was constructed based on the research hypothesis, and the fit of the TPB model and ETPB model were verified respectively, and the results are shown in Table 5. All the fit indicators of the two models meet the requirement, and ETPB model fit better.

Indicator	**Ref.**	**TPB Model**	**ETPB Model**
CMIN/DF	<3	3.232	2.326
RMSEA	<0.08	0.068	0.052
GFI	>0.9	0.931	0.934
AGFI	>0.9	0.900	0.911
CFI	>0.9	0.959	0.966

Table 5 Model fit test of TPB and ETPB model

4.3 Path analysis and hypothesis testing

The AMOS 24.0 was employed to construct a structural equation model, the results are shown in Table 6. TPB model revealed that behavioral attitudes and perceived behavioral control were significant positive predictors of bystanders' intention to intervene, with β-values of 0.577 and 0.526, respectively ($p < 0.01$). In contrast, subjective norm was not a significant predictors of intervention intention. In ETPB model, similar results was obtained, β-values of attitude and perceived behavioral 0.326 and 0.501, respectively. Moral norm had a significant positive predictive effect on the intention to intervene ($\beta=0.330$, $p<0.01$). Therefore, the research hypotheses H1, H3 and H4 verified, and the hypothesis H2 denied.

	TPB model				**ETPB Model**			
	β	**S.E.**	**C.R.**	**P**	**β**	**S.E.**	**C.R.**	**P**
AT→INT	0.557	0.072	8.134	***	0.326	0.076	4.469	***
SN→INT	0.012	0.038	0.239	0.811	-0.030	0.037	-0.624	0.533
PBC→INT	0.526	0.054	7.935	***	0.501	0.050	8.327	***
MN→INT	-	-	-	-	0.330	0.064	4.971	***

Table 6 Path analysis

5 DISCUSSIONS

This study presents a theoretical framework for bystanders' OHS intervention intentions based on ETPB model to help understand the behavior intention from users' cognitive perspective. The results indicated that behavioral attitudes and perceived behavioral control, as TPB model's standard variables, positively predicted users' willingness for intervention (supporting H1 and H3), while subjective norms did not (rejecting H2). Furthermore, moral norms, as an additional variable in ETPB model, also significantly predicted intervention intentions (supporting H4), and even outweighed the predictive role of behavioral attitudes. The study also indicates that bystander intervention intentions for online hate speech are similar but not the same as those observed in cyberbullying and offline bullying scenarios. In cyberbullying related studies, all the three TPB standard variables are significant predictor of the behavior (Hayashi, 2021; Brehmer, 2023), while for OHS, moral norm replacing subjective norm as a key predictor for intervention intention. This may result from the anonymity of the Internet and the widespread dissemination of hate speech, pro-social behaviors based on hate speech are more likely to be purely caused by moral considerations and altruism. Moreover, perceived behavioral control was found to be the most significant predictor on OHS intervention intentions, which was different from cyberbullying context. The assumption for this discrepancy is that cyberbullying is a repeated attack against a specific individual or group, which makes it easier for the platform to identify and intervene. The success rate of bystander intervention is also higher in this case. While there's grey area for hate speech identification, so bystanders must invest more time and efforts to intervene, and they may not only encounter cyberviolence but also be neglected by the SNS providers. Consequently, perceived behavioral control is relatively low in OHS intervention, and is of particular critical in intervention decision-making process.

In addition, experiences related to whether they and their friends had OHS experience were significantly and positively correlated with attitudes, subjective norms, perceived behavioral control and moral norms. This finding is consistent with some previous research (Henson, et al., 2020) that people who have had similar victimization experience are more aware of the harm and consequences of online violence on others, and have a better understanding of when and how to intervene, and therefore have a greater willingness to help others when they are bystanders. Educational level was also found to be significantly associated with moral norms and intention to intervene in the correlation analyses, and future research should investigate the mechanisms underlying such educational differences, which could contribute to the future tailoring of effective prevention and intervention support based on the different educational users, respectively.

6. Limitations and future directions

Based on the ETPB model, this study analyzed the data from 486 social media users' online questionnaires to understand the users' intention to intervene in online hate speech bystander behavior and the predictors, and concluded as follows: ① The ETPB model is applicable to the study of online hate speech bystander intervention intention, and the reliability of the scales and the fit to the model are good, so it can be used to explain the intervention intention of OHS from the perspective of cognition. (ii) Perceived behavioral control, moral norms and behavioral attitudes can play critical roles in predicting bystander intervention behavioral intentions, but the role of subjective norms is not significant. (iii) the predictors of OHS intervention is similar but different from

other types' of offline and online bullying due to event severity and intervention process difference, so deserver SNS providers' to further deep dive to improve the users' intention, which benefits both OHS detection algorithms and users' experience no matter when they are as OHS victims or bystanders.

The following suggestions are made for SNS based on the findings of this study to encourage users' intervention: 1) Strengthen SNS users' understanding on online hate speech definition and intervention. Given that attitudes and moral norms play an important role in predicting users' intervention willingness, SNS providers could provide clear reminder on site and message push when they suspect the users' are facing OHS. Anti-OHS campaign is another way for to enhance public awareness on OHS intervention. (ii) ② Enhance users' self-efficacy for positive intervention. Given that perceived behavioral control is the most important factor in predicting intervention intentions, simplifying the process of OHS reporting or flagging and providing timely feedback and rewards to users after intervention; protecting personal information of interveners in all aspects to avoid cyber-violence against the protectors; and provide counter-speech guidance to empower bystanders' intervention.

This study also has some limitations. First, the study used a cross-sectional study with self-reported intervention intentions, but self-reported intervention intentions sometimes do not yet accurately predict actual intervention behaviors. In Hurd (2022) study, 60% of the respondents who expressed a willingness to intervene finally chose not to send out an intervention post. Suggest to simulate the intervention scenarios on social media platforms to improve the ecological validity of this study. Second, intervention intentions may be influenced by participants' personality and mood (Erreygers, et al., 2016; Price, et al., 2014), and future research suggests combining the ETPB model with these variables improve the predictive model. Thirdly, despite the use of anonymity in this study and the fact that the majority of respondents said online hate speech was more prevalent and therefore non-intervention was acceptable, the possibility of social approval that some respondents chose to respond in line with social expectations still existed. Therefore, future research could further attempt to manage social approval bias through simulation experiments.

Reference

[1] Ajzen, I. The theory of planned behavior[J]. Organizational behavior and human decision processes, 1991, 50(2), 179-211.

[2] Ajzen I. The theory of planned behavior: Frequently asked questions[J]. Human behavior and emerging technologies. 2020,2(4), 314-24.

[3] Branscum P, Rush-Griffin S, Hackman CL, Castle A, Katague M. The role of moral norms as a determinant of intentions to engage in bystander intervention to prevent sexual assault[J]. Journal of community psychology, 2023, 51(1). 334-44.

[4] Bian Q, ChenD. 'Fragile' Intelligence and the 'Torn' World - Defining 'Hate Speech' in Major Western Social Media , Regulation and Algorithmic Dilemma[J]. China Book Review, 2021,(9), 15-32.

[5] Brehmer M. Perceived Moral Norms in an Extended Theory of Planned Behavior in Predicting University Students' Bystander Intentions toward Relational Bullying[J]. European journal of investigation in health, psychology and education, 2023, 13(7), 1202-18.

[6] Conner M, Armitage CJ. Extending the theory of planned behavior: A review and avenues for further research[J]. Journal of applied social psychology, 1998, 28(15), 1429-64.

[7] Cao X, Khan AN, Zaigham GH, Khan NA. The stimulators of social media fatigue among students: Role of moral disengagement[J]. Journal of Educational Computing Research, 2019, 57(5), 1083-107.

[8] Darley JM, Latané B. Bystander intervention in emergencies: diffusion of responsibility[J]. Journal of personality and social psychology, 1968, 8(4p1), 377-83.

[9] DeSmet A, Veldeman C, Poels K, Bastiaensens S, Van Cleemput K, Vandebosch H, De Bourdeaudhuij I. Determinants of self-reported bystander behavior in cyberbullying incidents amongst adolescents[J]. Cyberpsychology, Behavior, and Social Networking. 2014, 17(4), 207-15.

[10] Erreygers S, Pabian S, Vandebosch H, Baillien E. Helping behavior among adolescent bystanders of cyberbullying: The role of impulsivity[J]. Learning and Individual Differences, 2016, 48, 61-7.

[11] Gagliardone I, Gal D, Alves T, Martinez G. Countering online hate speech[R]. Unesco Publishing; 2015, 1-73.

[12] Gelber K, McNamara L. Evidencing the harms of hate speech[J]. Social Identities. 2016, 22(3), 324-41.

[13] Hawdon J, Oksanen A, Räsänen P. Exposure to online hate in four nations: A cross-national consideration[J]. Deviant behavior, 2017, 38(3). 254-66.

[14] Hayashi Y, Tahmasbi N. Psychological predictors of bystanders' intention to help cyberbullying victims among college students: An application of theory of planned behavior[J]. Journal of interpersonal violence. 2022, 37(13-14), NP11333-57.

[15] Henson B, Fisher BS, Reyns BW. There is virtually no excuse: The frequency and predictors of college students' bystander intervention behaviors directed at online victimization[J]. Violence Against Women, 2020, 26(5), 505-27.

[16] Hurd NM, Trawalter S, Jakubow A, Johnson HE, Billingsley JT. Online racial discrimination and the role of white bystanders[J]. American Psychologist, 2022, 77(1), 39.

[17] Kaiser FG, Scheuthle H. Two challenges to a moral extension of the theory of planned behavior: Moral norms and just world beliefs in conservationism[J]. Personality and individual differences, 2003, 35(5), 1033-48.

[18] Lazuras L, Barkoukis V, Ourda D, Tsorbatzoudis H. A process model of cyberbullying in adolescence[J]. Computers in Human Behavior, 2013, 29(3), 881-7.

[19] Leonhard L, Rueß C, Obermaier M, Reinemann C. Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook[J]. SCM Studies in Communication and Media, 2018, 7(4), 555-79.

[20] Lou T, Wang D, Chen H, Niu D. Different perceptions of belief: Predicting household solid waste separation behavior of urban and rural residents in China[J]. Sustainability. 2020, 12(18), 7778.

[21] Masullo Chen G, Lu S. Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments[J]. Journal of broadcasting & electronic media, 2017, 61(1), 108-25.

[22] Naab TK, Kalch A, Meitz TG. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior[J]. New Media & Society. 2018, 20(2), 777-95.

[23] Obermaier M. Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech[J]. New Media & Society, 2022, 14614448221125417.

[24] Paterson JL, Brown R, Walters MA. The short- and longer-term impacts of hate crimes experienced directly, indirectly, and through the media[J]. Personality and Social Psychology Bulletin, 2019, 45(7), 994-1010.

[25] Price D, Green D, Spears B, Scrimgeour M, Barnes A, Geer R, Johnson B. A qualitative exploration of cyber-bystanders and moral engagement[J]. Journal of Psychologists and Counsellors in Schools, 2014, 24(1), 1-17.

[26] Reichelmann A, Hawdon J, Costello M, Ryan J, Blaya C, Llorent V, Oksanen A, Räsänen P, Zych I. Hate knows no boundaries: Online hate in six nations[J]. Deviant Behavior, 2021, 42(9), 1100-11.

[27] Rivis A, Sheeran P, Armitage CJ. Expanding the affective and normative components of the theory of planned behavior: A meta-analysis of anticipated affect and moral norms[J]. Journal of applied social psychology, 2009, 39(12), 2985-3019.

[28] Santre S. Theory of planned behavior in cyberbullying: A literature review[J]. International Journal of Research Reviews in Applied Sciences, 2021, 8(11), 234-9.

[29] Soral W, Malinowska K, Bilewicz M. The role of empathy in reducing hate speech proliferation. Two contact-based interventions in online and off-line settings[J]. Peace and Conflict: Journal of Peace Psychology. 2022, 3, 361.

[30] Weber M, Viehmann C, Ziegele M, Schemer C. Online hate does not stay online–How implicit and explicit attitudes mediate the effect of civil negativity and hate in user comments on prosocial behavior[J]. Computers in human behavior, 2020, 104, 106192.

ABOUT THE AUTHOR

Chengying Yu was born in Shanghai, China, in 1990. She obtained a bachelor’s degree from McGill University in Canada, She currently studying applied psychology at the Institute of Psychology, Beijing. Her main research direction is user experience and behavior psychology.

E-mail: cy_yu66@126.com

Yiwen Chen is from Beijing, China. He was born in 1964. Graduated from Beijing Normal University, he obtained the master of science degree on Applied Mathematics,. He is now working as associate professor in the Division of Social and Engineering Psychology at the Institute of Psychology, Beijing. Main research fields: advertisement & consumer psychology.

E-mail: chenyw@psych.ac.cn