



Social bot literacy and responses to bot-generated information in AI-mediated environments: the roles of self-efficacy and perceived threat

Wei Fang and Chen Nie

DOI: <https://doi.org/10.47989/ir31263122>

Abstract

Introduction. Social bots constitute an increasingly prominent form of artificial intelligence within online information environments, shaping how information is generated, circulated, and assessed. This study investigates how social bot literacy relates to responses to bot-generated information, with particular attention to the mediating role of social bot self-efficacy and the moderating role of perceived threats from bots.

Method. A mixed-methods design was employed. Survey data were collected from 1,159 Sina Weibo users in China to examine associations among social bot literacy, social bot self-efficacy, perceived threats, and information response strategies. In addition, in-depth interviews were conducted to explore how bot-generated information is interpreted and evaluated in everyday information environments.

Analysis. Quantitative data were analysed using confirmatory factor analysis and moderated mediation models. Qualitative interview data were thematically analysed to contextualise information evaluation and response processes in AI-mediated settings.

Results. The results indicate that higher social bot literacy is associated with greater confidence in handling bot-generated information and with more active information response strategies. Social bot self-efficacy mediates this relationship, while perceived threat moderates the strength of the indirect effects. Interview findings further show how social bots are understood as information actors and how informational risks are negotiated.

Conclusion. The study advances Information Science by clarifying how social bot literacy shapes information evaluation and use in AI-mediated environments, with implications for information literacy development and responsible AI governance.

Introduction

Social bots, software-driven accounts designed to imitate human behaviour, are becoming increasingly prevalent on social media platforms (Ferrara et al., 2016; Ngo et al., 2023). Initially created to facilitate communication in areas such as business, daily life, and public crises (Bossu et al., 2023), these bots have also been utilized for malicious purposes, including the spread of disinformation and manipulation of public opinion during political events (Hagen et al., 2022; Martini et al., 2021; Sharma et al., 2022), economic affairs (Fan et al., 2020; Hajli et al., 2022), controversial scientific issues (Al-Rawi et al., 2021; Daume et al., 2023; Marlow et al., 2021), and health-related concerns (Shi et al., 2020; Suarez-Lledo and Alvarez-Galvez, 2022). These malicious uses have polluted individuals' information environments and pose significant threats to the integrity of democratic processes.

To address these negative impacts of social bots, various solutions have been proposed, particularly through technological and policy related interventions. Researchers have focused on analysing bot behaviour and profiles (Confessore et al., 2018; Keller et al., 2020) while developing advanced detection tools to combat bots' harmful influence (Ferrara, 2023; Nightingale and Farid, 2022). Beyond technological solutions, there have been calls to increase the responsibility of social media platforms in protecting users and reducing harm (Cresci et al., 2023; Hays et al., 2023; Tan et al., 2023). Moreover, regulatory measures addressing the mass production, dissemination, and consequences of social bots have gained attention as a key approach to mitigating the negative effects of bots (Gorwa and Guilbeault, 2020).

While technological and structural interventions have been effective in reducing the harm caused by bots, it is crucial to recognize that the influence of social bot depends on human users who interact with these bots. Users process, engage with, and respond to bot-generated content, and their actions can mediate the bots' ability to manipulate public opinion. Evidence has shown that human-bot interactions often involve sentiment exchange, message amplification, and user engagement (Duan et al., 2022; Kenny et al., 2024; Kušen and Strembeck 2019; Wald et al., 2013; Yan et al., 2021). Despite this, research on human-bot interactions remain limited (Yan and Yang, 2022). Furthermore, previous studies on social bots from a human-centred perspective have explored users' growing awareness, skills, and self-efficacy in bot detection (Fang and Nie, 2024; Ngo et al., 2023; Schmuck and von Sikorski, 2020; Yan et al., 2021). While these studies provide valuable insights into aspects of social bot literacy, they do not explicitly examine the relationship between users' literacy, detection abilities, and their responses to bots. Some findings suggest that users engage in certain responses, such as avoiding interaction, reporting bots, or using verification tools (Ngo et al., 2023; Wischnewski et al., 2024; Yan et al., 2023), but the underlying mechanisms driving these responses remain unclear. Given this, empowering users by enhancing their ability to critically evaluate and respond to online information appears to be a promising approach to mitigating the influence of social bots (Fang and Nie 2024; Lazer et al., 2018; Schmuck and von Sikorski 2020).

Building on these insights, this study examines the associations between individuals' social bot literacy and two forms of responses—protective strategies to safeguard one's online environment and pressuring strategies to prompt platform action—within a resistance practice framework (DeVrio et al., 2024). We further explore the potential mediating role of social bot self-efficacy and the moderating role of perceived threats from bots. Empirically, we combine survey data from 1,159 Sina Weibo users with in-depth interviews with 20 users. This mixed-method design enables us to test patterned associations statistically while using qualitative evidence to clarify how users interpret and enact these responses in practice. Together, the findings contribute to research on human-bot interaction, user empowerment, and digital resilience, and provide evidence to inform user-centred interventions alongside technological and regulatory approaches.

Theoretical framework

Users' responses to social bots

As awareness of social bots and their potentially harmful impact on the information environment grows (Schmuck and von Sikorski, 2020; Stocking and Sumida, 2018), many users have accumulated knowledge about detecting bots and utilizing verification tools and services (Ngo et al., 2023). However, after identifying social bots, how do users respond to mitigate the harms posed by these bots on the trustworthiness and credibility of information?

Research specifically addressing users' responses to identified social bots is still limited. Given that social bots are driven by algorithms, insights can be drawn from studies on users' responses to algorithmic harms. Algorithmic harm refers to "the adverse lived experiences resulting from [an algorithmic] system's deployment and operation" (Shelby et al., 2022). In response to such harms, users have employed various strategies to mitigate their effects. A recent study from DeVrio et al (2024) constructed a taxonomy of responses "from below" to algorithmic harm, revealing how individuals without structural power act to "employ, shift, or build power in response to algorithmic harms" (p.1095). The study identifies five broad categories of responses: (1) mitigating individual algorithmic harms, (2) pursuing legal avenues, (3) investigating harmful algorithmic behaviours, (4) refusing engagement with harmful systems, and (5) communicating algorithmic harm with others. Additionally, three overarching strategies demonstrate how users navigate power dynamics in their responses: protecting strategies, which involve users leveraging their power to counter algorithmic harms; pressuring strategies, where users exert influence on those controlling algorithms to shift power in their favour; and strengthening strategies, which enhance users' ability to protect and pressure (DeVrio et al., 2024). This classification offers valuable insights into how users might respond to social bots, which are similarly driven by algorithms, in an effort to mitigate the harm they pose to both personal and societal information environments.

Several empirical studies provide further evidence of users' reactions to social bots. When users suspect that certain accounts are bots, they may turn to automated detection tools to verify these suspicions (Ngo et al., 2023). Additionally, some users choose to avoid engaging with accounts they perceive as bots. A study examining Twitter users' interactions with social bots found that participants were less inclined to engage in activities such as following, retweeting, commenting, or quote-tweeting bot accounts. Moreover, they were more likely to ignore, block, or report these accounts (Wischniewski et al., 2024).

Beyond individual actions, users also seek structural measures to limit the impact of bots. In another study exploring users' perceptions of bots and preferred countermeasures, policies and regulations targeting social media companies and laws penalising the mass production of bots were the most endorsed responses (Yan et al., 2023). Additionally, users also supported interventions that enhance individual literacy, including self-education and public media literacy education programs, as well as efforts from the social media industry to mitigate bot-related harms (Yan et al., 2023).

The studies reviewed above suggest that users respond to social bots either through individual actions (e.g., bot verification) or by advocating for structural changes (e.g., policy interventions) (e.g., Ngo et al., 2023; Wischniewski et al., 2024; Yan et al., 2023). These responses generally fall into two categories: protective strategies, aimed at safeguarding control over one's information environment, and pressuring strategies, intended to shift responsibility toward platforms and bot developers. Drawing on the classification of responses to algorithmic harms (DeVrio et al., 2024) and prior empirical work, this study examines how users employ these two strategies when dealing with social bots.

To clarify the empirical basis of this subsection, we summarize the main methodological tendencies. Current evidence on users' responses to social bots mainly comes from survey-based and experimental studies, with outcomes typically measured as willingness to verify, disengage, report, or support regulation (Ngo et al., 2023; Wischniewski et al., 2024; Yan et al., 2023). Common instruments include Likert-type intention items, scenario-based exposure tasks, and self-reported behavioural preferences. Less common are behavioural-trace measures (e.g., observed reporting/blocking logs) and longitudinal designs that capture response change over time. Overall, the literature has identified key response patterns, but measurement remains concentrated on intention-focused indicators, with fewer studies combining observed behaviour and process-oriented evidence across contexts.

Social bot literacy and users' responses

Social bot literacy refers to an individual's knowledge, experience and skills related to social bots and their impact on individuals and society. Schmuck and von Sikorski (2020, p.3) define social bot literacy as "the awareness, attitude and ability of individuals to appropriately identify, evaluate, analyse and interact with social bots as well as to be able to reflect upon this process" (p.3). This framework encompasses the cognitive, attitudinal, and behavioural dimensions of literacy. It builds upon the broader concepts of digital literacy (e.g., Martin, 2006) and internet literacy (e.g., Livingstone and Helsper, 2010), while adapting these ideas to address the specific challenges posed by social bots.

Although there is limited research directly examining the relationship between users' social bot literacy and their responses to bots, several studies suggest an implicit connection between the two. Ngo et al (2023) explored the conditions under which users would identify accounts as social bots. In their study, aspects of social bot literacy—such as social bot experience, knowledge of bot identification tools, and experience using these tools—were included as control variables. Although that study did not directly examine literacy–response pathways, its inclusion of bot literacy as a background covariate indicates that literacy was considered relevant to explaining user responses. Similarly, another study examining users' engagement with or reactions to social bots also treated bot knowledge as a control variable, further hinting the influence of bot literacy on users' responses (Wischniewski et al., 2024).

This relationship is further supported by an experimental study investigating the cognitive effect of exposure to bots on users' responses (Yan et al., 2023). The findings indicate that users' experience with social bots, or exposure to them, tends to trigger stronger support for stricter regulations on bots. Moreover, as seen in previous discussions, users' responses to social bots include both passive (e.g., avoidance) and active (e.g., detection, regulation) strategies. Therefore, it appears that users' social bot literacy—encompassing their knowledge of bots, experience with bot detection tools, and skills in interacting with bots—correlates with both passive and active responses to mitigate the harms caused by social bots.

Given variation in operationalization across studies, we summarize common approaches to measuring social bot literacy. Research on social bot literacy primarily relies on self-report scales capturing awareness, knowledge, and experience with bots, often adapted from broader digital and media literacy traditions (Schmuck and von Sikorski, 2020). In several studies, literacy-related indicators are included as controls or background factors rather than modelled as focal explanatory constructs (Ngo et al., 2023; Wischniewski et al., 2024). Experimental work also uses exposure-based designs to examine bot-related perceptions and policy preferences (Yan et al., 2023), but direct measurement of literacy-to-response pathways is less common. In sum, commonly used instruments emphasize perceived literacy, while less common approaches include multi-method literacy assessment and integrated modelling of literacy with differentiated response strategies.

Based on the above insights, we propose the following hypotheses:

H1: There is a positive correlation between users' level of social bot literacy and their protective strategies for responding to social bots

H2: There is a positive correlation between users' level of social bot literacy and their pressuring strategies for responding to social bots

The mediating effect of social bot self-efficacy

Social bot self-efficacy, rooted in the broader framework of self-efficacy (e.g., Ajzen 2002; Bandura, 1982), refers to individuals' confidence in their ability to distinguish social bots from human accounts (Yan et al., 2021). Yan et al (2021) define social bot self-efficacy as the belief that one can successfully recognize most social bots when encountering them, effectively differentiate bots from humans, and identify clues that reveal bots resembling regular users. Similarly, (Schmuck and von Sikorski 2020, p.3) define perceived social bot self-efficacy as 'one's perceived confidence in their ability to detect a bot as a bot', measured by perceived ability, competence, and confidence in distinguishing social bots from human users on social media platforms. These definitions reflect users' confidence in their ability to identify social bots accurately.

Empirical studies consistently suggest a positive correlation between social bot literacy and social bot self-efficacy. Many users have developed social bot literacy— comprising awareness of bots, knowledge of bot detection devices, and expertise in identifying bots—through news coverage and personal experiences with bots (Fang and Nie, 2024). This growing literacy is positively associated with users' self-efficacy or confidence, in detecting bots (Fang and Nie, 2024; Ngo et al., 2023; Schmuck and von Sikorski, 2020; Stocking and Sumida, 2018; Yan et al., 2021). For example, a Pew survey revealed that young people, who reported the highest level of social bot literacy in terms of bot awareness, also displayed higher levels of confidence in identifying bots (Stocking and Sumida, 2018). Similarly, an experimental study investigating the relationship between social bot literacy, perceived bot control and threats, found that increased knowledge and skills regarding bots enhanced users' perceived ability to distinguishing bots from humans (Schmuck and von Sikorski, 2020).

Despite the abundance of research linking social bot literacy to bot self-efficacy, there is limited research examining the relationship between social bot self-efficacy and users' responses to bots. Most studies have focused on the link between bot self-efficacy and users' performance in detecting bots (e.g., Yan et al., 2021, 2023). These studies suggest that individuals with higher level self-efficacy tend to perform better in bot detection tasks before being exposed to bots, though this correlation diminishes after exposure (Yan et al., 2023). This pattern aligns with research highlighting the motivating role of self-efficacy in task performance (Bandura, 1986; Huang, 2013). Furthermore, as discussed earlier, once users detect social bots, they often respond by avoiding or disengaging with bot accounts (Wischnewski et al., 2024) and supporting stricter regulatory measures (Yan et al., 2023). Therefore, it is reasonable to infer that an increase in social bot self-efficacy may positively link to users' responses to bots.

Existing findings support the relevance of self-efficacy, but studies differ in whether they treat efficacy as a perceptual construct only or link it to behavioural outcomes. Research on social bot self-efficacy relies primarily on self-report Likert-type scales of perceived detection ability, competence, and confidence, with some studies pairing these measures with performance-based detection tasks in experimental settings. Compared with literacy measurement, self-efficacy instruments are generally more standardized around confidence judgments, but fewer studies track how efficacy beliefs relate to multiple response types within one integrated model. Mixed

designs that connect efficacy, behaviour, and contextual interpretation remain comparatively less common.

Based on these discussions, we propose the following hypotheses:

H3: There is a positive correlation between users' level of social bot literacy and their perceived level of social bot self-efficacy

H4: There is a positive correlation between users' level of social bot self-efficacy and protective strategies for responding to bots

H5: There is a positive correlation between users' level of social bot self-efficacy and pressuring strategies for responding to bots

The moderating effect of perceived threats from social bots

Perceived threats from bots refer to the extent to which individuals view social bots as potential threats to themselves and society (Schmuck and von Sikorski, 2020). Public awareness of the role of social bots in spreading misinformation has grown, driven by news coverage, personal encounter with bots on social media (Fang and Nie, 2024), and other sources. For instance, among the U.S. public, approximately two-thirds (66%) have heard of social bots, with only 34% reporting they know nothing about them (Stocking and Sumida, 2018). As social bots increasingly manipulate public opinion, concerns about their impact on information processing, opinion formation, and democratic processes are intensifying (Schmuck and von Sikorski, 2020; Starbird, 2019; Stocking and Sumida, 2018).

Some studies have treated perceived threats from bots as an outcome of social bot self-efficacy (e.g., Schmuck & von Sikorski, 2020), assuming that greater ability to identify and manage bots leads to clearer perceptions of their risks. However, recent research points to a more complex, potentially reciprocal link, where threat perceptions may also condition whether efficacy translates into action. For instance, Yan et al., (2023) found that before exposure to bots, social bot self-efficacy positively correlated with performance in identifying bots, but this correlation disappeared after exposure—indicating that changes in perceived threat levels after exposure could suppress the behavioural effects of self-efficacy. Similarly, Fang and Nie (2024) showed that perceived threats moderated the relationship between social bot literacy and perceived control. These findings suggest that threats act as a psychological filter: they may spur active resistance or prompt avoidance when perceived as overwhelming. Accordingly, this study treats perceived threats not as an outcome, but as a boundary condition shaping the efficacy–behaviour relationship.

To situate our moderation focus, we summarize how perceived threat has been measured in prior work. Perceived threat from social bots is typically measured through self-report scales of personal and societal risk, including concerns about information quality, opinion manipulation, and broader social consequences (Schmuck and von Sikorski, 2020). Experimental studies additionally manipulate exposure conditions to test shifts in risk appraisal and regulatory propensity (Yan et al., 2023). Most studies treat threat as a correlational outcome or accompanying perception; fewer test it explicitly as a moderator in pathways linking literacy or efficacy to user responses (Fang and Nie, 2024). Taken together, common approaches capture threat salience effectively, while less common approaches model threat as a boundary condition across multiple behavioural response pathways.

Based on this discussion, we propose the following research questions:

RQ1: Do perceived threats from bots moderate the association between social bot self-efficacy and users' protective strategies for responding to bots?

RQ2: Do perceived threats from bots moderate the association between social bot self-efficacy and users' pressuring strategies for responding to bots?

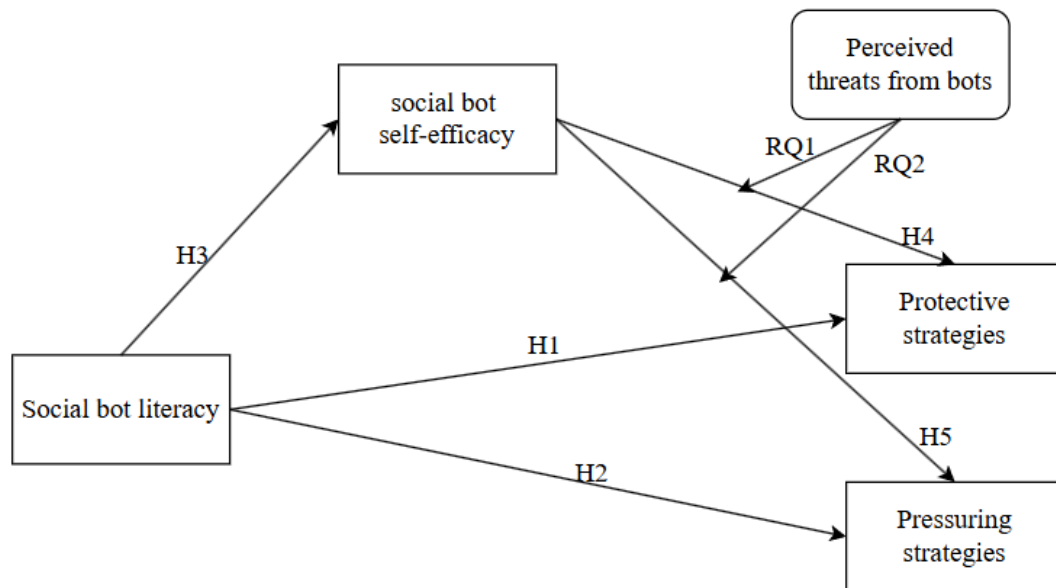


Figure 1. Theoretical model

Figure 1 presents the study's conceptual model. Social bot literacy is specified as the focal predictor and users' responses to bots are modelled as two distinct outcomes: protective strategies and pressuring strategies. Social bot self-efficacy is modelled as a mediator linking social bot literacy with both response strategies. Perceived threats from bots are modelled as a moderator of the self-efficacy-to-response paths, indicating that the strength of these associations may vary across levels of perceived threat.

Method

Participants and procedure

In September 2022, to test the research model and hypotheses, an anonymous online survey was conducted using a random sampling method on Wen Juan Xing, the largest online survey platform in China. The platform boasts a verified respondent database of 2.6 million users, ensuring a high level of authenticity and demographic diversity. It is widely recognized for its reliability and is frequently employed by both academic researchers and organizations. This study targeted users of Sina Weibo (hereafter referred to as Weibo), one of China's major social media platforms. Before starting the survey, all participants were informed about the study and provided consent on the introductory page of the questionnaire. A random sample was selected from the Wen Juan Xing database, resulting in 1,078 respondents. To extend the reach and increase the sample size, the survey link was also shared via WeChat, obtaining an additional 81 responses. After removing invalid responses, the final sample included 1,159 valid participants.

Table 1 displays the demographic characteristics of the participants. Approximately one-third (33.74%) of the participants were male, while the remaining two-thirds (66.26%) were female. The majority of participants were aged between 20 and 29 (57.03%) or between 30 and 39 (31.41%), together comprising over 80% of the sample. Participants in the 10–19 (7.33%) and 40 and above (4.23%) age groups made up the remaining 12%. In terms of educational background, the majority

of participants held a bachelor's degree (79.55%), followed by those with a vocational college degree (8.97%) and those with a high school diploma or lower (2.59%). The demographic characteristics of the participants align with those of the general Weibo user base, which is predominantly young, well-educated, and female (Thomala, 2023; Weibo, 2021). According to the official annual statistics released by Weibo, 55% of users are female and 45% are male. In terms of age, the 20–29 age group accounts for 48% of users, while the 30–39 age group accounts for 18%. Regarding educational attainment, approximately 68% of Weibo users have obtained a bachelor's degree or higher (Thomala, 2023). The slightly higher proportion of females (66.26% vs. 55%), 30–39-year-olds (31.41% vs. 18%), and participants with a bachelor's degree or above (88.54% vs. 68%) in our sample is understandable given our recruitment channels (Wen Juan Xing and WeChat), which tend to reach more highly educated individuals, and the well-documented tendency of women and slightly older users to be more active and responsive in online surveys (Sax et al., 2003; Kelfve et al., 2020). Overall, despite these modest deviations, the general trends in gender, age, and education in our sample are largely consistent with those of the platform's user base.

Demographic Variable	Item	Frequency	Percentage (%)	Cumulative percentage (%)
Gender	Male	391	33.74	33.74
	Female	768	66.26	100.00
Age	10–19	85	7.33	7.33
	20–29	661	57.03	64.37
	30–39	364	31.41	95.77
	≥ 40	49	4.23	100.00
Education	High school and below	30	2.59	2.59
	Vocational college	104	8.97	11.56
	Undergraduate	922	79.55	91.11
	Postgraduate	103	8.89	100.00
Total		1159	100.0	100.0

Table 1. Demographic profiles

Measures

All focal constructs were adapted from previously validated scales and contextualized to the Weibo setting. Because the measurement model was theory-driven and based on established instruments, validation focused on internal consistency and confirmatory factor analysis (CFA), rather than exploratory scale development.

We conducted confirmatory factor analysis (CFA) to test the factorial validity of latent constructs: social bot literacy, social bot self-efficacy, perceived threats from bots, protective strategies, and pressuring strategies. The model fit was acceptable to good: $\chi^2(44) = 325.796$, $p < .001$, GFI = 0.956, CFI = 0.975, RMSEA = 0.074 (90% CI [0.067, 0.082]), NFI = 0.971, TLI-NNFI = 0.962 and SRMR = 0.031. Although $\chi^2/df = 7.404$ exceeded the ideal threshold, this is common in large samples, and the overall fit remained satisfactory. All standardized loadings exceeded 0.70 ($p < .001$), confirming strong associations between indicators and latent factors. Convergent validity

was supported by AVE values ranging from 0.571 to 0.814 and CR values from 0.727 to 0.917, all above recommended cutoffs.

Discriminant validity mostly met the Fornell–Larcker criterion, except for social bot literacy (Square root of AVE = 0.789) and protective strategies (Square root of AVE = 0.756), whose square roots were slightly below their highest inter-construct correlations (0.800 and 0.785), indicating minor but acceptable concerns given theoretical distinctions.

Multicollinearity diagnostics showed all VIFs were well below 10, with the highest observed for social bot self-efficacy (VIF = 5.140), indicating moderate collinearity. However, all tolerance values ranged from 0.195 to 0.986, exceeding the 0.1 cutoff and confirming no serious multicollinearity issues. Therefore, all variables were retained for further analysis.

It is important to note that these CFA and collinearity diagnostics were intended to validate and consolidate observed items into latent variables for subsequent moderation and mediation analyses, rather than to build a fully specified structural model. Accordingly, a two-step analytic strategy was adopted: although the proposed theoretical framework could also be estimated using a full latent structural equation model, we first employed CFA to rigorously validate and refine the measurement model, ensuring adequate factorial validity and empirical distinctiveness among conceptually proximate constructs. Second, regression-based moderated mediation analyses were conducted to test the conditional indirect effects central to the research questions, which facilitates transparent estimation and interpretation of moderated mediation effects—particularly across levels of perceived threat—while avoiding additional model complexity that may obscure substantive interpretation. Together, the CFA and regression-based analyses provide complementary support for the robustness and interpretability of the proposed relationships.

Social bot literacy.

Drawing on previous studies (Schmuck and von Sikorski, 2020), Weibo bot literacy was assessed based on users' awareness of and interactions with social bots on the Weibo platform in China. Participants were asked to indicate their level of agreement with the following statements: "I am familiar with the characteristics of Weibo bots" and "I have personally identified Weibo bots among my social media followers" (1 = strongly disagree, 5 = strongly agree) (M = 3.310, SD = 1.082, Cronbach α = 0.757, AVE= 0.623, CR= 0.767, VIF=4.562, Tolerance=0.219).

Strategies for responding to social bots

Drawing on the framework developed by DeVrio et al (2024), participants' strategies for responding to social bots were classified into two types: protective strategies and pressuring strategies. To measure protective strategies, two items were averaged. Respondents indicated their agreement with the following statements using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree): 'I will ignore these social bot accounts' and 'I will block social bot accounts' (M = 3.7, SD = 1.24, Cronbach α = 0.72, AVE= 0.571, CR= 0.727, VIF=3.457, Tolerance=0.289). For pressuring strategies, another two items were averaged, with respondents expressing their agreement with these statements: 'I would prefer to report social bot accounts to the platform' and 'I will change my Weibo settings to prevent social bot accounts from interacting with me' (M = 3.45, SD = 1.28, Cronbach α = 0.80, AVE= 0.671, CR= 0.803, VIF=2.762, Tolerance=0.362).

Social bot self-efficacy.

Adapted from previous studies (Schmuck and von Sikorski, 2020; Yan et al., 2021,2023), perceived Weibo bot efficacy was measured by perceived self-efficacy. Participants were asked to rate their level of agreement with the following statements: 'I believe that I myself have the ability to detect social bots on Weibo platform' 'I think I have the ability to distinguish social bots on Weibo platform from real users' (1=strongly disagree, 5= strongly agree) (M = 3.43, SD=1.57, Cronbach α =0.896, AVE= 0.814, CR= 0.897, VIF=5.140, Tolerance=0.195).

Perceived threats from bots.

Adapted from previous studies (Fang and Nie, 2024; Schmuck and von Sikorski, 2020), perceived threats from bots were assessed based on Weibo users' perceptions of the risks posed by social bots at both the individual and societal levels. Respondents rated their agreement with the following statements using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree): 'Social bots on [the] Weibo platform pose a threat for me to have a comprehensive understanding of information' 'False information posted by social bots on [the] Weibo platform pose a threat for me to collecting information on a specific topic' 'Social bots on [the] Weibo platform are a threat to public opinion in the online space in China' 'Social bots on [the] Weibo platform are a threat to the offline life through affecting the online public opinion in China' (M 3.81, SD=1.18, Cronbach α = 0.916, AVE= 0.734, CR=0.917, VIF=3.490, Tolerance=0.287).

Control variables.

In line with prior research (Kenny *et al.*, 2024; Schmuck and von Sikorski, 2020), this study included three demographic variables as control factors: gender (66.26% female) (VIF=1.042, Tolerance=0.960), age (M =27.24, SD =6.59, VIF=1.051, Tolerance=0.951), and education level (M =3.955 ; SD =0.56, VIF= 1.014, Tolerance=0.986).

Statistical methods

In this study, we employed SPSS (version 26) and Hayes's PROCESS (version 4.0) to analyse the mediating and moderating effects of the examined factors. First, we utilized SPSS to provide descriptive statistics and compute the bivariate associations among all variables (Table 2). Second, hierarchical regressions were conducted to analyse H1, H2 and H3. Third, we used Model 4 to analyse the H4 and H5 (the mediation model). Additionally, we employed Model 14 to answer RQ1 and RQ2 (the moderated mediation model). The theoretical model is illustrated in Figure 1.

In addition to the survey, we conducted in-depth, semi-structured interviews with 20 Sina Weibo users to extend the study's explanatory scope and foreground users' own interpretations of bot-related experiences. Participants were recruited via Weibo and WeChat using snowball sampling, and discussions covered their experiences with bots, detection skills, perceived threats, and response strategies. Most interviews were conducted online, lasted 30–60 minutes, and were audio-recorded and transcribed. Using grounded-theory procedures, we coded the data iteratively and developed themes aligned with the study's core constructs (literacy, self-efficacy, perceived threat, and response strategies). These qualitative findings are presented as an integrated analytical component in Section "Qualitative Findings: Insights from In-Depth Interviews", where they are used to elaborate and contextualize the quantitative patterns rather than merely replicate them.

Results

Preliminary analyses

Descriptive statistics, including means, standard deviations, and correlations between potential variables, are shown in Table 2. Social bot literacy, social bot self-efficacy, and perceived threats from bots are positively associated with protective strategies ($r=0.704$, $P<0.001$; $r=0.688$, $P<0.001$; $r=0.785$, $P<0.001$), and are also positively associated with pressuring strategies ($r=0.649$, $P<0.001$; $r=0.620$, $P<0.001$; $r=0.717$, $P<0.001$). These significant correlations confirm the relationship between the variables in our hypotheses.

Variables	M	SD	1	2	3	4	5	6	7	8
1.Gender	–	–	–							
2.Age	27.23 9	6.588	-.170** *	–						
3.Education	3.948	0.557	-0.027	-0.013	–					
4.Social bot literacy	3.31	1.082	-0.016	0.007	0.060 *	–				
5.Social bot efficacy	3.427	1.192	-.074*	0.031	.093**	.625**	–			
6.Perceived threats	3.811	1.179	-.069*	-0.017	.079**	.731***	.714***	–		
7. Protective strategies	3.697	1.239	-0.048	0.042	.064*	.704** *	.688** *	.785** *	–	
8. Pressuring strategies	3.445	1.277	-.072*	0.052	0.012	.649** *	.620** *	.717***	.758** *	–

Note. *p < 0.05, **p < 0.01, ***p < 0.001.

Table 2. Means, standard deviations, and correlations of all variables.

Testing for model validation

Before testing our hypotheses and research questions, model validation was examined using criteria suggested by Hu and Bentler (1999). Estimations of the relationships among the full model showed excellent fit.

Testing for main direct effects

H1 proposed a positive correlation between users' level of social bot literacy and their protective strategies for responding to social bots. The regression analysis (see Table 3) revealed a significant positive relationship ($b=0.804$, $p<0.001$), indicating that higher levels of social bot literacy are associated with a greater likelihood of employing protective strategies. Specifically, respondents with higher social bot literacy were more inclined to engage in protective actions, such as avoiding or blocking bot accounts. Therefore, H1 is supported, demonstrating that social bot literacy plays a crucial role in driving protective responses to social bots.

H2 proposed a positive correlation between users' level of social bot literacy and their pressuring strategies for responding to social bots. The regression analysis (see Table 3) revealed a significant positive relationship ($b=0.767$, $p<0.001$), indicating that individuals with higher levels of social bot literacy are more likely to employ pressuring strategies, such as reporting bots or altering platform settings. Thus, H2 is supported, confirming that increased social bot literacy encourages users to take active measures to challenge social bots.

H3 hypothesized a positive correlation between users' level of social bot literacy and their perceived social bot self-efficacy. The results, as shown in Table 3, demonstrated a significant positive association ($b=0.789$, $p<0.001$), suggesting that higher social bot literacy is strongly linked to greater self-confidence in detecting and dealing with bots. Therefore, H3 is supported, further emphasizing the role of literacy in boosting users' perceived efficacy in managing social bots.

Predictors	Model 1 (Protective strategies)		Model 2 (Pressuring strategies)		Model 3 (Social bot self-efficacy)	
	β	t	β	t	β	t
Age	0.006	1.516	0.007	1.673	0.006	2.810**
Gender	-0.081	-1.456	-0.151	-2.482*	-0.066	-2.219*
Education	0.048	1.026	-0.064	-1.248	-0.019	-0.771
Social bot literacy	0.804	33.600***	0.767	29.136***	0.789	61.203***
R2	0.498		0.428		0.766	
F	286.550***		215.735***		944.332***	
*p < 0.05, **p < 0.01, ***p < 0.001						

Table 3. Hierarchical regression models of the main direct effects.

Testing for mediation effects

H4 proposed that social bot self-efficacy mediates the relationship between users' social bot literacy and their protective strategies. As shown in Table 4, we examined social bot literacy as the independent variable, protective strategies as the dependent variable, and social bot self-efficacy as the mediating variable within the mediation model. The mediation analysis was conducted in three steps, assessing the total and indirect effects.

In Model 1, social bot literacy positively influenced the use of protective strategies ($b=0.804$, $p<0.001$), indicating a significant total effect. Model 2 revealed a significant positive association between social bot literacy and social bot self-efficacy ($b = 0.684$, $p<0.001$). In Model 3, after adding social bot self-efficacy to the regression model for protective strategies, the coefficient for social bot literacy decreased to 0.516 ($p<0.001$), while social bot self-efficacy showed a significant positive relationship with the use of protective strategies ($b=0.421$, $p<0.001$).

Furthermore, the bias-corrected percentile bootstrap procedure confirmed that the indirect effect of social bot literacy on protective strategies through social bot self-efficacy was significant ($ab=0.288$, $SE=0.022$, 95% $CI=[0.209, 0.296]$). The model explained a substantial portion of the variance in protective strategy use ($R^2=0.597$, $F=341.491$, <0.001). These results meet the criteria for establishing a mediating effect, as outlined by MacKinnon (2008). Therefore, H4 is supported, demonstrating that social bot self-efficacy significantly mediates the relationship between social bot literacy and users' protective strategies.

Predictors	Model 1 (Protective strategies)		Model 2 (Social bot self-efficacy)		Model 3 (Protective strategies)	
	β	t	β	t	β	T
Age	0.006	1.516	0.003	0.768	0.005	1.311
Gender	-0.081	-1.456	-0.149	-2.553*	-0.018	-0.361
Edu	0.048	1.026	0.115	2.346*	-0.001	-0.016
NSBL	0.804	33.600***	0.684	27.142***	0.516	18.789***
SBE					0.421	16.795***
R2	0.498		0.398		0.597	
F	286.550***		190.910***		341.491***	
Note. NSBL: Social bot literacy; SBE: Social bot self-efficacy						
*p < 0.05, **p < 0.01, ***p < 0.001						

Table 4. The mediation effect of social bot literacy on protective strategies.

H5 predicted that social bot self-efficacy would mediate the relationship between social bot literacy and the use of pressuring strategies. As shown in Table 5, we conceptualized social bot literacy as the independent variable, pressuring strategies as the dependent variable, and social bot self-efficacy as the mediating variable within the mediation model. Model 1 indicated that social bot literacy positively influenced the use of pressuring strategies ($b=0.767$, $p<0.001$), demonstrating a significant total effect. Model 2 revealed a strong positive association between social bot literacy and social bot self-efficacy ($b=0.650$, $p<0.001$).

In Model 3, after incorporating social bot self-efficacy into the regression model for pressuring strategies, the regression coefficient for social bot literacy decreased to 0.510 ($p=0.001$), while social bot self-efficacy continued to show a significant positive association with pressuring strategies ($b=0.376$, $p<0.001$). The bias-corrected percentile bootstrap procedure further confirmed that the indirect effect of social bot literacy on the use of pressuring strategies through social bot self-efficacy was significant ($ab=0.257$, $SE=0.023$, 95% C [0.170, 0.265]). Together, social bot literacy and social bot self-efficacy explained a significant portion of the variance in pressuring strategy use ($R^2=0.502$, $F=232.374$, $p<0.001$). These results meet the criteria for establishing a mediating effect, as outlined by MacKinnon (2008).

Thus, H5 is supported, demonstrating that social bot self-efficacy mediates the relationship between social bot literacy and pressuring strategies. However, it is worth noting that the effect size for this mediation ($ab/c=33.519\%$) is smaller than that observed in H4 ($ab/c=35.815\%$), indicating a weaker mediation effect for pressuring strategies compared to protective strategies.

Predictors	Model 1 (Pressuring strategies)		Model 2 (Social bot self-efficacy)		Model 3 (Pressuring strategies)	
	β	t	β	T	β	t
Age	0.007	1.673	-0.003	0.768	0.0025	0.6010
Gender	-0.151	-2.482*	-0.149	-2.553*	-0.095	-1.670
Edu	-0.064	-1.248	0.115	2.346*	-0.107	-2.236*
NSBL	0.767	29.136***	0.650	27.142***	0.510	16.212***
SBE					0.376	13.094***
R2	0.428		0.398		0.502	
F	215.735***		190.910***		232.374***	
Note. NSBL: Social bot literacy; SBE: Social bot self-efficacy						
*p < 0.05, **p < 0.01, ***p < 0.001						

Table 5. The mediation effect of social bot literacy on pressuring strategies

Testing for moderated mediation effects

RQ1 asked whether perceived threats from bots moderate the relationship between social bot self-efficacy and users' protective strategies for responding to bots. To address this question, we followed the recommendations of Hayes (2017) and employed the PROCESS macro (Model 14) to assess the moderated mediation effect, as outlined in Table 6. Models 1, 2, and 3 examined the moderation effects of perceived threats on: (1) the relationship between social bot literacy and the use of protective strategies; (2) the relationship between social bot literacy and social bot self-efficacy; and (3) the relationship between social bot self-efficacy and the use of protective strategies.

The results of Model 1 revealed a significant main effect of social bot literacy on protective strategies ($b=0.332$, $p<0.001$). In Model 2, social bot literacy was found to significantly affect social bot self-efficacy ($b=0.684$, $p<0.001$). Model 3 demonstrated a significant main effect of social bot self-efficacy on protective strategies ($b=0.587$, $p<0.001$), with the effect being moderated by perceived threats ($b=-0.112$, $p<0.001$).

Predictor s	Model 1 (Protective strategies)		Model 2 (Social bot self-efficacy)		Model 3 (Protective strategies)	
	β	t	β	t	β	t
Age	0.009	2.769**	0.003	0.768	0.008	2.472*
Gender	0.026	0.572	-0.149	-2.553*	0.001	0.028
Edu	-0.004	-0.114	0.115	2.346*	-0.006	-0.162
NSBL	0.332	11.370***	0.684	27.142***	0.115	3.355**
SBE					0.587	10.760***
Pth	0.497	11.592***			0.752	16.903***
Pth*SBE	0.018	3.991***			-0.112	-7.641***
R ²	0.660		0.398		0.691	
F	372.124***		190.910***		367.283***	

Note. NSBL: Social bot literacy; SBE: Social bot self-efficacy; Pth: Perceived threat
 *p < 0.05, **p < 0.01, ***p < 0.001

Table 6. The moderated mediation effect of social bot literacy on protective strategies

As shown in Figure 2, simple slope tests indicated that for respondents with higher perceived threats from bots, the positive relationship between social bot self-efficacy and protective strategies was weaker compared to those with lower levels of perceived threat.

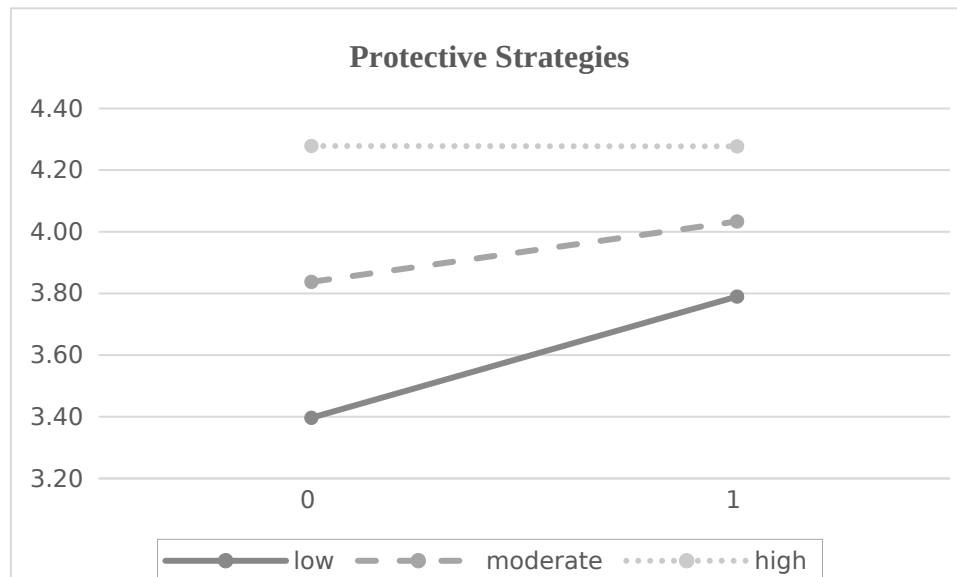


Figure 2. The moderation effect of perceived threat on protective strategies

To further investigate the moderated mediation effect, an analysis of the indirect effect was conducted at different levels of perceived threat. The results, presented in Table 7, show a significant moderated mediation model, indicating that the relationship between social bot literacy and protective strategies through social bot self-efficacy was moderated by perceived threats from bots (Moderated Mediation index=-0.126, Boot SE=0.0145, 95% CI=[-0.1543, -0.098]). Specifically, for respondents with higher perceived threats, the indirect effect was not

significant. However, for those with lower perceived threats, the indirect effect of social bot literacy on the use of protective strategies was both significant and stronger ($b=0.200$, Boot SE=0.025, 95% CI=[0.153, 0.250]) compared to respondents with moderate levels of perceived threats ($b=0.110$, Boot SE=0.021, 95% CI=[0.069, 0.153]). Based on the above analysis, the mediating effect varies across different levels, indicating a moderated mediation effect.

Perceived threat	Effect	Boot SE	Bootstrap 95% CI	
			LL	UL
M-1SD	0.2149	0.0312	0.1560	0.2760
M	0.089	0.0312	0.0292	0.1512
M+1SD	0.0574	0.0323	-0.0039	0.1218
Index of moderated mediation	-0.126	0.0145	-0.1543	-0.0980

Table 7. Indirect effect on different levels of perceived threat

RQ2 examined whether perceived threats from bots moderate the relationship between social bot self-efficacy and users' pressuring strategies for responding to bots. To address this question, we followed Hayes' (2017) recommendations and assessed the parameters for three regression models using the PROCESS macro (Model 14), focusing on the moderated mediation effect (see Table 8). Models 1, 2, and 3 respectively evaluated the moderation effects of perceived threats on: (1) the relationship between social bot literacy and the use of pressuring strategies, (2) the relationship between social bot literacy and social bot self-efficacy, and (3) the relationship between social bot self-efficacy and the use of pressuring strategies.

Model 1 revealed a significant positive relationship between social bot literacy and the use of pressuring strategies ($b=0.331$, $p<0.001$). Model 2 showed that social bot literacy was significantly associated with social bot self-efficacy ($b=0.684$, $p<0.001$). Model 3 demonstrated a significant main effect of social bot self-efficacy on pressuring strategies ($b=0.436$, $p<0.001$), with this effect also being moderated by perceived threats ($b=-0.077$, $p<0.01$). Additionally, the results from 5,000 bootstrap samples confirmed that the perceived threats significantly influenced the mediation, as the 95% confidence interval was entirely above zero. Additionally, the results of 95% CI from 5000 bootstrap samples were statistically greater than 0, illustrating that the perceived threat could moderate the relationship between social bot self-efficacy and use of pressuring strategies.

Predictor s	Model 1 (Pressuring strategies)		Model 2 (Social bot self-efficacy)		Model 3 (Pressuring strategies)	
	β	t	β	t	β	t
Age	0.010	2.626**	0.003	0.768	0.009	2.408*
Gender	-0.052	-0.959	-0.149	-2.553*	-0.070	-1.326
Edu	-0.112	-2.492*	0.1273	3.1508*	-0.114	-2.564*
NSBL	0.331	9.678***	0.684	27.142***	0.170	4.107***
SBE					0.436	6.619***
Pth	0.459	9.772***			0.642	11.948***
Pth*SBE	0.028	3.334**			-0.077	-4.324**
R2	0.558		0.398		0.574	
F	242.538***		190.910** *		221.873***	
Note. NSBL: Social bot literacy; SBE: Social bot self-efficacy; Pth: Perceived threat *p < 0.05, **p < 0.01, ***p < 0.001						

Table 8. The moderated mediation effect of social bot literacy on pressuring strategies

As illustrated in Figure 3, the simple slope tests revealed that, among respondents with higher perceived threats from bots, the positive relationship between social bot self-efficacy and pressuring strategies was weaker compared to those with lower levels of perceived threat.

To further evaluate the moderated mediation effect, an indirect effect analysis was performed at varying levels of perceived threat. The results, shown in Table 9, indicate a significant moderated mediation model where the relationship between social bot literacy and the use of pressuring strategies through social bot self-efficacy was moderated by perceived threats from bots (Moderated Mediation index=-0.052, Boot SE=0.013, 95% CI=[-0.079, -0.028]). For respondents with higher levels of perceived threat, the indirect effect was not significant. However, for those with moderate and lower levels of perceived threat, the indirect effect of social bot literacy on the use of pressuring strategies was significant (moderate level: b = 0.099, Boot SE = 0.024, 95% CI=[0.053, 0.146]) (lower level: b=0.161, Boot SE=0.026, 95% CI=[0.111, 0.212]). These results indicate that the mediating effect varies across different levels, indicating a moderated mediation effect.

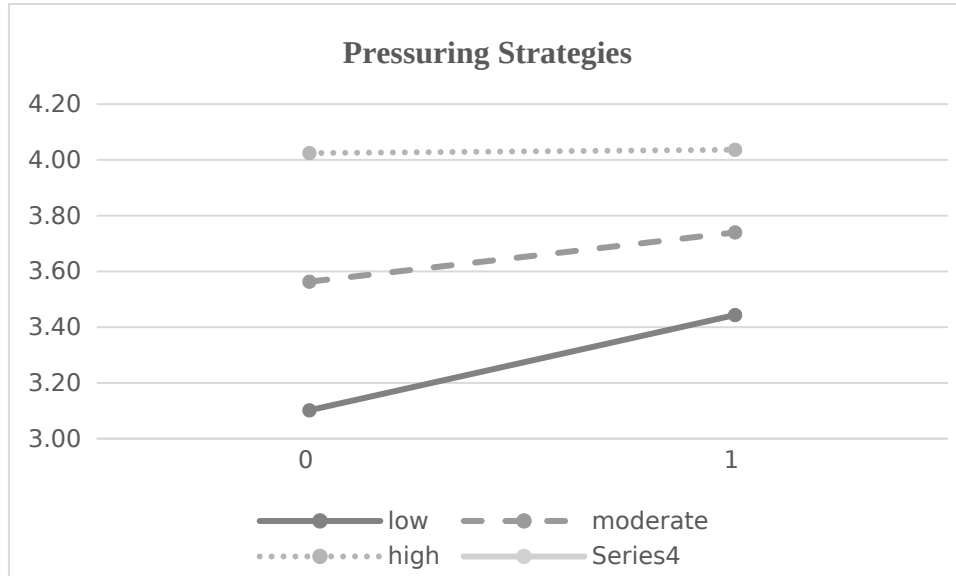


Figure 3. The moderation effect of perceived threat on pressuring strategies

Perceived threat	Effect	Boot SE	Bootstrap 95% CI	
			LL	UL
M-1SD	0.161	0.026	0.111	0.212
M	0.099	0.024	0.053	0.146
M+1SD	0.037	0.030	-0.023	0.097
Index of moderated mediation	-0.052	0.013	-0.079	-0.028

Table 9. Indirect effect on different levels of perceived threat

Finally, Figure 4 below visualizes our final moderated mediation model.

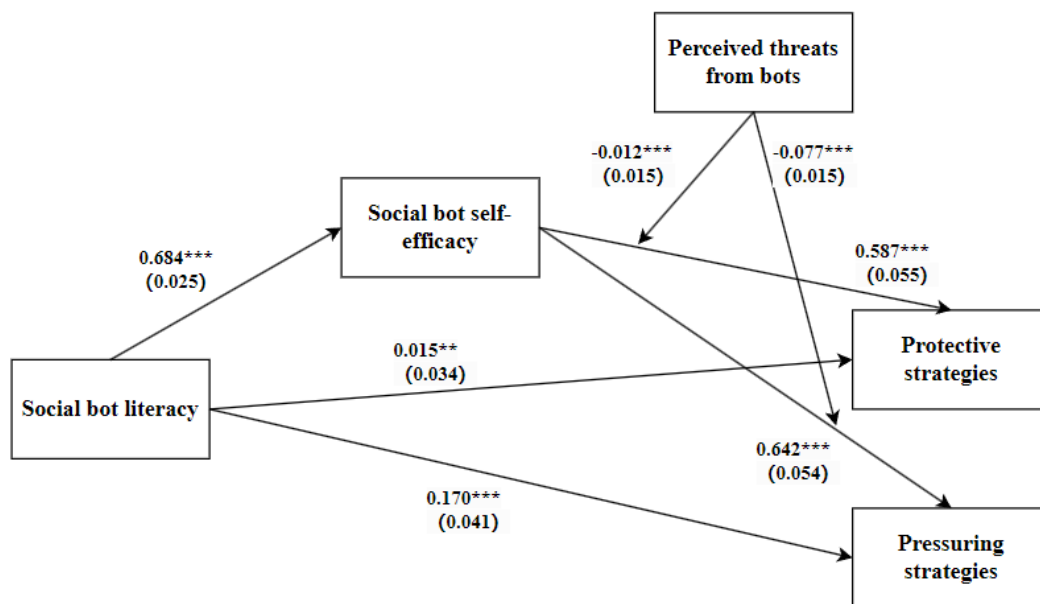


Figure 4. The moderation effect of perceived threat on pressuring strategies

Qualitative findings: insights from in-depth interviews

To strengthen the mixed-method design, we conducted in-depth, semi-structured interviews with 20 Sina Weibo users (8 males and 12 females), aged between 20 and 40 years, all holding at least a college degree. The interviews explored users' bot-related knowledge and detection skills, perceptions of threats from social bots, and strategies for responding to them. Through grounded-theory analysis, we developed themes that are presented as an analytical extension of the quantitative model, clarifying the conditions and meanings underlying the observed associations—especially those involving self-efficacy and threat perception.

Users' understanding of bot literacy and self-efficacy

This subsection examines how participants define and apply social bot literacy and how this knowledge underpins a sense of self-efficacy in managing bots.

(1) Users' awareness and knowledge of social bots

Participants generally demonstrated a solid level of literacy, reflected in both their awareness of bots' presence on Weibo and their knowledge of bot identification techniques. Most participants demonstrated clear awareness of the pervasive presence of bots on Weibo. Cheng (male, 24, law major) remarked, "I've heard of them—there are actually quite a lot of bots out there." Hu (female, 31, real estate professional) described encountering bots not only in her follower list but also in comments on posts by others. Awareness was often mentioned as preceding more refined recognition skills.

Participants frequently described concrete cues for identifying bots, such as abnormal usernames, generic profile images, repetitive posting patterns, and templated or illogical comment content. Meng (female, 22) observed: "The usernames of bot accounts are just weird. No real person would pick a name like that—it feels like a string of code." Zhang (female, 23) added: "Their comments are often similar, as if they're following a script. The language is off, and their usernames are a mix of Chinese characters and random numbers—that's a sign it might be a bot." Some participants offered typologies of bots based on observed behaviour. Yang (female, 23) categorised them as: marketing bots tied to specific accounts, celebrity-related bots flooding posts about certain stars, and spam-advertising bots posting under unrelated comments. Such

classifications suggest that participants' literacy extends beyond recognition to a functional understanding of bot ecosystems.

The sources of bot literacy varied. Some participants learned indirectly from news coverage or online discussions about bot detection. Others built their knowledge through repeated direct encounters with suspicious accounts. Jin (male, 22), who primarily followed anime and gaming content, recalled: 'At first, I didn't know how to tell bots apart and even got misled. But over time, I noticed patterns—like accounts always posting the same thing. Eventually, I realized those were bots'. In some cases, prior use of bot services provided insider perspectives. Melissa (female, 38), for example, had purchased fake followers and engagement for her own account, which sharpened her understanding of bot operations. In some cases, indirect and direct learning experiences reinforced each other. Duo (female, 26), for example, initially thought she had found like-minded users who quickly commented on her posts. Only later, after reading an article about keyword-triggered bot activity, did she realize the comments were likely generated by bots.

(2) Connections between bot literacy and users' confidence

A recurring theme was that higher literacy often corresponded with greater confidence in identifying and managing bots. Jin (male, 22), for instance, offered a detailed and confident description of his identification process:

First, I check the username—bots usually have weird combinations of letters or numbers, something a real person wouldn't use. Then I look at the profile picture, most of them are just generic or seem mass-produced. I also look at their comments—if they say something totally unrelated to the post or if the same kind of comments appear repeatedly from different accounts, then it's likely a bot. And if they praise something that most people dislike, that's also suspicious.

This multi-step method, built through repeated exposure, reinforced his belief in his ability to discern authentic from inauthentic accounts. This pattern was echoed by Duo (female, 26), who initially engaged passively with bots but, after learning about bot behaviours through media reports, became more confident in recognizing and managing them. These narratives indicate an iterative learning process in which indirect knowledge is tested and reinforced through direct encounters, progressively associated with enhanced perceived efficacy.

In summary, this subsection highlights how social bot literacy was described not only as static knowledge but also as a dynamic foundation associated with confidence in managing bots.

User strategies for responding to bots

(1) Protective strategies: ignoring, blocking, and selective engagement

Many participants described protective actions aimed at minimizing unwanted interaction with bots, with ignoring and blocking being the most frequently mentioned. Cai (male, 22), a programmer, summarized a common approach: 'If I realize it's a bot, I usually just ignore their comments. I might also report them or block the user'. While this reflects a general avoidance stance, other accounts revealed that responses were not uniform but instead contingent on the context in which bots were encountered.

For instance, Meng (female, 22) highlighted how the spatial location of bot activity shaped her reaction: 'If the bot shows up in my followers, I'll definitely block it. But if it's in someone else's list, I'll just ignore it'. Similarly, Duo (female, 26) distinguished between public and private encounters: 'If I see them in public threads, I'll ignore them. But if a bot comments on my post, I might delete the comment. If they follow me, I'll remove them. Basically, I ignore them unless they show up in my personal space'.

Overall, protective strategies appeared to vary depending on situational judgments. Intrusions into personal accounts often prompted blocking or deletion, whereas bots in public threads were more likely to be ignored. This context-sensitive logic highlights how perceived online boundaries mediate the link between social bot literacy, self-efficacy, and protective behaviours.

(2) Pressuring strategies: reporting and expectations of platform action

Beyond personal protection, some participants also described pressuring responses—including reporting bots, adjusting privacy settings, and voicing expectations toward platform governance. For some, reporting was seen as the most immediate and actionable step. Cheng (male, 24) emphasized this as his primary reaction: 'Of course I report them. That's the first thing I'd do'. However, his response extended beyond individual action to a broader governance perspective, arguing that responsibility lies not just with users but also with institutional actors: 'There are two main parties responsible for the spread of bots: the government and the platforms. The government should create regulations, and the platforms should use technology to detect and manage bot accounts'.

This view was echoed by Teng (male, 36), a media professional, who framed platform responsibility in terms of both capacity and accountability: 'The platform is the main player here. They benefit from traffic and have the capacity to identify bots in bulk. Ordinary users can't keep up with the volume—reporting bots one by one is unrealistic. It's the platform's responsibility'. These perspectives reveal an underlying belief that effective bot mitigation requires coordinated efforts between users and institutional actors, with platforms holding a central role.

However, not all participants shared this optimism about institutional intervention. Several expressed scepticism about platforms' willingness to act, especially when bot activity aligns with commercial incentives. Yang (female, 23) voiced such doubts bluntly: 'I don't think Weibo will do anything about it. Bots are a revenue stream—they sell accounts and make money from this. So why would they stop?'. This sentiment points to a perceived misalignment between platform profit motives and user interests, which may limit the impact of pressuring strategies.

These perspectives indicate that while some confident users engaged in pressuring strategies, their willingness to do so depended on their trust in institutional responsiveness. High-literacy, confident users may report or push for platform action, but such efforts hinge on perceived institutional responsiveness and accountability.

Perceived threats and user responses

This final subsection examines how perceived threats—both personal and societal—mediate the relationship between literacy, self-efficacy, and action. The findings reveal a nuanced, sometimes counterintuitive pattern: heightened threat perception can, in certain contexts, suppress rather than stimulate active resistance.

(1) Perceived threats at the individual and societal levels

At the individual level, many participants expressed concern about how social bots compromise their personal information environment. The most frequently cited issue was the overload of redundant promotional content—particularly marketing for brands or celebrities—generated by bots. Such content, they argued, diminished information authenticity and reduced efficiency in locating relevant information (e.g., Teng, male, 36). This concern was not limited to promotional overload; several participants also highlighted the intrusion of irrelevant bot-generated posts into their personal social networks, which they saw as a direct encroachment on their online space. Duo (female, 26) illustrated this distinction vividly:

If a bot ends up in my *followers* list, it's not a big deal—I usually just remove it. But if it's in my *following* list, it's annoying, because whatever they post might show up on

my homepage. I think people get angrier when bots are pushed into the accounts they follow rather than into their follower list.

Her account shows how perceived threat at the individual level was shaped not only by content volume but also by its proximity to one's own curated online environment.

At the societal level, concerns shifted from personal inconvenience to the broader implications for public discourse. Several participants voiced unease over bots' ability to manipulate public opinion by disseminating misleading or biased information. This influence was often described as subtle but cumulative. Zhang, for instance, likened it to an invisible pressure that shapes viewpoints without overtly declaring its intent:

I feel like they still influence my views—it's this invisible, silent effect. They generate a certain volume of voices, and that can skew perspectives or even control the direction of the discussion.

Such accounts suggest that participants recognized a dual-layered threat: the immediate, tangible disruption to personal information flow, and the diffuse, systemic influence on societal opinion climates.

(2) Perceived threats in relation to patterns of user responses

Interview data revealed that the perceived severity of these threats played a pivotal role in shaping how social bot literacy and self-efficacy translated into actual responses. For participants who appraised the threat as low, higher literacy and efficacy tended to encourage proactive measures. Yang (female, 23), a heavy Weibo user who spends 5–6 hours daily on the platform for both news and social interaction, had accumulated substantial bot-related knowledge and exhibited high self-efficacy. However, she viewed bots as posing minimal risk—mainly generating redundant content rather than meaningfully shaping opinions:

Bots don't really have opinions. They might seem to, but it's AI-generated stuff like 'I totally agree with your post'—it's not a real stance. They can't produce any meaningful value or opinions, so there's no real influence.

Her coping strategies reflected this low-threat appraisal:

If a bot follows me, I'll remove it. If it comments, I might delete the comment. But if it's under someone else's post, I just ignore it—as if it's not there.

In contrast, participants who perceived bots as highly threatening often reported more restrained or avoidant behaviours, even when they possessed high literacy and confidence. Zhang (female, 23) admitted: 'I usually just ignore them. I might have blocked one or two, but that's rare'. Similarly, Cai (male, 22), an automation major with deep technical knowledge of bot operations, expressed strong concerns about their disruptive potential and his passive response:

They flood you with redundant, biased content. It reduces the comprehensiveness, objectivity, and efficiency of getting information. The best way to deal with them is to ignore—bots have no emotions, but humans do. Arguing with bots just gets you upset, so it's better to ignore or report them.

Some participants went further, suggesting that extreme bot proliferation could drive them to abandon the platform altogether. Duo (female, 26) reflected this ultimate form of disengagement: 'If the threat got too big and beyond what I could control, I might just stop using the platform altogether'.

These narratives reveal a counterintuitive pattern consistent with the quantitative findings: higher perceived threat was not always linked to more active resistance. Instead, strong threat perceptions were frequently associated with avoidance, suggesting that excessive concern may reduce the likelihood of proactive coping.

Discussion and implications

Discussion

Our findings indicate that higher perceived social bot literacy is associated with stronger engagement in both protective and pressuring strategies. Users who see themselves as knowledgeable in identifying bots and experienced in interacting with them were more likely to avoid bot interactions or take actions such as adjusting account settings and reporting to platforms. This aligns with prior work showing that users' awareness, past encounters, and detection skills are linked to their coping behaviours (Schmuck & von Sikorski, 2020; Wischnewski *et al.*, 2024; Yan *et al.*, 2021; Kenny *et al.*, 2024; Ngo *et al.*, 2023). Similar associations have been reported in studies on algorithmic literacy and resistance to algorithmic harms (DeVrio *et al.*, 2024).

This study further shows that social bot self-efficacy mediates the relationship between literacy and user responses. Individuals who report higher literacy also tend to report greater confidence in distinguishing bots from human accounts, consistent with prior research linking knowledge and experience to self-efficacy (Fang and Nie, 2024; Ngo *et al.*, 2023; Schmuck and von Sikorski, 2020; Yan *et al.*, 2021). In turn, stronger self-efficacy is associated with both protective and pressuring strategies, whether through avoidance or reporting. Although research directly connecting bot self-efficacy to responses is limited, evidence on detection performance and post-identification reactions (Yan *et al.*, 2021, 2023; Wischnewski *et al.*, 2024) supports this pattern. Notably, the mediating effect of self-efficacy is stronger for pressuring than protective strategies, while the direct effect of literacy is also more pronounced for pressuring responses.

Our results indicate that perceived threats from bots serve as a boundary condition for the link between self-efficacy and user responses. Specifically, this association is weaker among users who view bots as highly threatening. Prior studies echo this pattern: exposure to bot threats has been shown to diminish the correlation between self-efficacy and detection performance (Yan *et al.*, 2023), and stronger threat perceptions have been linked to a reduced connection between literacy and perceived control over bots (Fang and Nie, 2024). These findings suggest that threat perceptions shape the extent to which knowledge and confidence are reflected in behavioural responses.

The qualitative findings further elaborate the quantitative model by showing how these associations are experienced in practice. Participants described bot literacy as a cumulative process shaped by both direct encounters and indirect learning, with perceived self-efficacy developing through repeated recognition and reflection. Their responses were strongly context-dependent: protective actions were more likely when bots entered one's personal digital space, whereas public-space exposure was more often met with passive avoidance; pressuring actions were more likely when users believed platforms or regulators would respond. The interviews also clarify the moderation pattern: lower perceived threat was associated with more active coping, while higher perceived threat was often associated with withdrawal or limited engagement, even among users reporting high literacy and confidence.

Across methods, the findings present a coherent but nuanced pattern. The survey model identifies robust associations among social bot literacy, self-efficacy, response strategies, and perceived threat, while interviews clarify how these associations are enacted in context. In interview accounts, protective responses were often stronger when bot activity entered users' personal account space; pressuring responses depended partly on perceived platform

accountability; and high perceived threat was at times linked to disengagement rather than stronger action. This integrative evidence gives greater prominence to users' situated reasoning and strengthens interpretation of the moderated pathways.

Implications

Theoretically, this study advances research on human–bot interaction by foregrounding user empowerment in response to automated accounts. Our findings indicate that social bot literacy and self-efficacy are closely linked to both protective and pressuring strategies, extending prior work on algorithmic resistance (DeVrio et al., 2024). Importantly, the results underscore the conditional role of perceived threats: while literacy and confidence generally relate to active responses, elevated threat perceptions may weaken these associations, echoing evidence from prior studies (Fang & Nie, 2024; Yan et al., 2023). By combining survey data with supplementary qualitative insights, this study contributes to a more nuanced understanding of the psychological mechanisms—literacy, efficacy, and perceived threat—that shape resistance practices.

From a practical perspective, the results highlight two complementary pathways for addressing bot-related challenges. First, at the individual level, educational initiatives can strengthen public awareness and confidence. Public campaigns that disseminate information about bots' features, risks, and detection tools, as suggested by Yang et al., (2019), can help users develop protective and pressuring strategies.

Limitations and future directions

This study examined the relationship between social bot literacy and user responses but did not address how literacy itself can be enhanced. Prior work has shown that factors such as news coverage (Schmuck & von Sikorski, 2020) and social media use (Fang & Nie, 2024) may influence literacy, yet systematic evidence remains limited. Future research could explore these factors using diverse approaches—surveys, interviews, and especially experimental designs—to identify effective ways of strengthening social bot literacy. Such work would provide valuable insights for literacy interventions and policy initiatives.

Second, this study focused on Weibo users, who tend to be younger, urban, and more educated (Weibo, 2021). While Weibo shares similarities with platforms such as X, bot behaviours may differ across other platforms, including e-commerce (e.g., Ali), review (e.g., Dianping), and video-sharing (e.g., TikTok, Kuai). Broader comparative studies across platforms and demographic groups would help clarify how context shapes human–bot interaction.

Third, protective and pressuring strategies were measured with only two items each. Although adapted from validated scales, the limited scope may constrain construct coverage. Future studies should employ richer measurement tools to capture the multidimensional nature of users' strategies.

Finally, given the cross-sectional design and reliance on regression analysis, the study can only identify associations rather than causal effects. Reverse causality is also possible—for instance, active engagement with bots may shape perceptions of literacy, rather than literacy alone influencing responses. In addition, unobserved variables, such as prior exposure to bot-related news or platform interventions, may affect both literacy and responses. Future research should address these issues by employing longitudinal surveys, experiments, or mixed methods to better disentangle these dynamics.

Conclusion

This study examines the associations between social bot literacy and users' responses, with attention to the mediating role of self-efficacy and the moderating effect of perceived threats. The findings show that higher literacy is linked to greater confidence in distinguishing bots,

which in turn correlates with more frequent protective and pressuring strategies. Yet, this association weakens when bots are perceived as highly threatening, underscoring the complex interplay between knowledge, confidence, and risk perception. Theoretically, the study contributes to understanding resistance practices in human–bot interaction by highlighting the conditions under which literacy and efficacy relate to user strategies. Practically, the results offer guidance for policymakers and platform designers: public education campaigns can raise awareness and improve detection skills, while social media platforms should provide accessible detection tools and streamlined reporting channels to support user engagement.

Acknowledgements

The authors gratefully acknowledge the anonymous reviewers for their insightful comments and constructive feedback, which have significantly enhanced this manuscript. We also extend our sincere appreciation to all survey participants for their time and valuable contributions.

About the authors

Wei Fang is an Associate Professor in the College of Communication Arts and Sciences at Beijing Information Science & Technology University, China. She received her PhD from the University of Bristol, UK. Her research focuses on the effects of AI technologies (e.g., algorithms and social bots) on wellbeing and social change. She can be contacted at fangwei@bistu.edu.cn.

Chen Nie is an Associate Professor at the School of Public Administration at Beihang University, China. He received his PhD from the University of Bristol, UK. His research focuses on human–AI interactions among young people. He is the corresponding author of this article and can be contacted at niechen@buaa.edu.cn.

References

- Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology*, 32(4), 665–683. <https://doi.org/10.1111/j.1559-1816.2002.tb00236.x>
- Al-Rawi, A., Kane, O., & Bizimana, A. (2021). Topic modelling of public Twitter discourses, part bot, part active human user, on climate change and global warming. *Journal of Environmental Media*, 2(1), 31–53. https://doi.org/10.1386/jem_00039_1.
- Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2), 122. <https://doi.org/10.1037/0003-066X.37.2.122>.
- Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology*, 4(3), 359–373. <https://doi.org/10.1521/jscp.1986.4.3.359>.
- Bossu, R., Corradini, M., Cheny, J., & Fallou, L. (2023). A social bot in support of crisis communication: 10-years of @ LastQuake experience on Twitter. *Frontiers in Communication*, 8. <https://doi.org/10.3389/fcomm.2023.992654>
- Confessore, N., Dance, G.J., Harris, R., & Hansen, M. (2018). The follower factory. *The New York Times*, 27. <https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html>. (Archived by the Internet Archive at <https://web.archive.org/web/20180128001510/https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html>)

- Cresci, S., Yang, K., Spognardi, A., Di Pietro, R., Menczer, F., & Petrocchi, M. (2023). Demystifying misconceptions in social bots research. *ArXiv Preprint ArXiv:2303.17251*.
<https://doi.org/10.48550/arXiv.2303.17251>.
- Daume, S., Galaz, V., & Bjersér, P. (2023). Automated framing of climate change? the role of social bots in the twitter climate change discourse during the 2019/2020 Australia Bushfires. *Social Media+ Society*, 9(2). <https://doi.org/10.1177/20563051231168370>
- DeVrio, A., Eslami, M., & Holstein, K. (2024). Building, Shifting, & Employing Power: A taxonomy of responses from below to algorithmic harm. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1093–1106.
<https://dl.acm.org/doi/10.1145/3630106.3658958>.
- Duan, Z., Li, J., Lukito, J., Yang, K.-C., Chen, F., Shah, D. V., & Yang, S. (2022). Algorithmic agents in the hybrid media system: Social bots, selective amplification, and partisan news about COVID-19. *Human Communication Research*, 48(2), 1-18.
<https://doi.org/10.1093/hcr/hqac012>
- Fan, R., Talavera, O., & Tran, V. (2020). Social media bots and stock markets. *European Financial Management*, 26(3), 753–777. <https://doi.org/10.1111/eufm.12245>.
- Fang, W., & Nie, C. (2024). Social media use, social bot literacy, perceived threats from bots, and perceived bot control: a moderated-mediation model. *Behaviour & Information Technology*, 43(13), 3271–3287. <https://doi.org/10.1080/0144929X.2023.2276801>
- Ferrara, E. (2023). Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*, 28(6). <https://doi.org/10.5210/fm.v28i6.13185>
- Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96–104. <https://doi.org/10.1145/2818717>.
- Gorwa, R., & Guilbeault, D. (2020). Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, 12(2), 225–248. <https://doi.org/10.1002/poi3.184>.
- Hagen, L., Neely, S., Keller, T.E., Scharf, R., & Vasquez, F.E. (2022). Rise of the machines? Examining the influence of social bots on a political discussion network. *Social Science Computer Review*, 40(2), 264–287. <https://doi.org/10.1177/0894439320908190>.
- Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2022). Social bots and the spread of disinformation in social media: the challenges of artificial intelligence. *British Journal of Management*, 33(3):1238–1253. <https://doi.org/10.1111/1467-8551.12554>.
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.
- Hays, C., Schutzman, Z., Raghavan, M., Walk, E., & Zimmer, P. (2023). Simplistic collection and labeling practices limit the utility of benchmark datasets for Twitter bot detection. In *Proceedings of the ACM web conference 2023*. <https://doi.org/10.48550/arXiv.2301.07015>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>.
- Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education*, 28, 1–35. <https://doi.org/10.1007/s10212-011-0097-y>.

- Kelfve, S., Kivi, M., Johansson, B., & Lindwall, M. (2020). Going web or staying paper? The use of web-surveys among older people. *BMC Medical Research Methodology*, 20, 252.
<https://doi.org/10.1186/s12874-020-01138-0>
- Keller, F.B., Schoch, D., Stier, S., & Yang, J. (2020). Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political Communication*, 37(2), 256-280.
<https://doi.org/10.1080/10584609.2019.1661888>.
- Kenny, R., Fischhoff, B., Davis, A., Carley, K.M., & Canfield, C. (2024). Duped by bots: why some are better than others at detecting fake social media personas. *Human Factors*, 66(1), 88-102.
<https://doi.org/10.1177/00187208211072642>
- Kušen, E., & Strembeck, M. (2019). Something draws near, I can feel it: An analysis of human and bot emotion-exchange motifs on Twitter. *Online Social Networks and Media*, 10, 1-17.
<https://doi.org/10.1016/j.osnem.2019.04.001>.
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science*, 359(6380), 1094-1096.
<https://doi.org/10.1126/science.aao2998>.
- Livingstone, S., & Helsper, E. (2010). Balancing opportunities and risks in teenagers' use of the internet: The role of online skills and internet self-efficacy. *New Media & Society*, 12(2), 309-329. <https://doi.org/10.1177/1461444809342697>.
- Marlow, T., Miller, S., & Roberts, J. T. (2021). Bots and online climate discourses: Twitter discourse on President Trump's announcement of US withdrawal from the Paris Agreement. *Climate Policy*, 21(6), 765-777. <https://doi.org/10.1080/14693062.2020.1870098>
- Martin, A. (2006). A European framework for digital literacy. *Nordic Journal of Digital Literacy*, 1(2), 151-161. <https://doi.org/10.18261/ISSN1891-943X-2006-02-06>
- Martini, F., Samula, P., Keller, T. R., & Klinger, U. (2021). Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data & Society*, 8(2), 54050782.
<https://doi.org/10.1177/20539517211033566>
- Ngo, T., Wischniewski, M., Bernemann, R., Jansen, M., & Krämer, N. (2023). Spot the bot: Investigating user's detection cues for social bots and their willingness to verify Twitter profiles. *Computers in Human Behavior*, 146, 107819.
<https://doi.org/10.1016/j.chb.2023.107819>.
- Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. In *Proceedings of the National Academy of Sciences*, 119(8), e2120481119.
<https://doi.org/10.1073/pnas.2120481119>.
- Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*, 44(4), 409-432.
<https://doi.org/10.1023/A:1024232915870>
- Schmuck, D., & von Sikorski, C. (2020). Perceived threats from social bots: The media's role in supporting literacy. *Computers in Human Behavior*, 113, 106507.
<https://doi.org/10.1016/j.chb.2020.106507>.
- Sharma, K., Ferrara, E., & Liu, Y. (2022). Characterizing online engagement with disinformation and conspiracies in the 2020 US presidential election. In *Proceedings of the International*

- AAAI Conference on Web and Social Media, 16(1), 908–919. <https://doi.org/10.1609/icwsm.v16i1.19345>.
- Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. <https://doi.org/10.1145/3600211.3604673>
- Shi, W., Liu, D., Yang, J., Zhang, J., Wen, S., & Su, J. (2020). Social bots' sentiment engagement in health emergencies: A topic-based analysis of the COVID-19 pandemic discussions on Twitter. *International Journal of Environmental Research and Public Health*, 17(22), 8701. <https://doi.org/10.3390/ijerph17228701>
- Starbird, K. (2019). Disinformation's spread: Bots, trolls, and all of us. *Nature*, 571(7766), 449–450. <https://doi.org/10.1038/d41586-019-02235-x>
- Stocking, G., & Sumida, N. (2018). Social media bots draw public's attention and concern. Pew Research Center. <https://www.pewresearch.org/journalism/2018/10/15/social-media-bots-draw-publics-attention-and-concern/> (Archived by the Internet Archive at <https://web.archive.org/web/20181015174009/https://www.pewresearch.org/journalism/2018/10/15/social-media-bots-draw-publics-attention-and-concern/>)
- Suarez-Lledo, V., & Alvarez-Galvez, J. (2022). Assessing the role of social bots during the COVID-19 pandemic: Infodemic, disagreement, and criticism. *Journal of Medical Internet Research*, 24(8), e36085. <https://doi.org/10.2196/36085>
- Tan, Z., Feng, S., Sclar, M., Wan, H., Luo, M., Choi, Y., & Tsvetkov, Y. (2023). BotPercent: Estimating bot populations in Twitter communities. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.2302.00381>
- Thomala, L.L. (2023). Most Popular News on Weibo in China 2022. Statista. <https://www.statista.com/statistics/1377073/china-most-popular-news-on-weibo/#statisticContainer>. (Archived by the Internet Archive at <https://web.archive.org/web/20231001000000/https://www.statista.com/statistics/1377073/china-most-popular-news-on-weibo/>)
- Wagner, C., Mitter, S., Körner, C., & Strohmaier, M. (2012). When social bots attack: Modeling susceptibility of users in online social networks. In *Proceedings of the 2012 IEEE/ACM International Conference on Multimodal Search and Mining*. https://ceur-ws.org/Vol-838/paper_11.pdf (Archived by the Internet Archive at https://web.archive.org/web/20130101000000/https://ceur-ws.org/Vol-838/paper_11.pdf)
- Wald, R., Khoshgoftaar, T.M., Napolitano, A., & Sumner, C. Predicting susceptibility to social bots on twitter. In *Proceedings of the 2013 IEEE 14th international conference on information reuse and integration*. <https://doi.org/10.1109/IRI.2013.6642447>.
- Weibo. (2021). Weibo 2020 User Development Report. <https://data.weibo.com/report/reportDetail?id=456>. (Archived by the Internet Archive at <https://web.archive.org/web/20210101000000/https://data.weibo.com/report/reportDetail?id=456>)
- Wischnewski, M., Ngo, T., Bernemann, R., Jansen, M., & Krämer, N. (2024). "I agree with you, bot!" How users (dis)engage with social bots on Twitter. *New Media & Society*, 26(3), 1505–1526. <https://doi.org/10.1177/14614448211072307>

- Yan, H. Y., Yang, K., Menczer, F., & Shanahan, J. (2021). Asymmetrical perceptions of partisan political bots. *New Media & Society*, 23(10), 3016–3037.
<https://doi.org/10.1177/1461444820942744>
- Yan, H. Y., Yang, K., Shanahan, J., & Menczer, F. (2023). Exposure to social bots amplifies perceptual biases and regulation propensity. *Scientific Reports*, 13(1), 20707.
<https://doi.org/10.1038/s41598-023-46630-x>
- Yang, K. C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1), 48–61. <https://doi.org/10.1002/hbe2.145>.

Copyright

Authors contributing to *Information Research* agree to publish their articles under a [Creative Commons CC BY-NC 4.0 license](https://creativecommons.org/licenses/by-nc/4.0/), which gives third parties the right to copy and redistribute the material in any medium or format. It also gives third parties the right to remix, transform and build upon the material for any purpose, except commercial, on the condition that clear acknowledgment is given to the author(s) of the work, that a link to the license is provided and that it is made clear if changes have been made to the work. This must be done in a reasonable manner, and must not imply that the licensor endorses the use of the work by third parties. The author(s) retain copyright to the work. You can also read more at: <https://publicera.kb.se/ir/openaccess>