Online disinhibition is not a master key: an examination of online disinhibition mechanisms

Ruohan Wen and Asako Miura
Osaka University, Osaka, Japan

Received 21 July 2024
Revised 1 November 2024
30 December 2024
5 March 2025
Accepted 5 March 2025

Abstract

Purpose – This study aims to empirically examine how online disinhibition influences individuals' behavior on the Internet based on the motivation-based online disinhibition model, a refinement of Suler's (2004) online disinhibition theory. This model suggests that individuals' behaviors are influenced by motivational factors, with online disinhibition moderating this process.

Design/methodology/approach – We observed how individuals expressed their attitudes by clicking "Like" when reading a fictional thread with multiple replies. In Study 1a, participants were shown two types of replies in a thread: inflammatory and normal posts. In Study 1b, we followed up with the participants from Study 1a by asking them how sensational and annoying they perceived each post. In Study 2, we examined whether participants clicked "Like" on a post that contained extreme language.

Findings – In Study 1a, the influence of online disinhibition on Like-clicking did not significantly differ across post types. In Study 1 b, when participants perceived posts as sensational or annoying, they were more likely to click "Like" if they experienced high levels of online disinhibition. For posts that were neither sensational nor annoying, online disinhibition did not affect the likelihood of clicking "Like." In Study 2, although online disinhibition was significantly correlated with Like-clicking on the extreme post, this effect was not observed among participants who held a negative attitude toward the extreme post.

Originality/value – This study rigorously examined the causal relationships among motivational factors, online disinhibition and behaviors, challenging previously overgeneralized explanations of online disinhibition theory regarding online behaviors.

Keywords Online disinhibition, Motivation-based online disinhibition model, "Like" button, Attitude expression, Social media

Paper type Research paper

1. Introduction

People sometimes behave on the Internet in ways they would not behave in real life (Walther, 1996). For example, some people may exhibit deceptive, destructive behavior online to cause chaos, also known as "online trolling" (Buckels *et al*., 2014). Bargh *et al.* (2002) argued that the Internet allows people to share intimate personal details with strangers they may never see again in anonymous online environments. Suler (2004) interpreted such differences between real and online environments as the results of the *online disinhibition effect*. In contrast to inhibition, where behavior is constrained or restrained through self-consciousness, anxiety about social situations, and worries about public evaluation, disinhibition is characterized by an absence or reversal of these factors (Joinson, 2007).

In today's digital era, the exponential rise in Internet usage has necessitated a deeper understanding of online users' unique features. Online disinhibition theory elucidates Internet users' psychological and behavioral features from the perspective of "inhibition release," influencing various academic disciplines for nearly two decades (Cheung *et al*., 2021). It has

Internet Research
Vol. 35 No. 7, 2025
pp. 52-70
Emerald Publishing Limited
e-ISSN: 2054-5657
p-ISSN: 1066-2243
DOI 10.1108/INTR-07-2024-1143

This study was supported by the JSPS KAKENHI Grant Number JP23KJ1483 and approved by the university's Ethics Committee (approval date: March 7, 2023, approval number: HB022-120).

primarily been linked to negative behaviors, such as cyberbullying (Udris, 2014; Wright *et al.*, 2019), cyber aggression (Kurek *et al.*, 2019), and sexual-erotic online behavior (Hernández *et al.*, 2021). Additionally, online disinhibition can also promote positive behaviors, such as self-disclosure (Hollenbaughand Everett, 2013; Lapidot-Lefler and Barak, 2015; Schouten *et al.*, 2007). Moreover, understanding online disinhibition can aid Internet professionals in recognizing how attributes such as anonymity and invisibility influence deviant behaviors online. This can help adjust system designs so that such behaviors can be mitigated by monitoring users' levels of online disinhibition (Cheung *et al.*, 2021).

Though online disinhibition has widely established links with various unique online behaviors, the specific mechanism through which online disinhibition influences these behaviors remains ambiguous. Accordingly, this study aims to investigate the research question: How does online disinhibition influence online behavior? To address it, we first established a motivation-based online disinhibition model (Wen and Miura, 2023) as the theoretical framework, drawing from existing research. We then focused on Internet users' Like-clicking behavior on social media, conducting two experiments to examine the effect of online disinhibition on users' Like-clicking behavior. This study enriches the literature by providing empirical evidence of the influence mechanism of online disinhibition, refining its theoretical framework, and improving the understanding of online behavior.

2. Theoretical background

2.1 The mechanisms of online disinhibition

2.1.1 Benign–toxic disinhibition model. We first summarize the main theoretical frameworks from previous research. The first widely used framework is the "benign–toxic disinhibition model." Suler (2004) divided online disinhibition into two distinct parts based on its results: benign and toxic disinhibitions. Benign disinhibition produces positive results (such as self-disclosure and prosocial behavior), whereas toxic disinhibition produces negative results (such as cyberbullying and online trolling). Based on this naïve theoretical framework, the concept of online disinhibition has been widely used to explain various online behaviors and phenomena. For example, Lapidot-Lefler and Barak analyzed online flaming and prosocial behaviors through toxic and benign disinhibition perspectives, respectively (Lapidot-Lefler and Barak, 2012, 2015). Udris (2014) developed an online disinhibition scale that includes two factors: benign and toxic disinhibition, and investigated their association with cyberbullying. Further, Kordyaka *et al.* (2020) utilized this scale to examine the relationship between online disinhibition and toxic behaviors in video gaming.

Nevertheless, researchers have recently highlighted the conceptual confusion within this model. Stuart and Scott (2021) argued that online disinhibition itself should not be distinguished as positive or negative according to its results, as it merely represents the mental state in which individuals experience acting, thinking, or feeling differently online when compared to face-to-face interactions. Wen and Miura (2023) further argued that disinhibition can only be considered good (benign) or bad (toxic) when it manifests as good or bad behavior. That is, the benign–toxic disinhibition model, which posits that benign and toxic disinhibitions induce positive and negative behaviors respectively, represents a circular argument.

2.1.2 Disinhibition–behavior model. To address this issue, Stuart and Scott (2021) thoroughly refined the concept of online disinhibition, differentiating it more clearly from its potential outcomes. They also proposed a simplified model, linking online disinhibition directly to behaviors such as self-disclosure, trolling, and cyberbullying. Schouten *et al.* (2007) similarly revealed that online disinhibition directly promoted online self-disclosure, and Kurek *et al.* (2019) showed that it was a significant positive predictor of cyber aggression. These approaches, which establish a direct causal link from online disinhibition to certain behaviors, could be generalized as the disinhibition−behavior model (Wen and Miura, 2023).

While this model provides an intuitive and broadly applicable framework, it remains theoretically incomplete. For example, this model implies that various behaviors, such as self-

disclosure and online trolling, will automatically manifest when individuals are influenced by online disinhibition. However, given the pervasive use of the Internet in modern societies, assuming that behaviors on the Internet are merely passive reactions influenced by the online environment and experiences of online disinhibition is unrealistic. A more precise perspective could be that individuals actively use the Internet to fulfill their specific objectives or needs. For example, in the context of Internet addiction, it is often contended that individuals exhibiting Internet addiction intentionally indulge in the virtual world to escape reality (Chou *et al.*, 2005). This demonstrates that the source of diverse disinhibitory behaviors on the Internet should be the subjective motivation of individuals seeking a particular purpose rather than the online environment or online disinhibition itself. Given its neglect of the influence of subjective motivation, the disinhibition–behavior model requires further refinement.

2.1.3 Motivation-based online disinhibition model. Considering these limitations, Wen and Miura (2023) proposed an improved model called the motivation-based online disinhibition (MOD) model. It posits that individuals' motivations—whether generated intrinsically (e.g. social needs or protection of their self-esteem) or aroused by extrinsic stimuli (e.g. being rebuked at work or meeting unreasonable others on the Internet)—are the determining factors that motivate them to exhibit certain behaviors, such as self-disclosure on the Internet to relieve their stress or catharsis by abusing others. Whether these motivations could be transformed into behaviors is moderated by online disinhibition. When individuals exhibit high online disinhibition levels, their motivations are more likely to transform into behaviors. Put differently, only individuals who are motivated to display certain behaviors are more likely to act on them when they experience high online disinhibition. Conversely, if individuals have no motivation to display certain behaviors, they will not suddenly act accordingly because they are online or experiencing high online disinhibition.

The MOD model assumes that individuals do not passively act in disinhibitory ways because they are impacted by online disinhibition; rather, they intentionally deploy various online services, which provide a more comprehensive theoretical framework for understanding Internet users' behaviors. Having established this theoretical model, we now turn to employing an experimental approach to fill the empirical evidence gap in the MOD model and to verify its validity by comparing it to the disinhibition–behavior model.

2.2 The "like" button in social media

To achieve this aim, we focused on a lightweight communication act of clicking "Like" on an online platform. The "Like" button has become a widely utilized feature on traditional social media platforms, such as Facebook and X (formerly Twitter), as well as news websites and their respective comment sections. Hayes *et al.* (2016) conceptualized the "Like" button as a paralinguistic digital affordance (PDA) that facilitates communication and interaction without specific language associated. In this study, we focus on the following two aspects of Like-clicking. First, we consider clicking "Like" as expressing attitudes in the online community. Eranti and Lonkila (2015) highlighted that the ease of the "Like" function significantly reduces the threshold for online political participation, allowing users to conveniently express their support toward certain content. Hayes *et al.* (2016) argued that clicking "Like" enables users to convey slight positive attitudes, such as "subtle recognition" or "affirmation of someone's post." Consequently, the frequency of clicking "Like" can reflect the degree of expressed positive attitudes. In this context, we conceptualized Like-clicking as *expressions of positive valence opinions*.

Second, we consider how the object of "Like" affects the interpretation of Like-clicking. On the Internet, extreme expressions and hateful speech that are rarely encountered in real life (Castaño-Pulgarín *et al.*, 2021) can sometimes escape universal condemnation and may even gather "Likes" from certain individuals. Clicking "Like" on such content may represent an indirect form of expression. In this context, we conceptualized clicking "Like" on content that is extreme or contradicts general values as *expressions of deviant opinions*.

The online disinhibition theory and disinhibition–behavior model indicate that the Internet's relaxed atmosphere encourages individuals to express their opinions, thereby facilitating both types of Like-clicking behavior. This study conducted more detailed experiments to examine how online disinhibition and motivation influence two types of Like-clicking and to compare the insights of the MOD model with the disinhibition–behavior model.

3. Hypotheses

We preregistered two studies based on the MOD model. The first study examined the influence of different extrinsic stimuli on using "Likes" as *expressions of positive valence opinions*. On social media, we sometimes witness rude language, harsh criticisms, anger, hatred, and even threats (Suler, 2004). We defined the posts containing such aggressive, offensive, or provocative language as *inflammatory content*. Inflammatory content is not always merely an outlet for emotions or meaningless swear words; sometimes, its extremity can also resonate with the viewpoints of specific groups and attract them (Zimmerman *et al.*, 2024). When individuals encounter inflammatory content that they agree with, the extreme and deviant wording against social desirability will stimulate cognitive judgment about whether it is appropriate to interact with them. Those with low online disinhibition may recognize the inappropriateness of interacting with such posts due to concerns about social desirability or potential damage to their image, thereby restraining their desire to click "Like"; conversely, individuals with high online disinhibition are less restrained by these considerations and therefore more likely to click "Like." Therefore, an individual's use of "Like" to express positive valence opinions on inflammatory content will be significantly influenced by their level of online disinhibition.

However, online posts are not always highly inflammatory or about challenging social values. We defined ordinary or commonplace posts as *normal content*. As these posts contain less information that conflicts with general values, interacting with them does not typically activate concerns, such as social desirability or personal reputation. Consequently, whether individuals have high or low online disinhibition would not significantly impact their likelihood of clicking "Like." Based on these considerations, we propose the following hypothesis:

H1. Online disinhibition has a stronger effect on clicking "Like" for inflammatory posts compared to normal posts.

The first study examined the influence of objective features of content on Like-clicking. However, people may exhibit different intrinsic reactions (i.e. emerging with different intrinsic motivations) to the same content based on their individual traits. Therefore, the second study examined the impact of individuals' intrinsic motivations on using "Likes" as *expressions of deviant opinions*. The target post focused on an inflammatory post containing extreme language. According to the MOD model, we hypothesized that a positive attitude toward the extreme post would act as intrinsic motivation, determining an individual's likelihood of clicking "Like" on it. This process is moderated by online disinhibition: individuals under low online disinhibition might restrain from Like-clicking because they recognize the deviant nature of the post, whereas those under high online disinhibition are more likely to click "Like." Conversely, those who view the post negatively are less likely to click "Like" due to a lack of motivation, regardless of their level of disinhibition. Thus, we propose the following hypotheses:

H2a. Individuals' Like-clicking on an extreme post is determined by their attitudes toward it.

H2b. Online disinhibition amplifies the likelihood of clicking "Like" on an extreme post for individuals who hold a positive attitude toward it.

Statistical analyses in this study were conducted using R, and the R code and the appendix material are available at https://osf.io/fwmjr. The preregistration information is available at https://osf.io/enrjx.

4. Study 1a

4.1 Study design

The present study proposed an improved method for determining participants' Like-clicking. Here, we utilized the "hotspot" function in Qualtrics to create more authentic experiences of clicking "Like". We drew long images of threads containing one initial post and multiple reply posts. In each reply post, we drew a thumb icon and set an invisible hotspot on it in Qualtrics. As participants clicked on the position of the thumb icon, the hotspot was activated, switching to a visible, translucent green state. This means clicking "Like" on the post. Participants could click on any activated hotspot again and it would return to being invisible. This means withdrawing "Like". This method makes participants feel more like they are performing real Like-clicking on social media, equipping our study with better ecological validity than a traditional questionnaire.

To provide participants with a broad spectrum of posts, particularly those involving highly inflammatory and deviant content, we utilized a highly topical and controversial political discussion as the theme of the thread. According to Fine and Hunt (2023), social media has become a common tool for discussing political topics, particularly negative messages and political attacks exerting stronger influences. In March 2023, the Japanese government introduced a policy about culling cows due to milk overproduction. At the time, the news about a high school introducing edible cricket powder into its school meals stirred significant controversy across the Japanese Internet. Given that Japan does not have a tradition of eating crickets or insects, introducing such food into the school diet for minors was deemed unacceptable to most Japanese. Subsequently, cricket-eating even became a meme, and many individuals expressed their dissatisfaction with the government and politicians.

Given that both issues were related to food, we utilized the cow-culling policy and cricket-eating meme to compose the thread. We drew specific content from real discussions on the Japanese anonymous online forum 5ch. The initial post introduced the cow-culling policy. We manipulated the contents of the reply posts according to the operational definitions of inflammatory and normal posts. For the inflammatory posts, we selected nine posts that contained sarcastic, radical, and harsh language within the scope of ethical permission. For the normal posts, we selected nine that contained usual, conventional, and safe comments. The two post types were presented alternately, and participants could click "Like" on any post they wished to. According to H1, online disinhibition would exert a stronger effect on the count of "Likes" on inflammatory posts rather than on normal posts.

4.2 Method

We preregistered the following survey and conducted it on September 20, 2023. We entrusted the crowdsourcing company CrowdWorks Co., Ltd., with the task of recruiting Japanese general Internet users aged 18–70 years without imposing restrictions on gender or education level. Before the survey, we specified that it would be conducted anonymously and not involve any questions violating participants' privacy. Next, we obtained participants' consent before proceeding with the survey; participants could stop at any time during the survey. As the thread featured some extreme content, we explained the purpose of the study and asked participants for their consent again after completing the survey. The answers provided by participants who disagreed with the use of the data were deleted, even though they were still entitled to the reward. The survey lasted approximately 6.5 min, and each participant received 100 JPY as a reward. The survey in Study 1a was conducted in the following steps:

4.2.1 The frequency at which participants utilize online bulletin boards or news sites. Participants were required to rate their frequency of using online bulletin boards or news sites

in daily life on a scale from 1 (almost never) to 10 (very frequently). Hereinafter, it is referred to as InfoPlatforms usage rate.

4.2.2 MMOD and the directed question scale. We employed the multidimensional measure of online disinhibition (MMOD; Wen and Miura, 2024) to assess participants' online disinhibition levels. As an improved scale of the measure of online disinhibition (Stuart and Scott, 2021), MMOD comprises three factors: the "unique perspective on online environment," "change of alienation cognition," and "change of relationship cognition," which could measure online disinhibition more comprehensively than the previous one. Participants were required to answer MMOD items from 1 (strongly disagree) to 6 (strongly agree); they were also given an "I don't know" option (which would be coded as a missing value). A directed question scale (DQS; Maniaci and Rogge, 2014) was set at the end of MMOD, requiring participants to select "I don't know".

4.2.3 Introduction of the event and attitudes toward cricket-eating. A part of a news article from The Nikkei (2022) was quoted to introduce the cricket-eating incident. After reading the news, participants were asked to answer four items (e.g. "I have a reluctance to eat crickets" as a reversal item) regarding their attitudes toward cricket-eating, from 1 (strongly disagree) to 7 (strongly agree). These items were created in the preliminary survey. Hereinafter, the average of the four items will be referred to as attitude toward cricket-eating.

4.2.4 Practice and reading of the fictitious thread. We informed participants that they would be required to read a thread. Before formally reading the post, we provided participants with instructions and practice opportunities. We conveyed the following instructions to participants: "There are "Like" buttons below each post. Click on them if you wish to. However, if you do not commonly do it, it is not obligatory to click any "Like" button." Next, participants were required to click the "Like" button once in the practice session. Subsequently, participants were required to read the fictitious thread we created.

4.2.5 Demographic questions. Finally, participants were required to answer demographic questions including age, gender, and education level.

4.3 Results

4.3.1 Descriptive analysis. A total of 410 individuals completed the survey. According to the preregistered exclusion criteria, we list-wise deleted data from 28 participants due to incorrect DQS answers or missing values in MMOD or demographic items. Additionally, we preregistered a criterion to exclude participants who spent less than 60 s reading the thread, which was expected to result in the exclusion of 15.5% of the data. Upon reevaluation, we realized that the anticipated reading time was overestimated and that the thread could be thoroughly read in approximately 45 s. Consequently, we adjusted the criterion to a 45-s threshold, resulting in the exclusion of 37 participants (9.7%). The final sample included 345 participants (M_{age} = 40.61, SD = 10.34; 64.1% female). The structural validation of the MMOD revealed acceptable model fit indices (Table S1). We calculated the total number of "Likes" for both post types and performed a descriptive analysis. Table 1 presents the correlation coefficients, descriptive statistics, and Cronbach's αs.

4.3.2 HLM of Like-clicking. Given that participants read both inflammatory and normal posts in a thread, which created a multi-level data structure with within-participants (post types) and between-participants (MMOD, demographics, and InfoPlatforms usage rate) variables, a hierarchical linear model (HLM) was considered an appropriate analytical approach (McCoach, 2010). We preregistered the following HLM to test H1. We used the number of "Likes" as the dependent variable. We aimed to examine the general effects of the MMOD, post types, the interaction between MMOD and post types, demographic variables, and InfoPlatforms usage rate. Therefore, these variables were employed as fixed effects. The individual ID, representing inter-individual differences, was employed as a random effect (cf. McCoach, 2010). Table 2 presents the standardized HLM coefficients, and Figure 1 presents the interaction effect between MMOD and the post types. The findings revealed a significant

Table 1. Descriptive statistics and correlation matrix of study 1a

	1	2	3	4	5	6	7
1. Likes on inflammatory posts							
2. Likes on normal posts	0.49**						
3. MMOD	0.13*	0.15**					
4. InfoPlatforms usage rate	0.06	0.07	0.05				
5. Attitude toward cricket-eating	−0.23**	−0.02	−0.10	0.01			
6. Age	−0.09	−0.10	−0.09	−0.04	0.10		
7. Gender	−0.10	−0.11*	−0.12*	0.07	−0.17**		
8. Education level	−0.03	−0.05	−0.02	0.09	−0.02	−0.05	−0.09
M	0.79	1.84	3.41	7.12	3.04	40.61	
SD	1.36	1.69	0.54	2.06	1.26	10.34	
α			0.72		0.86		

Note(s): MMOD: Multidimensional measure of online disinhibition, Gender (0 = Male, 1 = Female), Education level (0 = Below bachelor's degree, 1 = Bachelor's degree or above), *$p < 0.05$, **$p < 0.01$
Source(s): Authors' own creation/work

Table 2. HLM results

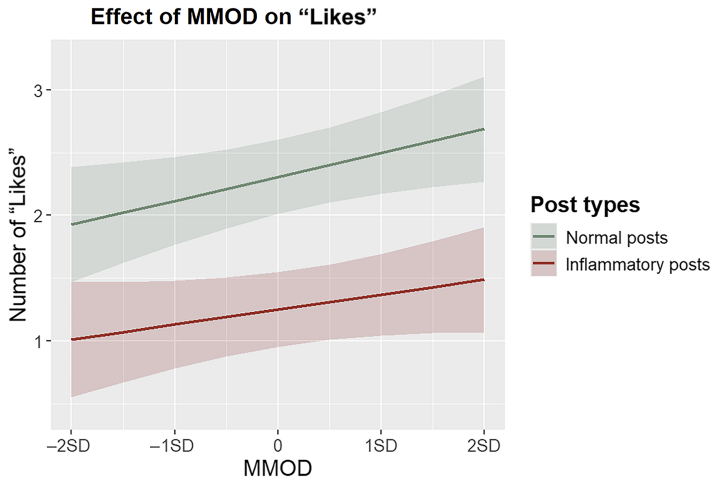
Variable	Estimate	Std. error	*t*-value	*p*-value
(Intercept)	−0.04	0.09	−0.43	0.67
MMOD	0.07	0.05	1.45	0.15
Post types (=1)	0.65	0.05	12.50	<0.01
InfoPlatforms usage rate	0.05	0.04	1.27	0.20
Attitude toward cricket-eating	−0.11	0.04	−2.60	0.01
Age	−0.10	0.04	−2.23	0.03
Gender (=1)	−0.27	0.09	−2.96	<0.01
Education level (=1)	−0.19	0.09	−2.19	0.03
MMOD × Post types	0.04	0.05	0.85	0.40

Note(s): MMOD: Multi-dimensional measure of online disinhibition, Post types (0 = Inflammatory posts, 1 = Normal posts), Gender (0 = Male, 1 = Female), Education level (0 = Below bachelor's degree, 1 = Bachelor's degree or above)
Source(s): Authors' own creation/work

main effect of the post types ($p < 0.01$), indicating that the normal posts received more "Likes" than the inflammatory ones. However, the main effect of MMOD ($p = 0.15$) and the interaction effect between MMOD and the post types were insignificant ($p = 0.40$), indicating no significant difference in the effects of online disinhibition across the two post types. This result does not support H1, which posited that the effect of online disinhibition on interacting with inflammatory posts would be stronger.

4.4 Discussion

In Study 1a, the HLM revealed an insignificant difference in the influence of online disinhibition on the "Likes" of two types of posts, which contradicted the predictions of the MOD model. We attributed this to the unrefined manipulation of motivations. To conceptualize the motivational factors, we artificially categorized 18 posts into inflammatory and normal posts based on their content. We assumed that the inflammatory posts could serve as stronger stimuli that promote participants' disinhibitory motivation. However, this broad and rough categorization might be inadequate to exactly reflect the stimulation degree of each post to the participants. Not all inflammatory posts necessarily



Source(s): Authors' own creation/work

Figure 1. Interaction effect between MMOD and post types

provided strong extrinsic stimuli to the participants. For example, despite conveying an intense tone, Post 11 expressed somewhat reasonable views, with its intensity being slightly lower than those of other inflammatory posts. Similarly, some of the normal posts might have aroused the participants' sentiments. Post 6, which exhibited a significant relationship with MMOD ($r = 0.16$, $p < 0.01$), mentioned the issue of potential tax increases, a topic of widespread concern among Japanese citizens, which might have stimulated the participants to click "Like." Furthermore, even for the same post, participants' perceptions could vary greatly among individuals. In Study 1a, the experimental manipulation oversimplified the complexity of the posts' impact and the diversity of participants' reactions, which ultimately contributed to the failure to capture their disinhibitory motivation accurately. Therefore, we suggest that future research should be aware of the diversity in Internet users' responses to certain content and focus on assessing their motivations from a subjective perspective. In the next phase, Study 1b, we will attempt to address this issue by incorporating participants' subjective perceptions of each post to reevaluate the inflammatory nature of the posts and assess their disinhibitory motivation more accurately.

5. Study 1b

5.1 Study design

Considering that post types fail to represent the realistic degree of individuals' disinhibitory motivations, we collected additional data on post characteristics based on Study 1a to further explore other features that could conceptualize motivation on Like-clicking. In Study 1 b, we reconsidered the features that characterize posts as "*inflammatory*" from two perspectives considering our specific experimental context. First, we hypothesized that the presence of sensational language in inflammatory posts can elicit strong emotions, potentially serving as a motivational factor for Like-clicking. Next, following Seigner *et al.* (2023), who posited that provocative language online could, paradoxically, foster audience engagement, we hypothesized that the unfriendly tones and extreme language featured in inflammatory posts would provoke annoyance. While annoyance is typically considered a negative emotional valence, it can also predict participants' emotional arousal, which serves as a motivational factor for Like-clicking. In Study 1 b, we sent a follow-up survey invitation to the

participants in Study 1a to gather their perceptions of how *sensational* and *annoying* each post was.

As the participants' perceptions of each post were available, we performed an in-depth analysis of their Like-clicking for each post. Our approach involved employing a generalized linear mixed model (GLMM, cf. Moscatelli *et al.*, 2012) in which whether participants clicked "Like" on each post served as the dependent variable. The independent variables included MMOD, post perceptions, the interaction effect between MMOD and the post perceptions, and other control variables.

5.2 Method

On September 29, 2023 (approximately nine days after completing Study 1a), we sent follow-up survey invitations to the participants whose data were included in the analyses via the private messaging function on CrowdWorks. The survey lasted for approximately 4 min and participants received 60 JPY as a reward. Within four days (September 29 to October 2), 320 participants (92.6% response rate) had responded to our invitation.

In the questionnaire, we once again presented the news cited from *The Nikkei* and the initial post from the thread, followed by filler items such as participants' interest in the topic. Next, we presented each reply post sequentially and inquired about participants' perceptions of each post when they initially encountered it. This included how sensational the post was, from 1 (not sensational at all) to 7 (very sensational), and how annoying the post was, from 1 (not annoying at all) to 6 (very annoying).

5.3 Results

We merged the datasets of Studies 1a and 1b based on each participant's unique response ID. As the categorical variable failed to adequately capture participants' motivation levels, we shifted our focus to using their sensational and annoying scores for each post as predictive factors. We employed two GLMMs. The dependent variable was whether participants clicked "Like" on each post. The independent variables included MMOD, sensational (or annoying) scores, the interaction effect between MMOD and sensational (or annoying) scores, attitudes toward cricket-eating, age, gender, education levels, and InfoPlatforms usage rate. The individual ID, representing inter-individual differences, was set as a random effect.

However, the initial GLMMs failed to converge. To address this, we excluded education level due to the underrepresentation of non-college samples, as well as age and InfoPlatforms usage rate due to their relatively small coefficients. Using the revised GLMMs with the remaining variables, the models successfully converged. To compare the different models, we performed three-step hierarchical GLMMs. According to the disinhibition−behavior model, Step 1 included MMOD, attitude toward cricket-eating, and gender as independent variables. Step 2 introduced the sensational (or annoying) scores. Step 3 introduced the interaction effect between MMOD and the sensational (or annoying) scores, following the MOD model.

The results of hierarchical GLMMs in the analyses of the sensational (referred to as the sensational model) and annoying (referred to as the annoying model) scores are presented in Tables 3 and 4. Numbers in parentheses next to the coefficients represent their standard deviations. In both hierarchical approaches, we observed stepwise decreases in the AIC and BIC values from Steps 1 to 2 and Steps 2 to 3. This trend indicated a continuous improvement in the model fit. Specifically, the analysis of variance comparing the likelihood ratios between Steps 2 and 3 revealed that the inclusion of the interaction effects significantly enhanced the performance of the model (both p-values <0.01). Consequently, these findings indicated that the MOD model offered higher explanatory power than the disinhibition−behavior model in statistics.

Moving forward, we focused on the models from Step 3 in both hierarchical approaches. In both models, MMOD exhibited significant positive effects, with coefficients of 0.21 ($p < 0.01$) and 0.32 ($p < 0.01$). These results indicated that a higher online disinhibition

Table 3. Sensational model

Independent variables	Dependent variable: whether individuals clicked "Like" on each post Step 1	Step 2	Step 3
(Intercept)	−1.98 (0.12)**	−2.06 (0.13)**	−2.06 (0.13)**
MMOD	0.18 (0.07)*	0.18 (0.08)*	0.21 (0.08)**
Attitude toward cricket-eating	−0.13 (0.07)	−0.14 (0.08)	−0.14 (0.08)
Gender (=1)	−0.35 (0.15)*	−0.33 (0.16)*	−0.35 (0.16)*
Sensational		−0.42 (0.04)**	−0.43 (0.05)**
MMOD × Sensational			0.13 (0.04)**
AIC	4339.59	4251.71	4244.01
BIC	4372.88	4291.67	4290.62

Note(s): Numbers in parentheses indicate standard error. MMOD: Multi-dimensional measure of online disinhibition, Gender (0 = Male, 1 = Female), AIC: Akaike information criterion, BIC: Bayesian information criterion, *$p < 0.05$, **$p < 0.01$
Source(s): Authors' own creation/work

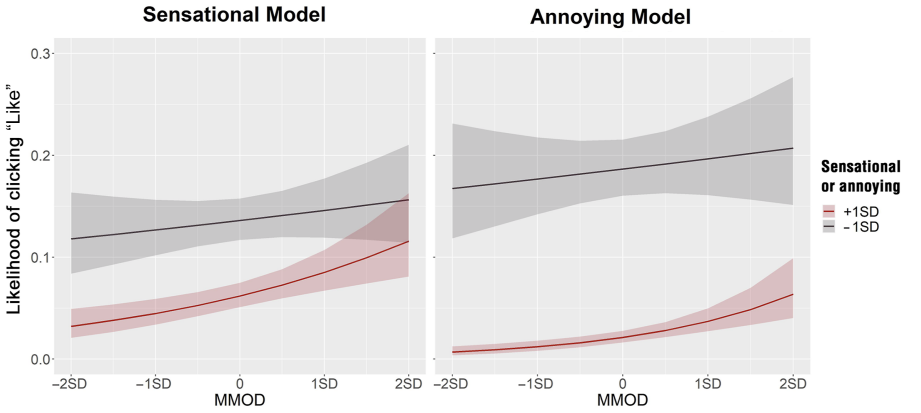
Table 4. Annoying model

Independent variables	Dependent variable: whether individuals clicked "Like" on each post Step 1	Step 2	Step 3
(Intercept)	−1.98 (0.12)**	−2.39 (0.15)**	−2.43 (0.15)**
MMOD	0.18 (0.07)*	0.18 (0.09)*	0.32 (0.09)**
Attitude toward cricket-eating	−0.13 (0.07)	−0.12 (0.09)	−0.13 (0.09)
Gender (=1)	−0.35 (0.15)*	−0.33 (0.18)	−0.35 (0.18)*
Annoying		−1.11 (0.06)**	−1.18 (0.06)**
MMOD × Annoying			0.25 (0.05)**
AIC	4339.59	3838.53	3817.79
BIC	4372.88	3878.48	3864.40

Note(s): Numbers in parentheses indicate standard error. MMOD: Multi-dimensional measure of online disinhibition, Gender (0 = Male, 1 = Female), AIC: Akaike information criterion, BIC: Bayesian information criterion, *$p < 0.05$, **$p < 0.01$
Source(s): Authors' own creation/work

correlated with an increased likelihood of Like-clicking. Conversely, the sensational and annoying scores exhibited significant negative effects, with coefficients of −0.43 ($p < 0.01$) and −1.18 ($p < 0.01$). This indicated that people were less inclined to express a positive attitude toward deviant content than more conventional content. These results correlated with the theoretical and empirical insights of previous studies (e.g. Wen and Miura, 2023), indicating that only a few individuals engaged in deviant behaviors.

Importantly, the interaction effects between MMOD and the sensational (or annoying) scores in the models exhibited significant positive effects on Like-clicking, with coefficients of 0.13 ($p < 0.01$) and 0.25 ($p < 0.01$). Figure 2 illustrates these interaction effects. These results indicate that online disinhibition exerts a stronger effect on promoting Like-clicking when participants perceive posts as sensational or annoying. For posts that participants did not perceive as sensational or annoying, strong online disinhibition did not lead to a higher likelihood of Like-clicking compared to weak online disinhibition. Put differently, strong online disinhibition did not necessarily induce an indiscriminate increase in clicking "Like" to express positive attitudes, contradicting the prediction made by the disinhibition–behavior model. Rather, an individual's perception of the content plays a crucial role. When individuals perceived a post as strongly inflammatory, they recognized that clicking "Like" on such content was not desirable. In such cases, online disinhibition moderates the transformation of



Source(s): Authors' own creation/work

Figure 2. Interaction effect between MMOD and the sensational or annoying scores

Like-clicking motivation into actual behavior, with stronger online disinhibition increasing the likelihood of Like-clicking. In summary, these findings essentially support H1, which posits that online disinhibition would exert a stronger effect on interacting with posts when the target was perceived as inflammatory. They also validate the MOD model's perspective that online disinhibition moderates disinhibitory motivation in the occurrence of online behaviors rather than directly determining online behavior.

5.4 Discussion

Study 1b gathered more detailed data on the participants' perceptions toward each post and used these data to predict motivation. Although the participants' sensational and annoying scores were collected after they had engaged in the Like-clicking decisions, which might raise questions about rigorously establishing causality in predicting motivations for Like-clicking, we argue that the objective characteristics of the posts, such as wordings and expressions, predominantly shaped these perceptions. Consequently, these perceptions were unlikely to significantly fluctuate over time or be influenced by Like-clicking at that moment.

Regarding the significant interaction effect between MMOD and the sensational or annoying scores, we derived a conclusion that generally supported the MOD model. When the participants experienced strong inflammatory feelings toward certain content, higher online disinhibition predicted a greater likelihood of attitude expression by clicking "Like." This outcome complemented a more intuitive and common-sense explanation for online disinhibition. Disinhibition refers to the reduction or disappearance of inhibition, which means that its effect becomes noticeable only when individuals act contrary to their usual inhibitions. In our studies, clicking "Like" on posts characterized as sensational or annoying represented such behavior. Conversely, when a behavior does not contravene social norms or provoke negative feelings—such as clicking "Like" on commonplace or safe posts—the presence or absence of the disinhibition effect in an individual does not significantly influence their behaviors.

However, Studies 1a and 1 b had certain limitations. The MOD model indicated that stronger motivation correlated with a higher likelihood of expressing an attitude by clicking "Like." Given that the sensational and annoying scores negatively predicted the Like-clicking tendency, they might not accurately represent participants' true motivations. Furthermore, the data-collection sequences in Studies 1a and 1 b presented a challenge in establishing a strict

causal relationship between motivations and Like-clicking from a theoretical standpoint. Consequently, the experimental design must be further refined and optimized to address these issues.

6. Study 2

6.1 Study design

We preregistered a new experiment to test H2a and H2b and examined the MOD model using a stricter process. We adapted the materials from Study 1a to create a new thread containing one initial post and seven replies. Among these replies, Post 5 used particularly extreme language to criticize the government. We focused on participants' "Likes" on this extreme post. To minimize the potential order effect, the second to fourth posts were intentionally designed to contain less substantive content.

To address the insufficient rigor of examining causality in Studies 1a and 1 b, we adjusted the thread-reading process. Before participants made a Like-clicking decision on the extreme post, a series of questions about their attitudes toward the subject were asked. Through this adjustment, we could rigorously examine the causal effect of participants' attitudes toward Like-clicking.

6.2 Method

In Study 2, the questionnaire was identical to that of Study 1a, except for the thread-reading stage. In this stage, the thread was divided into two parts. Participants started by reading the first part, which contained the initial post and Posts 2 to 5. Posts 2 to 4 were similar to those in Study 1a, containing content and "Like" buttons. Post 5, the extreme post, was only presented with its content, excluding the "Like" button.

Participants could click the "next page" button to proceed. On the next page, we utilized nine items developed in the preliminary survey to assess participants' attitudes toward the extreme post (e.g., "I feel resonant with it."), rated on a scale of 1 (strongly disagree) to 4 (strongly agree). After responding, participants continued to the second part of the thread, which displayed Post 5 to the final post, with each post presented with content and "Like" buttons.

The survey was conducted on October 23, 2023, using the same recruitment method, ethical procedure, and consent acquisition process as in Study 1a. The survey lasted for approximately 6 min and participants received 90 JPY each as a reward.

6.3 Results

6.3.1 Descriptive analysis. A total of 592 individuals completed the survey. According to the preregistered exclusion criteria, we list-wise deleted data from 37 participants due to incorrect DQS answers or missing values in MMOD or demographic items, as well as data from 137 participants who spent less than 30 s reading the first part of the thread. Additionally, two respondents who clicked all "Likes" buttons were excluded as outliers, since such behavior is exceedingly rare in real scenarios.

These exclusions, particularly removing participants with reading times under 30 s, could introduce bias. Participants with short reading times might represent users who are uninterested in the topic and quickly skim through content in real-world settings. Nevertheless, these exclusions ensure that the remaining data reflects samples that have thoroughly read the thread, allowing us to rule out the possibility that the absence of Like-clicking on extreme posts is due to non-engagement or insufficient reading time. This approach enables a more accurate investigation of the effect of online disinhibition on Like-clicking behavior.

After exclusions, 425 valid responses remained (M_{age} = 40.93, SD = 10.70; 62.8% female). Table 5 presents the correlation coefficients, descriptive statistics, and Cronbach's α

Table 5. Descriptive statistics and correlation matrix of study 2

	1	2	3	4	5	6	7
1. Like on the extreme post							
2. Attitude toward the extreme post	0.48**						
3. MMOD	0.13**	0.15**					
4. InfoPlatforms usage rate	0.06	0.08	0.13**				
5. Attitude toward cricket-eating	−0.22**	−0.25**	−0.13**	−0.03			
6. Age	−0.09	−0.10	−0.09	−0.04	0.10		
7. Gender	−0.08	−0.04	−0.19**	−0.01	−0.13**		
8. Education level	−0.01	0.02	0.04	0.02	−0.05	−0.04	0.05
M		2.24	3.33	7.10	3.25	40.93	
SD		0.64	0.56	2.00	1.17	10.70	
α		0.93	0.73		0.82		

Note(s): Like on the extreme post (0 = No, 1 = Yes), MMOD: Multi-dimensional measure of online disinhibition, Gender (0 = Male, 1 = Female), Education level (0 = Below bachelor's degree, 1 = Bachelor's degree or above), *$p < 0.05$, **$p < 0.01$
Source(s): Authors' own creation/work

values of the variables. To assess the robustness of the exclusion criteria, a sensitivity analysis was conducted using data from 555 participants, including those with short reading times and all-Like behaviors. The descriptive and correlation analyses (Table S2) revealed no substantial changes in the relationships between variables.

6.3.2 Logistic regressions. We preregistered a logistic regression analysis where Like-clicking on the extreme post served as the dependent variable. We used MMOD, the attitude toward the extreme post, the interaction effect between MMOD and attitude toward the extreme post, attitude toward cricket-eating, age, gender, education levels, and InfoPlatforms usage rate as the independent variables. Table 6 presents the results.

The findings revealed that MMOD and attitude toward the extreme post significantly influenced the likelihood of clicking "Like" on the extreme post, with odds ratios of 2.15 ($p = 0.02$) and 10.54 ($p < 0.01$), respectively. The notably high odds ratio of attitude toward the extreme post confirmed that it was the predominant predictor of Like-clicking. This result supports H2a, which posits that participants' attitudes toward an extreme post determined their Like-clicking. However, we observed a significant negative interaction effect between

Table 6. Results of the logistic regression

	Estimate	z-value	*p*-value	Odds ratio
(Intercept)	−3.23 (0.54)	−5.97	<0.01	
MMOD	0.77 (0.32)	2.36	0.02	2.15
Attitude toward the extreme post	2.36 (0.37)	6.40	<0.01	10.54
Attitude toward cricket-eating	−0.27 (0.20)	−1.31	0.19	0.77
InfoPlatforms usage rate	−0.00 (0.20)	0.00	1.00	1.00
Age	0.04 (0.19)	0.19	0.85	1.04
Gender (=1)	−0.48 (0.40)	−1.22	0.22	0.62
Education (=1)	−0.05 (0.39)	−0.13	0.89	0.95
MMOD × Attitude toward the extreme post	−0.68 (0.25)	−2.75	0.01	0.51

Note(s): Numbers in parentheses indicate standard error. MMOD: Multi-dimensional measure of online disinhibition, Gender (0 = Male, 1 = Female), Education level (0 = Below bachelor's degree, 1 = Bachelor's degree or above)
Source(s): Authors' own creation/work

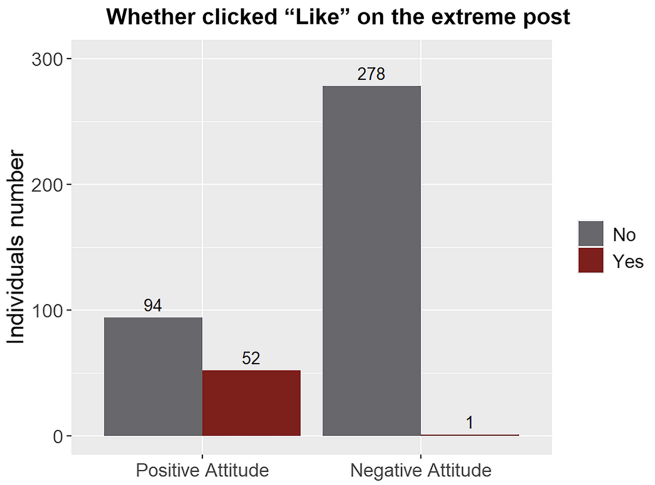
MMOD and attitude toward the extreme post, indicating that the influence of online disinhibition decreased as participants' Like-clicking motivation for the extreme post increased. This finding contradicted the MOD model, which we attributed to the influence of outliers in the statistical model, specifically participants who held negative attitudes toward the extreme post but still clicked "Like".

We performed additional preregistered analyses to address this problem. First, based on the score of attitude toward the extreme post (ranging from 1 to 4), participants were categorized into positive-attitude ($N = 146$; scores >2.50) and negative-attitude ($N = 279$; scores ≤2.50) groups. Figure 3 illustrates the Like-clicking frequency of both groups. This result indicated that only one individual in the negative-attitude group clicked "Like" on the extreme post. This finding supported our prior assumption about the presence of outliers and indirectly validated the MOD model's perspective: when participants held negative attitudes toward the extreme post (i.e. were less motivated to click "Like"), they were unlikely to click "Like."

We performed an additional logistic regression analysis on the positive-attitude group [1]. We used whether participants clicked "Like" on the extreme post as the dependent variable, MMOD, attitude toward the extreme post, attitude toward cricket-eating, age, gender, education level, and InfoPlatforms usage rate as independent variables. Table S3 presents the results. The analysis revealed that MMOD had an insignificant effect ($p = 0.42$), with a coefficient of 0.16 and an odds ratio of 1.18. This result does not support H2b, which posits that online disinhibition would exert a stronger effect on clicking "Like" on an extreme post when participants hold a positive attitude. This indicated that online disinhibition might not have a substantial impact on the expression of deviant attitudes, even with a certain motivation level.

6.3.3 Exploratory analyses. The correlation analysis revealed a significant relationship between MMOD and clicking "Like" on the extreme post, potentially indicating that online disinhibition directly affects Like-clicking behavior, as proposed in the disinhibition–behavior model. Therefore, we further explored the numerical relationships between MMOD and clicking "Like" on the extreme post.

Initially, we performed a logistic regression analysis using whether participants clicked "Like" on the extreme post as the dependent variable and MMOD as the independent variable. The results revealed a significant influence of MMOD on clicking "Like" on the extreme post,



Source(s): Authors' own creation/work

Figure 3. Stacked bar chart of the "Likes" on the extreme post

with a standardized path coefficient of 0.39 ($p < 0.01$). Subsequently, we performed a mediation analysis using whether participants clicked "Like" on the extreme post as the dependent variable, MMOD as the independent variable, and attitude toward the extreme post as the mediator. This analysis revealed a significant mediation model (average causal mediated effect = 0.02, $p < 0.01$), explaining 71.5% of the variance through the mediating variable ($p = 0.05$). By introducing the mediator, the direct effect of MMOD on clicking "Like" on the extreme post was reduced to 0.01 ($p = 0.55$). This finding indicates that the connection between MMOD and clicking "Like" on the extreme post might be mediated by participants' positive attitude toward it, which further suggests that MMOD did not exhibit a direct causal relationship with Like-clicking the extreme post.

6.4 Discussion

In Study 2, the participants' motivations were measured before they made Like-clicking decisions, which allowed for a more rigorous causal examination of the MOD model. However, this procedure interrupted the continuity of the thread-reading, thereby sacrificing the ecological validity of the experiment. It appears to be a challenging task to balance rigorous causality with a natural reading experience. Since the moderating effect of online disinhibition was successfully verified in Studies 1a and 1b, we shifted our focus toward quantifying the participants' motivation and examining its causal impact on Like-clicking.

In Study 2, we determined the target to be a particularly extreme post featuring language that was extremely offensive and deviant. Clicking "Like" on such a post was considered an expression of a deviant attitude. A series of analyses convincingly demonstrated that the underlying motivation, rather than online disinhibition, primarily determined the expression of deviant attitudes via Like-clicking. We believe that these findings offer a rational explanation for the expression of deviant attitudes. On the Internet, anonymously clicking "Like" on extreme content that resonates with is a relatively safe engagement. However, it is challenging to imagine that an individual would absurdly engage in clicking "Like" solely because of the "disinhibition" effect without any approval of the content. Thus, these results challenged the position of the disinhibition–behavior model used in previous studies.

Interestingly, our exploration further revealed that the popularity of the disinhibition–behavior model was not without reason. An exploratory analysis initially reproduced the significant relationship between online disinhibition and clicking "Like" on the extreme post. However, the introduction of attitude toward the extreme post as a mediating variable completely nullified the direct effect of disinhibition, potentially indicating the absence of a direct causal link. We must clarify that the mediation model was only an exploratory analysis of the numerical relationship among variables. There may be a reasonable correlation between online disinhibition and a positive attitude toward the extreme post: the individuals who exhibited less inhibition online were prone to displaying higher tolerance for extreme expressions, while a causal relationship of online disinhibition toward a positive view of posts with extreme language required further investigation. This mediation analysis illustrated that conclusions based solely on correlational data that overlooked causality could be misleading and fragile.

Notably, we did not assert the non-correlation between online disinhibition and the expression of deviant attitudes. The experiment target in Study 2 was focused on a single post, which could be more susceptible to bias from sampling, the data exclusion process, extreme values, or accidental mistaken "Likes." Therefore, a more comprehensive examination is required to further explore this relationship.

7. General discussion

7.1 Implications for theory

The present study's most significant theoretical implication is demonstrating that the online disinhibition effect is not a universal mechanism but is influenced by specific contextual and

motivational factors, which challenges the traditional disinhibition–behavior model. Originally, Suler's (2004) online disinhibition theory posited that the pervasive phenomenon of people behaving more openly online could be attributed to the "disappearance of inhibition." Under this framework, the disinhibition–behavior model simply posits that online disinhibition directly leads to disinhibitory behavior, which appears to be intuitive and is supported by empirical evidence. Indeed, if we considered only the initial parts of the hierarchical GLMMs from Study 1b or the exploratory mediation analysis from Study 2, we would also observe findings that support this model. However, a fatal limitation of this model is its failure to convincingly demonstrate a rigorous causal relationship in which online disinhibition determines online behavior. This could allow researchers to use fundamentally correlational evidence to make an overgeneralization explanation for various correlated behaviors (Burton *et al*., 2021). For example, it might be claimed that "experiencing online disinhibition might lead to more friendly behaviors, but it might also lead to more hostile behaviors (e.g. Stuart and Scott, 2021)." While such statements are thought-provoking, they also remain ambiguous and self-serving, potentially overgeneralizing the theory and reducing its explanatory power. To address these issues, we emphasize the importance of adopting a causal approach to explore the mechanisms of online disinhibition and advocate for a more precise examination through the perspective of the MOD model.

In this study, we conducted a series of experiments that generally supported the MOD model. Studies 1a and 1b indicated that disinhibition does not work when the object of expression of positive attitudes is inherently insignificant, commonplace, or safe. Study 2 revealed that in cases where participants disapprove of the stance of a target post, the proposal that online disinhibition influences their Like-clicking is fundamentally flawed. These findings underscore the significance of motivational factors in the mechanisms of online behavior, resonating with previous studies on face-to-face communication. Johnson and Downing (1979) found that mere anonymity—essentially a form of disinhibition from identity—could not directly predict aggressive behaviors; instead, situational cues were crucial. Anonymous individuals in nurse uniforms, where prosocial motivations were activated, exhibited less aggression. Conversely, those wearing KKK clothes, where antisocial motivations were activated, exhibited more aggression. Regarding these findings, we attribute them not only to a self-explanatory conclusion that "the online disinhibition effect varies from different situations" but also propose a more critical stance that online disinhibition does not provide a one-size-fits-all explanation for all online behavior. In the absence of motivation, or when behavior aligns with social norms and does not challenge cognitive inhibition, online disinhibition should not be regarded as a predictive factor. By establishing a more precise boundary for the application of online disinhibition, researchers can better understand online behaviors within a more systematic framework.

7.2 Implications for practice

As a practical contribution, this study highlights a crucial direction for addressing aggressive behavior on the Internet. While previous research has primarily emphasized that the disinhibitory mental state contributes to Internet users' aggressive behaviors (e.g. Kurek *et al.*, 2019), this study posits that the root of these behaviors lies in individuals' disinhibitory motivational factors. Therefore, social media administrators should pay special attention to content that may provoke emotional reactions and limit its exposure and dissemination to prevent triggering other users' inflammatory motivation. Clinical workers should focus on individuals' inherent motivations and use psychological interventions to alleviate aggressive impulses or desires in real-world settings. These approaches offer effective strategies for addressing aggressive behavior on the Internet at a fundamental level.

Another practical contribution is the innovative methodology proposed in this study for simulating online behavior. By integrating graphic elements with interactive "Like" buttons, this approach creates an immersive thread-reading experience that closely mimics real social

media interactions. Additionally, as it eliminates the need to build a server or set up front-end interactive functions, the experimental setup significantly reduces both costs and technical requirements. Researchers can utilize this method to capture participants' real-time, spontaneous reactions, enhancing ecological validity compared to traditional methods that assess participants' likelihood of engaging in certain behaviors.

7.3 Limitations

The present study has the following limitations.

First, the motivation behind online behavior warrants further investigation. The MOD model introduced individuals' motivational factors, complicating the antecedents of online behavior compared to the disinhibition-behavior model. Given that human motivation is extremely intricate, interpreting online behavior from the perspective of the MOD model may require incorporating an excessive number of factors into the statistical model. To avoid overcomplicating the statistical analysis, this study adopted a simplified approach, conceptualizing motivation from the perspectives of extrinsic stimuli or intrinsic motivations. Future research should focus on developing a comprehensive yet concise approach to assessing the motivations behind certain online behaviors. Such an improvement would enable a more detailed investigation of how motivation and online disinhibition influence online behavior.

Second, although the thread-reading task in this study improved ecological validity, it remained an artificial experimental approach with inherent limitations. For example, the experimental environment was controlled and could not replicate real-time changes in the number of "Likes," resulting in lower ecological validity compared to data obtained from real social media. Additionally, the aggressive context of the thread may have made participants more likely to be influenced by social desirability bias (Van de Mortel, 2008). Future studies employing observational approaches with higher ecological validity could provide more supplementary evidence for this field and strengthen the generalizability of the findings.

Notes

1. Logistic regression analysis on the negative-attitude group was also preregistered; however, because only one sample clicked "Like" in this group, this analysis was not conducted.

References

Bargh, J.A., McKenna, K.Y.A. and Fitzsimons, G.M. (2002), "Can you see the real me? Activation and expression of the 'true self' on the Internet", *Journal of Social Issues*, Vol. 58 No. 1, pp. 33-48, doi: 10.1111/1540-4560.00247.

Buckels, E.E., Trapnell, P.D. and Paulhus, D.L. (2014), "Trolls just want to have fun", *Personality and Individual Differences*, Vol. 67, pp. 97-102, doi: 10.1016/j.paid.2014.01.016.

Burton, J.W., Cruz, N. and Hahn, U. (2021), "Reconsidering evidence of moral contagion in online social networks", *Nature Human Behaviour*, Vol. 5 No. 12, pp. 1629-1635, doi: 10.1038/s41562-021-01133-5.

Castaño-Pulgarín, S.A., Suárez-Betancur, N., Vega, L.M.T. and López, H.M.H. (2021), "Internet, social media and online hate speech. Systematic review", *Aggression and Violent Behavior*, Vol. 58, 101608, doi: 10.1016/j.avb.2021.101608.

Cheung, C.M.K., Wong, R.Y.M. and Chan, T.K.H. (2021), "Online disinhibition: conceptualization, measurement, and implications for online deviant behavior", *Industrial Management and Data Systems*, Vol. 121 No. 1, pp. 48-64, doi: 10.1108/IMDS-08-2020-0509.

Chou, C., Condron, L. and Belland, J.C. (2005), "A review of the research on Internet addiction", *Educational Psychology Review*, Vol. 17 No. 4, pp. 363-388, doi: 10.1007/s10648-005-8138-1.

Eranti, V. and Lonkila, M. (2015), "The social significance of the Facebook like button", *First Monday*, Vol. 20 No. 6, doi: 10.5210/fm.v20i6.5505.

Fine, J.A. and Hunt, M.F. (2023), "Negativity and elite message diffusion on social media", *Political Behavior*, Vol. 45 No. 3, pp. 955-973, doi: 10.1007/s11109-021-09740-8.

Hayes, R.A., Carr, C.T. and Wohn, D.Y. (2016), "One click, many meanings: interpreting paralinguistic digital affordances in social media", *Journal of Broadcasting and Electronic Media*, Vol. 60 No. 1, pp. 171-187, doi: 10.1080/08838151.2015.1127248.

Hernández, M.P., Schoeps, K., Maganto, C. and Montoya-Castilla, I. (2021), "The risk of sexual-erotic online behavior in adolescents: which personality factors predict sexting and grooming victimization?", *Computers in Human Behavior*, Vol. 114, 106569, doi: 10.1016/j.chb.2020.106569.

Hollenbaugh, E.E. and Everett, M.K. (2013), "The effects of anonymity on self-disclosure in blogs: an application of the online disinhibition effect", *Journal of Computer-Mediated Communication*, Vol. 114 No. 3, pp. 283-302, doi: 10.1111/jcc4.12008.

Johnson, R.D. and Downing, L.L. (1979), "Deindividuation and valence of cues: effects on prosocial and antisocial behavior", *Journal of Personality and Social Psychology*, Vol. 37 No. 9, pp. 1532-1538, doi: 10.1037/0022-3514.37.9.1532.

Joinson, A.N. (2007), "Disinhibition and the Internet", in *Psychology and the Internet*, Academic Press, pp. 75-92, doi: 10.1016/B978-012369425-6/50023-0.

Kordyaka, B., Jahn, K. and Niehaves, B. (2020), "Towards a unified theory of toxic behavior in video games", *Internet Research*, Vol. 30 No. 4, pp. 1081-1102, doi: 10.1108/INTR-08-2019-0343.

Kurek, A., Jose, P.E. and Stuart, J. (2019), "I did it for the LULZ': how the dark personality predicts online disinhibition and aggressive online behavior in adolescence", *Computers in Human Behavior*, Vol. 98, pp. 31-40, doi: 10.1016/j.chb.2019.03.027.

Lapidot-Lefler, N. and Barak, A. (2012), "Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition", *Computers in Human Behavior*, Vol. 28 No. 2, pp. 434-443, doi: 10.1016/j.chb.2011.10.014.

Lapidot-Lefler, N. and Barak, A. (2015), "The benign online disinhibition effect: could situational factors induce self-disclosure and prosocial behaviors?", *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, Vol. 9 No. 2, doi: 10.5817/CP2015-2-3.

Maniaci, M.R. and Rogge, R.D. (2014), "Caring about carelessness: participant inattention and its effects on research", *Journal of Research in Personality*, Vol. 48, pp. 61-83, doi: 10.1016/j.jrp.2013.09.008.

McCoach, D.B. (2010), "Hierarchical linear modeling", in *The Reviewer's Guide to Quantitative Methods in the Social Sciences*, Routledge, pp. 123-140, doi: 10.4324/9780203861554.

Moscatelli, A., Mezzetti, M. and Lacquaniti, F. (2012), "Modeling psychophysical data at the population-level: the generalized linear mixed model", *Journal of Vision*, Vol. 12 No. 26, pp. 1-17, doi: 10.1167/12.11.26.

Schouten, A.P., Valkenburg, P.M. and Peter, J. (2007), "Precursors and underlying processes of adolescents' online self-disclosure: developing and testing an 'Internet-attribute-perception' model", *Media Psychology*, Vol. 10 No. 2, pp. 292-315, doi: 10.1080/15213260701375686.

Seigner, B.D.C., Milanov, H., Lundmark, E. and Shepherd, D.A. (2023), "Tweeting like Elon? Provocative language, new-venture status, and audience engagement on social media", *Journal of Business Venturing*, Vol. 38 No. 2, 106282, doi: 10.1016/j.jbusvent.2022.106282.

Stuart, J. and Scott, R. (2021), "The measure of online disinhibition (MOD): assessing perceptions of reductions in restraint in the online environment", *Computers in Human Behavior*, Vol. 114, 106534, doi: 10.1016/j.chb.2020.106534.

Suler, J. (2004), "The online disinhibition effect", *CyberPsychology and Behavior*, Vol. 7 No. 3, pp. 321-326, doi: 10.1089/1094931041291295.

The Nikkei (2022), "Introduction of edible cricket powder in school lunches - a Nationwide First in Tokushima", 28 November, available at: https://www.nikkei.com/article/DGXZQOCC24BFE0U2A121C2000000/ (accessed 25 February 2025).

Udris, R. (2014), "Cyberbullying among high school students in Japan: development and validation of the Online Disinhibition Scale", *Computers in Human Behavior*, Vol. 41, pp. 253-261, doi: 10.1016/j.chb.2014.09.036.

Van de Mortel, T.F. (2008), "Faking it: social desirability response bias in self-report research", *Australian Journal of Advanced Nursing*, Vol. 25 No. 4, pp. 40-48, available at: https://search.informit.org/doi/pdf/10.3316/informit.210155003844269?download=true

Walther, J.B. (1996), "Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction", *Communication Research*, Vol. 23 No. 1, pp. 3-43, doi: 10.1177/009365096023001001.

Wen, R. and Miura, A. (2023), "Online disinhibition: reconsideration of the construct and proposal of a new model", *Osaka Human Sciences*, Vol. 9, pp. 63-80, doi: 10.18910/90710.

Wen, R. and Miura, A. (2024), "Development of the multi-dimensional measure of online disinhibition and examination of its validity and reliability", *Osaka Human Sciences*, Vol. 10, pp. 19-38, doi: 10.18910/94827.

Wright, M.F., Harper, B.D. and Wachs, S. (2019), "The associations between cyberbullying and callous-unemotional traits among adolescents: the moderating effect of online disinhibition", *Personality and Individual Differences*, Vol. 140, pp. 41-45, doi: 10.1016/j.paid.2018.04.001.

Zimmerman, F., Bailey, D.D., Muric, G., Ferrara, E., Schöne, J., Willer, R., Halperin, E., Navajas, J., Gross, J.J. and Goldenberg, A. (2024), "Attraction to politically extreme users on social media", *PNAS Nexus*, Vol. 3 No. 10, pgae395, doi: 10.1093/pnasnexus/pgae395.

Supplementary material
The supplementary material for this article can be found online.

Corresponding author
Ruohan Wen can be contacted at: runningwz@gmail.com