Manuscript for:

Title: Personal Moderation Configurations on Facebook: Exploring the Role of FoMO, Social Media Addiction, Norms, and Platform Trust

Author: Shagun Jhaver

Journal: First Monday

Version: Author's Submitted Manuscript (ASM)

Status: Accepted on Aug 21, 2025

Personal Moderation Configurations on Facebook: Exploring the Role of FoMO, Social Media Addiction, Norms, and Platform Trust

Shagun Jhaver, Rutgers University

shagun.jhaver@rutgers.edu

Abstract

Personal moderation tools on social media platforms let users control their news feeds by configuring acceptable toxicity thresholds for their feed content or muting inappropriate accounts. This research examines how four critical psychosocial factors – fear of missing out (FoMO), social media addiction, subjective norms, and trust in moderation systems – shape Facebook users' configuration of these tools. Findings from surveying a nationally representative sample of 1,061 participants show that FoMO and social media addiction make Facebook users more vulnerable to content-based harms by reducing their likelihood of adopting personal moderation tools to hide inappropriate posts. In contrast, descriptive and injunctive norms positively influence the use of these tools whereas trust in Facebook's moderation systems also significantly affects users' engagement with personal moderation. This analysis highlights qualitatively different pathways through which FoMO and social media addiction make affected users disproportionately unsafe and offers solutions to address this challenge.

Keywords

content moderation; platform governance; online harm; psychosocial factors; FoMO

1. Introduction

Social media sites empower users worldwide by letting them create and share the content of their choice. The rules determining acceptable content on these platforms often reflect the cultural norms predominant in Silicon Valley, where most major digital platforms are located (Gillespie, 2018). However, norms of appropriate conduct vary widely across different cultures and

communities, so relying on a universal platform-driven approach to regulate online content overlooks the diverse requirements of millions of users (Chandrasekharan et al., 2018; Gorwa et al., 2020). Recognizing this, some scholars have called for an alternative approach, *personal moderation*, defined as a "form of moderation in which users can configure or customize some aspects of their moderation preferences on social media." (Jhaver et al., 2023)

Today, personal moderation tools are critical tools that social media platforms offer to let users choose their moderation preferences (Feng et al., 2024; Jhaver et al., 2023). These tools play a crucial role in reducing exposure to content-based harms, such as hate speech, harassment, and violent content (Jhaver et al., 2022). Examples of these tools include *mute functionality*, which lets users stop an account from appearing in their news feed, and *word filters*, which allow users to configure a set of keywords they do not want to see on their feed. Such tools empower users to align the content they view with their tolerance for sensitive spectatorship (Tait, 2008).

Recognizing the utility of these tools, previous research has thus far analyzed how users configure them (Alqabandi et al., 2024; Feng et al., 2024; Jhaver & Zhang, 2023); designed and built novel personal moderation tools (Jhaver et al., 2022); and examined how changes in their design can produce more valuable outcomes for users (Jhaver et al., 2023). However, we do not yet clearly understand the psychological dispositions and social dynamics that contribute to users' willingness or reluctance to adopt such tools in the first place. Given that these tools substantially shape user experience and safety on social media sites, it is vital that we understand the factors that affect their use.

I fill this critical gap by exploring how four key psychosocial factors regarding media use contribute to Facebook users' configuration of personal moderation tools. These factors include

(1) *Fear of Missing Out (FoMO),* (2) *social media addiction,* (3) *subjective norms, and* (4) *trust in content moderation systems*. These factors (detailed below) have been widely examined in prior literature regarding their relation to individuals' social media attitudes and behaviors, and they correlate with related factors of interest, e.g., prior research has linked FoMO to problematic social media use (Fang et al., 2020). Yet, their impact of users' engagement with moderation mechanisms remains empirically underexplored. The analysis of the influence of these four psychosocial factors will allow me to build a first and exploratory multivariate model to assess the adoption of personal moderation tools.

Building upon my findings from a nationally representative survey in the US, I discuss how FoMO and social media addiction add new dimensions of vulnerability to online harms for the affected users. I reflect on how platforms and communities can address this challenge by reinforcing subjective norms regarding the use of these tools and fostering trust in content moderation systems. While prior framings of personal moderation have often emphasized its delegation of content curation responsibility solely to the individual end-users (Jhaver et al., 2023; Jhaver & Zhang, 2023), this research complicates that view by pointing to the institutional, affective, and relational forces that co-produce users' interactions with personal moderation tools.

2. Background and Related Work

2.1. Content-based Harm, Online Safety, and Personal Moderation Tools

This research focuses on *content-based harm*, which refers to harm caused by viewing undesirable online content (Jhaver et al., 2022). Content-based harm has been documented across social media platforms like Reddit and Twitter (Sowles et al., 2018), video-sharing platforms like YouTube and TikTok (Lewis et al., 2012), gaming platforms like Twitch and Xbox, and elsewhere (Jhaver et al., 2018). Prior research has shown that continuous exposure to violent, hateful, or otherwise troubling posts can negatively affect mental health, including inducing panic attacks and secondary trauma (Roberts, 2019).

Platforms usually address content-based harm with content moderation. When users post content that violates platform rules, platforms impose sanctions, such as removing that content or banning that user's account (Schoenebeck et al., 2021). Traditionally, those issuing sanctions may be commercial content moderators employed by the platform or community content moderators, who are volunteer end-users invested in their community's success (Seering et al., 2019). Previous studies have explored the limitations and challenges introduced by these existing strategies to mitigate content-based harms and promote online safety. For example, Pater et al. (2016) showed that online platforms often have inconsistent and non-exhaustive standards for addressing such harms. This can lead to frustration among targets who report content they believe to be in violation, only to discover later that the attack falls outside the scope of the remediation (Blackwell et al., 2017).

Some scholars have taken a victim-centric approach to addressing content-based harm, studying victims' experiences and perspectives on various aspects, including the classification of harm (Blackwell et al., 2017), the impact of harm (Scheuerman et al., 2021), and effective ways to address it (Jhaver et al., 2018). Xiao et al. (2023) found that current approaches fail to sufficiently remove disturbing materials afflicting victims, like fake news, alt-right trolls, and revenge porn, and further perpetuate harm by directing offenders' attention to the punishment they receive instead of the damage they cause. By examining the diverse needs of victims, including subjective differences in what content is deemed problematic (Jhaver et al., 2023), researchers have explored interventions beyond platform-enacted content moderation. Examples include outsourcing the filtering of problematic content to friends (Mahar et al., 2018), providing tools to help victims gather authentic evidence of harm to share publicly (Sultana et al., 2021), or encouraging offenders to apologize to their victims (Xiao et al., 2023).

<Figure 1 here>

I add to these prior efforts by focusing on the use of personal moderation tools that let users *proactively prevent* content-based harms (Jhaver et al., 2023). While some critics argue that customized moderation mechanisms could reinforce echo chambers and violate freedom of expression, most scholars point out that they are critically needed to protect vulnerable groups (Jhaver et al., 2018; Sultana et al., 2021). Prior inquiries on designing and deploying these tools have demonstrated that they can increase users' safety perceptions, accommodate their individualized moderation preferences, and enhance their participation online without infringing on free speech values (Jhaver et al., 2018; Jhaver et al., 2023). A recent nationally representative survey of 984 US adults showed that end-users prefer personal moderation tools over the default platform-enacted moderation to prevent content-based harms resulting from exposure to hate

speech, violent content, and sexually explicit content (Jhaver & Zhang, 2023). The current article advances this line of research by examining the psychosocial factors impacting users' propensity to configure these tools to reduce content-based harms.

Jhaver et al. (2023) classified personal moderation tools into two types: content-based and account-based. *Content-based tools* let users configure moderation preferences based on the content of each post (see Figure 1). For example, *word filter tools* allow any user to configure a set of undesired keywords (Figure 1, left); once set up, posts containing any of these keywords are automatically hidden from the user's news feed (Jhaver et al., 2022). Another category of content-based tools is *sensitivity controls* (Figure 1, right), which let users configure their moderation preferences on a Likert-type scale over factors like content sensitivity or toxicity (Jhaver et al., 2023). Given the growing deployment of toxicity sliders across major social media platforms such as Instagram and Tumblr, this study examines the use and configuration of *toxicity* sliders.

Unlike content-based tools, *account-based tools* (Figure 2) let users restrict their interaction with an individual account or a set of accounts (Geiger, 2016; Jhaver et al., 2018). For example, muting an account hides all subsequent posts from it to a user's news feed. Since muting is available as a feature on almost all major platforms, this work examines users' preferences for muting inappropriate accounts.

<Figure 2 here>

2.2. Fear of Missing Out (FoMO)

The Fear of Missing Out (FoMO) is defined as "a pervasive apprehension that others might be having rewarding experiences from which one is absent." (Przybylski et al., 2013) Błachnio and Przepiórka (2018) characterize FoMO as "a fundamental human motivation that consists in [of] craving interpersonal attachments." Such attachments can be impeded by social exclusion, which is frequently associated with the experience of social pain (Lai et al., 2016). Studies show that it is a widespread phenomenon, with 56% - 70% of adults suffering from FoMO (Murphy, 2013; Westin & Chiasson, 2021). Prior research has identified FoMO as a public health concern, linking it to stress, depression, anxiety (Elhai et al., 2016; Tugtekin et al., 2020), headaches (Baker et al., 2016), and decreased sleep (Milyavskaya et al., 2018).

While FoMO was initially conceptualized in the offline context (Przybylski et al., 2013), it has found widespread applicability regarding social media use (Bloemen & De Coninck, 2020; Reer et al., 2019). Over the past decade, researchers have established its evidentiary relationship with online vulnerability (Thompson et al., 2021), fake news and misinformation sharing (Talwar et al., 2019), and social media fatigue (Malik et al., 2020). Some researchers have also examined the connection between platform users' FoMO and their interactions with system design. For example, Westin and Chiasson (2021) showed that social media users are systematically pressured into privacy-compromising behaviors, such as posting more information more often, due to FoMO. Popovac and Hadlington (2020) demonstrated that FoMO is a significant predictor of online risk-taking behaviors, such as sexting and sharing passwords with friends, among adolescents. I build upon this research to examine whether FoMO influences users' online safety practices by reducing their ability to restrict *any* content from their feeds.

I argue that FoMO might influence users' inclination to limit inappropriate content because of their fear of missing future potential information. Users with greater FoMO may also prefer to avoid eliminating any social ties to stay connected for future interaction opportunities. Therefore, I explore how FoMO affects users' configuration of personal moderation tools that proactively reduce online harm.

2.3. Social Media Addiction

Social media addiction refers to "the irrational and excessive use of social media to the extent that it interferes with other aspects of daily life." (Hou et al., 2019) Prior literature has shown that users suffering from social media addiction exhibit most behavioral addiction symptoms, including tolerance, withdrawal, conflict, salience, relapse, and mood modification (Griffiths et al., 2014). Social media addiction is associated with greater loneliness, anxiety, and suicidality and with declines in academic performance, self-esteem, and life satisfaction (Hawi & Samaha, 2016; Latikka et al., 2022). Increased social media use also raises individuals' potential exposure to various online vulnerabilities, including online harassment, incidents of data misuse, interactions with strangers with harmful intentions, and exposure to inappropriate content due to spending more time on these sites (Brandtzæg et al., 2010; Sasson & Mesch, 2014; Staksrud et al., 2013).

Contemporary social theories offer frameworks to explain the links between social media addiction and safety practices (Craig et al., 2020). According to Problem Behavior Theory (Jessor & Jessor, 1977), risky behaviors tend to co-occur, and certain individuals have distinct characteristics that heighten their susceptibility to engaging in such behaviors (Craig et al., 2020). For example, Huang et al. (2023) demonstrated a pathway between social media addiction

and food addiction with the involvement of psychological distress. I examine whether social media addiction relates to another risky behavior: configuring inadequate moderation controls. Further, drawing from social learning theory, individuals who spend more time on social media may observe more offensive behaviors and, through role modeling and reinforcement, may come to see them as more acceptable (Lee et al., 2023). Repeated exposure to offensive behaviors may also result in the "disinhibition effect," (Barlett & Gentile, 2012) i.e., inappropriate behaviors may become more normalized over time. This suggests that users with social media addiction may see offensive content as acceptable and hesitate to remove it from their feeds proactively. To examine this, I study how social media addiction influences users' choices regarding personal moderation tools.

2.4. Subjective Norms

According to Ajzen (1991), subjective norms refer to the social pressure associated with performing a given behavior. This pressure may stem from two categories of subjective norms: descriptive and injunctive (Ajzen, 2020). Descriptive norms are beliefs about whether important others themselves perform the behavior under consideration, whereas injunctive norms are expectations about whether others approve or disapprove of that behavior (Ajzen, 2020; Kiesler et al., 2012). The underlying assumption is that people generally engage in behaviors that are encouraged and embraced within their social sphere.

Prior research has reported that subjective norms significantly affect individuals' behaviors in a variety of social media contexts, such as posting behaviors (Arpaci, 2020), privacy regulation (Neubaum et al., 2023), and preventing unruly conversations (Matias, 2019). Subjective norms also contribute to engagement in negative behaviors, such as taking risky selfies (Chen et al.,

2019), excusing the use of aggressive language (Allison et al., 2019), and conducting cyberbullying (Heirman & Walrave, 2012). Recent literature has specifically acknowledged the role of subjective norms in people's responses to fake news, health misinformation, and conspiratorial content (Bautista et al., 2022; Koo et al., 2021). Building upon this literature, I examine how subjective norms shape users' preferences for addressing content-based harms through configuring personal moderation tools.

2.5. Trust in the Moderation System

Personal moderation tools are automated tools whose operations rely on the efficacy of algorithmic mechanisms that drive them. Prior research has shown users' awareness of the use of automation in enacting personal moderation (Jhaver et al., 2018; Jhaver et al., 2023). Users of any AI application must have the confidence that they can depend on the AI agent to accomplish their objectives in situations of uncertainty (Okamura & Yamada, 2020). The "Computers are Social Actors" paradigm suggests that viewing AI applications as human-like collaborators rather than tools can clarify our understanding of human trust in AI (Seeber et al., 2020). Prior literature on user acceptance of AI systems emphasizes interpersonal trust as an essential component (Gillath et al., 2021); such trust encompasses the willingness of one party to accept vulnerability based upon positive expectations of the behavior of another party, irrespective of the ability to monitor that party (Lewicki et al., 2006).

Accordingly, I conceptualize trust in personal moderation tools as the extent to which a user is confident and comfortable in the actions or decisions made by these tools and is, therefore, willing to rely on them. This includes faith in the judgment and fairness of the platform's

moderation system to determine the appropriateness of submitted posts and confidence that these tools' decisions would align with users' own determinations.

Recent controversies over Facebook's privacy invasions have affected users' trust in its moderation system (Brown, 2020). This could, in turn, influence the use of the platform's safety tools. Prior research has examined how users' trust in algorithmic moderation tools shapes their perspectives about content moderation decisions (Jhaver et al., 2019; Molina & Sundar, 2022; Schulenberg et al., 2023). I build upon this research to examine the extent to which users' trust in the moderation apparatus affects their configuration of personal moderation tools.

Integrating the concepts reviewed in this section, I examine the following research question in this article:

> *How do FoMO, social media addiction, subjective norms, and trust in moderation systems influence end-users' configuration of personal moderation tools?*

3. Method

For this study, which was considered exempt from review by the Rutgers University's IRB, I recruited participants via Lucid,[i] a survey company that provides researchers access to demographically representative national samples. This survey's inclusion criteria encompassed all adult internet users within the United States. Compensation for participants was facilitated through the Lucid system.

I framed this survey's questionnaire (see Appendix) around Facebook because its widespread popularity made it more likely that many users would be aware of terms relevant to the concepts this study explores, such as 'Facebook friends' and 'news feeds.' By focusing on Facebook, this

paper also adds to vital research (Alhabash & Ma, 2017; Mena, 2020; Paradise & Sullivan, 2012; Théro & Vincent, 2022) into this key social network site's (actual and perceived) social implications and use.

I designed survey questions to examine how four key psychosocial factors related to media use – FoMO, social media addiction, subjective norms, and trust in moderation systems – shape Facebook users' adoption or configuration of two personal moderation tools: (1) sensitivity controls and (2) the muting function. In doing so, I adapted survey instruments from pertinent prior research to evaluate specific measures, which I explain below. To enhance the survey's validity, I solicited input on early versions of the questionnaire from peers and students within my institution. Eight individuals external to the project, trained in diverse fields like Computer Science, Psychology, and Media Studies, offered responses and suggestions regarding question phrasing and survey structure. I integrated these insights into the survey design. Next, I conducted a trial run of the survey with 30 participants from Lucid, which prompted further refinement of the questionnaire.

I conducted the survey through the online survey platform Qualtrics, with the survey going live on Jan 4, 2024. Since this survey focuses on Facebook use, I screened for participants who had used Facebook over the past year. I also implemented data-cleaning steps to improve the quality of analyzed survey responses. For example, I removed responses where participants engaged in straightlining (Kim et al., 2018), i.e., selecting identical answers (e.g., 'Strongly agree') to items in every question that uses the same response scale. Table 1 shows the demographic details of my final sample, which comprised 1,061 participants following data refinement. This table also compares the demographics of my sample to those of adult internet users in the United States.

<Table 1 here>

3.1. Measures

I built hierarchical linear regression models to examine the relationships between this study's independent variables (FoMO, Facebook addiction, subjective norms (injunctive and descriptive), and trust in moderation) and dependent variables (sensitivity control setting and likelihood of muting). I describe below how I used survey items to measure each variable along with the variable's mean (M), standard deviation (SD), and Cronbach's alpha (α) values from the processed survey data. I also note the socio-demographic variables I controlled for in these models.

3.1.1. Fear of Missing Out (FoMO)

Participants responded to the ten-item Fear of Missing Out scale (Przybylski et al., 2013) with answer choices ranging from 1 = "Not at all true of me" to 5 = "Extremely true of me." This scale measures the extent of apprehension of missing out on the rewarding experiences of others. Example items include, "When I miss out on a planned get-together, it bothers me." and "I get worried when I find out my friends are having fun without me." The scale produced an average score ranging from 1 to 5, with higher scores indicating increased levels of FoMO. Consistent with past research (Przybylski et al., 2013), this FoMO scale showed good internal consistency (M = 2.01, SD = .73, α = .85).

3.1.2. Facebook Addiction

Facebook addiction was measured using the widely used 6-item Facebook Bergen Addiction Scale (FBAS) (Andreassen et al., 2012). Participants were prompted to "Please answer the following questions with regard to your Facebook use over the past year." I retained the use of

'Facebook' in these items because the rest of the survey also referred to Facebook as the site of focus. Each item in the FBAS corresponds to one of the six central components of addiction according to the Griffiths et al. (2014) model: salience, mood modification, tolerance, withdrawal symptoms, conflict, and relapse. For example, the item concerning withdrawal asked, "Have you become restless or troubled if you have been prohibited from using Facebook?" Participants responded on a 5-point Likert scale ranging from 1 = "Very rarely" to 5 = "Very often." These items were averaged to create a single measure and showed good internal consistency (M = 2.03, SD = .84, α = .84).

3.1.3. Subjective Norms

Subjective norms regarding the configuration of personal moderation tools were assessed based on items adopted from prior research (Bautista et al., 2022; Pundir et al., 2021). Participants rated each of these items on a Likert-type scale ranging from 1 (Strongly disagree) to 7 (Strongly agree).

Injunctive norms regarding sensitivity controls and the muting function were assessed by the following items, respectively: (a) "Suppose Facebook offers a moderation setting that limits the number of upsetting or offensive posts appearing on my news feed. Most people who are important to me would expect me to turn on this setting,"[ii] and (b) "Most people who are important to me would expect me to mute a Facebook friend's account if it frequently posts upsetting or offensive content." These two items were averaged to create an index for injunctive norms (M = 3.84, SD = 1.66, α = .74)

Similarly, descriptive norms regarding sensitivity controls and the muting function were assessed by the following items, respectively: (a) "Suppose Facebook offers a moderation setting that

limits the number of upsetting or offensive posts appearing on the news feed. Most people who are important to me would turn on this setting in their Facebook profile," and (b) "Most people who are important to me would mute a Facebook friend's account if it frequently posted upsetting or offensive content." Averaging these items created an index for descriptive norms (M = 4.31, SD = 1.53, α = .78).

3.1.4. Trust in the Moderation Process

This variable was operationalized using four items, each on a Likert-type scale ranging from 1 (Strongly disagree) to 7 (Strongly agree). First, faith in the judgment and fairness of the platform's moderation system was assessed by the following items: (a) "Facebook's content moderation process can be trusted to judge how upsetting or offensive each post is," and (b) "Facebook's content moderation process is fair and impartial in determining how upsetting or offensive each post is." Second, the expected alignment of users' content evaluations with Facebook's moderation was measured by asking: "If Facebook's content moderation process determines that a post is upsetting or offensive, I am likely to find it upsetting or offensive." Finally, users' confidence in Facebook's account-based moderation tools operating as expected was measured by asking: "I feel confident that if I mute an account on Facebook, its posts will no longer appear on my news feed." These four items were averaged to create a measure of general trust in Facebook moderation (M = 3.91, SD = 1.25, α = .77).

<Figure 3 here>

3.1.5. Dependent Variables Related to Personal Moderation

The adoption of personal moderation tools was assessed using two items. First, I asked for sensitivity controls: "Suppose that Facebook provides a setting that lets you control the volume of posts you may find upsetting or offensive on your news feed. How would you configure this setting?" Options included: (a) Allow, (b) Limit, and (c) Limit even more (Figure 3). This design and options were inspired by a similar setting offered by Instagram (Figure 1). Second, I asked for account-based tools: (a) "Suppose that you encounter a Facebook friend frequently posting upsetting or offensive content that appears on your news feed. How likely are you to mute this friend's account?" I added a note with this question that explained how "muting" works. This item was rated on a 7-point Likert-type scale ranging from 1 (Extremely unlikely) to 7 (Extremely likely).

3.1.6. Control Variables

Prior literature demonstrated the relationship between socio-demographic variables and attitudes toward media regulation (Gunther, 2006; Jhaver & Zhang, 2023; Lambe, 2002). Drawing from this literature, I controlled for age, education, gender, race, and political affiliation (1 = "very liberal," 7 = "very conservative") in my analyses.

4. Results

I began by examining the descriptive statistics regarding the configuration of personal moderation tools. Results show that 34.3%, 38.9%, and 26.8% of participants would prefer to configure their content-based moderation setting to 'Allow', 'Limit', and 'Limit even more'

levels, respectively (Figure 4). Further, 29.5% of participants are at least slightly unlikely to mute an offensive Facebook friend, whereas 47.8% are at least slightly likely to mute in such a case (Figure 5).

<Figure 4 here>

<Figure 5 here>

4.1. Support for Stricter Content-based Personal Moderation

I computed hierarchical linear regression to test how different factors shape preferences for configuring Facebook's sensitivity controls. I created a model in which the dependent variable was the strictness of the tool setting that participants selected. In Step 1, I included the control variables age, gender, race, education, and political affiliation. In Step 2, I introduced five independent variables: (1) FoMO, (2) Facebook addiction, (3) Injunctive norms, (4) Descriptive norms, and (5) Trust in Facebook moderation (*Table 2*).

The regression model (Model 1) shows a significant negative influence of Facebook addiction on preference for setting up stricter sensitivity controls (β = -.081, p < .05). Further, injunctive norms (β = .174, p < .001), descriptive norms (β = .207, p < .001), and trust in Facebook moderation (β = .145, p < .001) all positively influence a preference for configuring a stricter setting. FoMO did not have a significant influence on participants' configurations of this content-based moderation tool.

<Table 2 here>

4.2. Likelihood of Engaging in Account-based Personal Moderation

Next, I computed hierarchical linear regression to test how different factors shape participants' likelihood of muting an account that frequently posts upsetting or offensive content. I created a model in which the dependent variable was the likelihood of muting such an account. As in Model 1, in Step 1, I included the control variables age, gender, race, education, and political affiliation. In Step 2, I introduced five independent variables: (1) FoMO, (2) Facebook addiction, (3) Injunctive norms, (4) Descriptive norms, and (5) Trust in Facebook moderation (*Table 2*).

This regression model (Model 2) shows a significant negative influence of fear of missing out (FoMO) on the likelihood of muting (β = -.095, p < .01). Additionally, injunctive norms (β = .158, p < .001), descriptive norms (β = .282, p < .001), and trust in Facebook moderation (β = .134, p < .001) all positively influence the muting action. Facebook addiction did not influence participants' likelihood of muting.

5. Discussion

Personal moderation tools represent a growing shift in the governance of online spaces – platforms are now increasingly delegating content curation responsibility to end-users themselves. These tools promise greater control, flexibility, and protection from content-based harms (Feng et al., 2024; Jhaver et al., 2023; Jhaver & Zhang, 2023). Yet, as this study shows, simply making these tools available does not guarantee that users would benefit from them. Instead, their adoption is deeply conditioned by users' psychological orientations, normative environments, and levels of trust in platform governance.

My analysis shows that the fear of missing out (FoMO) is associated with a reduced likelihood of muting an account frequently posting offensive or upsetting content. When considering whether to mute an account, users make a tradeoff: should they avoid exposure to harmful content from that account or risk losing access to relevant content posted by that account? A significant relationship between FoMO and muting behavior identified here is a theoretically significant finding: it suggests that *FoMO makes users potentially more vulnerable to content-based harms by deterring in-the-moment actions against offensive accounts*. This adds another layer to previous findings on how FoMO contributes to online safety practices, such as privacy-

compromising and risk-taking behaviors (Popovac & Hadlington, 2020; Westin & Chiasson, 2021).

I also found that *social media addiction is linked to a significant reduction in the strictness levels users select for their sensitivity controls*. This hesitation to embrace a vital safety affordance aligns with Problem Behavior Theory (Jessor & Jessor, 1977), which predicts that risky behaviors tend to co-occur. As prior research (Brandtzæg et al., 2010; Sasson & Mesch, 2014; Staksrud et al., 2013) points out, social media addiction raises exposure to inappropriate content just because the addicted users spend more time on these sites. My analysis suggests that these users' vulnerability is further increased by their hesitance to configure stricter controls in moderation toggles. It is likely that, as Barlett and Gentile (2012) observe, repeated exposure to inappropriate content normalizes it for addicted users, and this disinhibition affects their moderation configurations. This raises concern that *social media addiction could create a vicious cycle whereby affected users are disproportionately exposed to online harms and become less likely to take actions that reduce this exposure.* Therefore, design and policy efforts must focus on reducing the prevalence of social media addiction. For instance, platforms could offer heavy users personalized psychological and mental health information about social media usage, prompting them to set and track their site usage goals (Cham et al., 2019).

How can platforms encourage users to engage in safety practices grounded in personal moderation choices? My analysis indicates that both *descriptive and injunctive norms are significantly associated with users setting up stricter settings in moderation toggles and muting inappropriate accounts*. These results open up a design space for platforms to incorporate normative information in their sites as a key strategy to promote online safety. For example, platforms may show users aggregated descriptive statistics of how their linked accounts (e.g.,

Facebook “friends” or Twitter “followees”) set up their sensitivity controls. These information nudges may normalize and enhance the use of personal moderation tools and reduce users’ exposure to content-based harms. These findings also suggest that users and communities can encourage the adoption of personal moderation tools among their social connections as an online safety mechanism by rendering the norms about their use more salient, e.g., by sharing how they help reduce content-based harm. While personal moderation tools empower end-users, configuring them requires digital literacy and cognitive labor (Jhaver et al., 2023). Thus, it is crucial that besides these tools, platforms also invest in improving site-wide moderation procedures to address online harms equitably.

My findings also show that *trust in Facebook moderation significantly relates to users’ adoption of stricter content-based moderation toggles and their muting of inappropriate accounts.* Recent controversies about social media moderation decisions (Brown, 2020) and increased scrutiny of moderation infrastructures by news media, lawmakers, and scholars have diluted the general public’s trust in social media platforms (Gillespie, 2018). The relationship between this trust and users’ moderation practices documented here highlights the institutional dimension of online safety – even when the moderation power is devolved to individual users, confidence in the judgment, fairness, and reliability of platform infrastructures remains a crucial determinant of users’ action. Therefore, it is vital that platforms take concrete and conspicuous actions to clarify their commitment to user safety and highlight the robustness of their content moderation mechanisms.

Taken together, my findings highlight that users do not make personal moderation decisions in a vacuum; instead, they are influenced by a range of social, affective, and institutional forces. Moreover, psychological vulnerabilities likely result in uneven uptake of personal moderation

tools, such that users most in need of these tools may end up not using them. Thus, any moderation approach to online safety that emphasizes personal choice must move beyond merely providing technical affordances and toward cultivating the normative, educational, and trust-building conditions necessary for meaningful user empowerment.

5.1. Limitations and Future Work

The survey method deployed in this study was not intended to explore participants' specific motivations behind their different preferences for personal moderation. Further, this cross-sectional study cannot draw definitive conclusions about causal relationships. Instead, my investigation serves as an initial exploration of how some of the most relevant psychosocial factors about social media use identified in prior literature relate to users' moderation actions. Future research can build upon this work by delving into other factors, such as digital literacy and past experiences with moderation tools, that may affect participants' moderation configurations.

This survey asked participants to report how they would react in hypothetical scenarios. Participants may behave differently in their actual social media use. Future studies with access to users' activity logs and moderation tool settings can offer more reliable evidence.

Finally, the survey questions were deliberately tied to a specific platform, Facebook, to ensure that participants could better comprehend and more concretely answer survey questions about their social media activities and preferences. Future work should evaluate how these findings apply to other social media platforms.

6. Conclusion

This study shows how FoMO and social media addiction decrease users' inclination to engage in proactive approaches to reducing content-based harm. Thus, it highlights a qualitatively different pathway, i.e., greater exposure to upsetting content, through which users experiencing FoMO and social media addiction could be disproportionately vulnerable to online harm. This paper also explores new directions for how platforms and community members can help such users. First, my findings suggest that platforms should offer these tools by default, and end-users and communities should normalize their use. Second, this study motivates design and policy efforts to reduce FoMO and social media addiction. Third, it demonstrates how user trust in content moderation systems plays a crucial role in the adoption of defensive tools and motivates further acceleration of efforts to foster this trust.

About the Author

Shagun Jhaver is a social computing scholar whose research focuses on improving content moderation on digital platforms. He is currently studying how internet platform design and moderation policies can address societal issues such as online harassment, misinformation, and the rise of hate groups. Jhaver is an assistant professor at Rutgers University's School of Communication & Information in the Library & Information Science Department.

Acknowledgements

I thank Kiran Garimella and Koustuv Saha for their feedback on this project. This work was supported by the National Science Foundation Award 2329394.

[i] https://lucidtheorem.com

[ii] I posed this question as a hypothetical because Facebook currently does not offer such a setting, although its sister company Instagram offers it.

References

Icek Ajzen. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

Icek Ajzen. (2020). The theory of planned behavior: Frequently asked questions. *Human Behavior and Emerging Technologies*, *2*(4), 314-324. https://doi.org/10.1002/hbe2.195

Saleem Alhabash, & Mengyan Ma. (2017). A Tale of Four Platforms: Motivations and Uses of Facebook, Twitter, Instagram, and Snapchat Among College Students? *Social Media + Society*, *3*(1), 2056305117691544. https://doi.org/10.1177/2056305117691544

Kimberley R. Allison, Kay Bussey, & Naomi Sweller. (2019). 'I'm going to hell for laughing at this': Norms, Humour, and the Neutralisation of Aggression in Online Communities. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), Article 152. https://doi.org/10.1145/3359254

Fatima Alqabandi, Graham Tierney, Christopher Bail, Sunshine Hillygus, & Alexander Volfovsky. (2024). Experiments offering social media users the choice to avoid toxic political content.

Cecilie Schou Andreassen, Torbjørn Torsheim, Geir Scott Brunborg, & Ståle Pallesen. (2012). Development of a Facebook Addiction Scale. *Psychological Reports*, *110*(2), 501-517. https://doi.org/10.2466/02.09.18.Pr0.110.2.501-517

Ibrahim Arpaci. (2020). The Influence of Social Interactions and Subjective Norms on Social Media Postings. *Journal of Information & Knowledge Management*, *19*(03), 2050023. https://doi.org/10.1142/s0219649220500239

Zachary G Baker, Heather Krieger, & Angie S LeRoy. (2016). Fear of missing out: Relationships with depression, mindfulness, and physical symptoms. *Translational Issues in Psychological Science*, *2*(3), 275.

Christopher P Barlett, & Douglas A Gentile. (2012). Attacking others online: The formation of cyberbullying in late adolescence. *Psychology of popular media culture*, *1*(2), 123.

John Robert Bautista, Yan Zhang, & Jacek Gwizdka. (2022). Predicting healthcare professionals' intention to correct health misinformation on social media. *Telematics and Informatics*, *73*, 101864. https://doi.org/10.1016/j.tele.2022.101864

Agata Błachnio, & Aneta Przepiórka. (2018). Facebook intrusion, fear of missing out, narcissism, and life satisfaction: A cross-sectional study. *Psychiatry Research*, *259*, 514-519. https://doi.org/https://doi.org/10.1016/j.psychres.2017.11.012

Lindsay Blackwell, Jill P. Dimond, Sarita Schoenebeck, & Cliff Lampe. (2017). Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *PACMHCI*, *1*(CSCW), 24–21-24–21.

Noor Bloemen, & David De Coninck. (2020). Social Media and Fear of Missing Out in Adolescents: The Role of Family Characteristics. *Social Media + Society*, *6*(4), 2056305120965517. https://doi.org/10.1177/2056305120965517

Petter Bae Brandtzæg, Marika Lüders, & Jan Håvard Skjetne. (2010). Too Many Facebook "Friends"? Content Sharing and Sociability Versus the Need for Privacy in Social

Network Sites. *International Journal of Human–Computer Interaction*, *26*(11-12), 1006-1030. https://doi.org/10.1080/10447318.2010.516719

Allison J. Brown. (2020). “Should I Stay or Should I Leave?”: Exploring (Dis)continued Facebook Use After the Cambridge Analytica Scandal. *Social Media + Society*, *6*(1), 2056305120913884. https://doi.org/10.1177/2056305120913884

Sainabou Cham, Abdullah Algashami, Manal Aldhayan, John McAlaney, Keith Phalp, Mohamed Basel Almourad, & Raian Ali. (2019). Digital Addiction: Negative Life Experiences and Potential for Technology-Assisted Solutions. In Á. Rocha, H. Adeli, L. P. Reis, & S. Costanzo, *New Knowledge in Information Systems and Technologies* Cham.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, & Eric Gilbert. (2018). The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 32-32.

Shuang Chen, Lara Schreurs, Sara Pabian, & Laura Vandenbosch. (2019). Daredevils on social media: A comprehensive approach toward risky selfie behavior among adolescents. *New Media & Society*, *21*(11-12), 2443-2462. https://doi.org/10.1177/1461444819850112

Wendy Craig, Meyran Boniel-Nissim, Nathan King, Sophie D. Walsh, Maartje Boer, Peter D. Donnelly, Yossi Harel-Fisch, Marta Malinowska-Cieślik, Margarida Gaspar de Matos, Alina Cosma, Regina Van den Eijnden, Alessio Vieno, Frank J. Elgar, Michal Molcho, Ylva Bjereld, & William Pickett. (2020). Social Media Use and Cyber-Bullying: A Cross-National Analysis of Young People in 42 Countries. *Journal of Adolescent*

Health, *66*(6, Supplement), S100-S108. https://doi.org/https://doi.org/10.1016/j.jadohealth.2020.03.006

Jon D. Elhai, Jason C. Levine, Robert D. Dvorak, & Brian J. Hall. (2016). Fear of missing out, need for touch, anxiety and depression are related to problematic smartphone use. *Computers in Human Behavior*, *63*, 509-516. https://doi.org/https://doi.org/10.1016/j.chb.2016.05.079

Jie Fang, Xingchao Wang, Zhonglin Wen, & Jianfeng Zhou. (2020). Fear of missing out and problematic social media use as mediators between emotional support from social media and phubbing behavior. *Addictive Behaviors*, *107*, 106430. https://doi.org/https://doi.org/10.1016/j.addbeh.2020.106430

KJ Feng, Xander Koo, Lawrence Tan, Amy Bruckman, David W McDonald, & Amy X Zhang. (2024). Mapping the Design Space of Teachable Social Media Feed Experiences. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems,

R. Stuart Geiger. (2016). Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space [https://doi.org/10.1080/1369118X.2016.1153700]. *Information, Communication & Society*, *19*(6), 787-803. https://doi.org/10.1080/1369118X.2016.1153700

Omri Gillath, Ting Ai, Michael S. Branicky, Shawn Keshmiri, Robert B. Davison, & Ryan Spaulding. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*, 106607. https://doi.org/https://doi.org/10.1016/j.chb.2020.106607

Tarleton Gillespie. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Robert Gorwa, Reuben Binns, & Christian Katzenbach. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance [https://doi.org/10.1177/2053951719897945]. *Big Data and Society*, *7*(1).

Mark D. Griffiths, Daria J. Kuss, & Zsolt Demetrovics. (2014). Chapter 6 - Social Networking Addiction: An Overview of Preliminary Findings. In K. P. Rosenberg & L. C. Feder (Eds.), *Behavioral Addictions* (pp. 119-141). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-407724-9.00006-9

Albert C. Gunther. (2006). Overrating the X-Rating: The Third-Person Perception and Support for Censorship of Pornography. *Journal of Communication*, *45*(1), 27-38. https://doi.org/10.1111/j.1460-2466.1995.tb00712.x

Nazir S. Hawi, & Maya Samaha. (2016). The Relations Among Social Media Addiction, Self-Esteem, and Life Satisfaction in University Students. *Social Science Computer Review*, *35*(5), 576-586. https://doi.org/10.1177/0894439316660340

Wannes Heirman, & Michel Walrave. (2012). Predicting adolescent perpetration in cyberbullying: An application of the theory of planned behavior. *Psicothema*, *24*(4), 614-620.

Yubo Hou, Dan Xiong, Tonglin Jiang, Lily Song, & Qi Wang. (2019). Social media addiction: Its impact, mediation, and intervention. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *13*(1), Article 4. https://doi.org/10.5817/CP2019-1-4

Po-Ching Huang, Janet D. Latner, Kerry S. O'Brien, Yen-Ling Chang, Ching-Hsia Hung, Jung-Sheng Chen, Kuo-Hsin Lee, & Chung-Ying Lin. (2023). Associations between social media addiction, psychological distress, and food addiction among Taiwanese

university students. *Journal of Eating Disorders*, *11*(1), 43. https://doi.org/10.1186/s40337-023-00769-0

Richard Jessor, & Shirley L. Jessor. (1977). *Problem behavior and psychosocial development : a longitudinal study of youth*. Academic Press. https://cir.nii.ac.jp/crid/1130282270780560128

Shagun Jhaver, Darren Scott Appling, Eric Gilbert, & Amy Bruckman. (2019). "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW). https://doi.org/10.1145/3359294

Shagun Jhaver, Quan Ze Chen, Detlef Knauss, & Amy X. Zhang. (2022, 2022). Designing Word Filter Tools for Creator-Led Comment Moderation. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems,

Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, & Eric Gilbert. (2018). Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.*, *25*(2). https://doi.org/10.1145/3185593

Shagun Jhaver, Alice Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, & Amy Zhang. (2023). Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.*(CSCW), Article 289. https://doi.org/10.1145/3610080

Shagun Jhaver, & Amy Zhang. (2023). Do Users Want Platform Moderation or Individual Control? Examining the Role of Third-Person Effects and Free Speech Support in Shaping Moderation Preferences. *New Media & Society*. https://doi.org/10.1177/14614448231217993

Sara Kiesler, Robert Kraut, & Paul Resnick. (2012). Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design*.

Yujin Kim, Jennifer Dykema, John Stevenson, Penny Black, & D. Paul Moberg. (2018). Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys. *Social Science Computer Review*, *37*(2), 214-233. https://doi.org/10.1177/0894439317752406

Alex Zhi-Xiong Koo, Min-Hsin Su, Sangwon Lee, So-Yun Ahn, & Hernando Rojas. (2021). What Motivates People to Correct Misinformation? Examining the Effects of Third-person Perceptions and Perceived Norms. *Journal of Broadcasting & Electronic Media*, *65*(1), 111-134. https://doi.org/10.1080/08838151.2021.1903896

Carlo Lai, Daniela Altavilla, Ambra Ronconi, & Paola Aceto. (2016). Fear of missing out (FOMO) is associated with activation of the right middle temporal gyrus during inclusion social cue. *Computers in Human Behavior*, *61*, 516-521. https://doi.org/https://doi.org/10.1016/j.chb.2016.03.072

Jennifer L. Lambe. (2002). Dimensions of Censorship: Reconceptualizing Public Willingness to Censor. *Communication Law and Policy*, *7*(2), 187-235. https://doi.org/10.1207/S15326926CLP0702_05

Rita Latikka, Aki Koivula, Reetta Oksa, Nina Savela, & Atte Oksanen. (2022). Loneliness and psychological distress before and during the COVID-19 pandemic: Relationships with social media identity bubbles. *Social science & medicine*, *293*, 114674. https://doi.org/https://doi.org/10.1016/j.socscimed.2021.114674

Michelle Hui Lim Lee, Manveen Kaur, Vinorra Shaker, Anne Yee, Rohana Sham, & Ching Sin Siau. (2023). Cyberbullying, Social Media Addiction and Associations with

Depression, Anxiety, and Stress among Medical Students in Malaysia. *International Journal of Environmental Research and Public Health*, *20*(4), 3136. https://www.mdpi.com/1660-4601/20/4/3136

Roy J. Lewicki, Edward C. Tomlinson, & Nicole Gillespie. (2006). Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions. *Journal of Management*, *32*(6), 991-1022. https://doi.org/10.1177/0149206306294405

Stephen P. Lewis, Nancy L. Heath, Michael J. Sornberger, & Alexis E. Arbuthnott. (2012). Helpful or Harmful? An Examination of Viewers' Responses to Nonsuicidal Self-Injury Videos on YouTube. *Journal of Adolescent Health*, *51*(4), 380-385. https://doi.org/https://doi.org/10.1016/j.jadohealth.2012.01.013

Kaitlin Mahar, Amy X. Zhang, & David Karger. (2018). *Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation* Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal QC, Canada. https://doi.org/10.1145/3173574.3174160

Aqdas Malik, Amandeep Dhir, Puneet Kaur, & Aditya Johri. (2020). Correlates of social media fatigue and academic performance decrement: A large cross-sectional study. *Information Technology & People*, *34*(2), 557-580.

Nathan J. Matias. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, *116*(20), 9785–9789-9785–9789.

Paul Mena. (2020). Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet*, *12*(2), 165-183. https://doi.org/https://doi.org/10.1002/poi3.214

Marina Milyavskaya, Mark Saffran, Nora Hope, & Richard Koestner. (2018). Fear of missing out: prevalence, dynamics, and consequences of experiencing FOMO. *Motivation and Emotion*, *42*(5), 725-737. https://doi.org/10.1007/s11031-018-9683-5

Maria D. Molina, & S. Shyam Sundar. (2022). Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society*, *0*(0), 14614448221103534. https://doi.org/10.1177/14614448221103534

Samantha Murphy. (2013). *Report: 56% of Social Media Users Suffer From FOMO*. Mashable. Retrieved 09/17/2023 from https://mashable.com/archive/fear-of-missing-out#83r7t4gHriq4

German Neubaum, Miriam Metzger, Nicole Krämer, & Elias Kyewski. (2023). How Subjective Norms Relate to Personal Privacy Regulation in Social Media: A Cross-National Approach. *Social Media + Society*, *9*(3), 20563051231182365. https://doi.org/10.1177/20563051231182365

Kazuo Okamura, & Seiji Yamada. (2020). Adaptive trust calibration for human-AI collaboration. *PloS one*, *15*(2), e0229132. https://doi.org/10.1371/journal.pone.0229132

Angela Paradise, & Meghan Sullivan. (2012). (In)visible threats? The third-person effect in perceptions of the influence of Facebook. *Cyberpsychology, Behavior, and Social Networking*, *15*(1), 55-60. https://doi.org/10.1089/cyber.2011.0054

Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, & Casey Fiesler. (2016, 2016). Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. Proceedings of the 19th International Conference on Supporting Group Work,

Maša Popovac, & Lee Hadlington. (2020). Exploring the role of egocentrism and fear of missing out on online risk behaviours among adolescents in South Africa. *International Journal of Adolescence and Youth*, *25*(1), 276-291. https://doi.org/10.1080/02673843.2019.1617171

Andrew K. Przybylski, Kou Murayama, Cody R. DeHaan, & Valerie Gladwell. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in Human Behavior*, *29*(4), 1841-1848. https://doi.org/10.1016/j.chb.2013.02.014

Vartika Pundir, Elangbam Binodini Devi, & Vishnu Nath. (2021). Arresting fake news sharing on social media: a theory of planned behavior approach. *Management Research Review*, *44*(8), 1108-1138. https://doi.org/10.1108/MRR-05-2020-0286

Felix Reer, Wai Yen Tang, & Thorsten Quandt. (2019). Psychosocial well-being and social media engagement: The mediating roles of social comparison orientation and fear of missing out. *New Media & Society*, *21*(7), 1486-1505. https://doi.org/10.1177/1461444818823719

Sarah T Roberts. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Hagit Sasson, & Gustavo Mesch. (2014). Parental mediation, peer norms and risky online behavior among adolescents. *Computers in Human Behavior*, *33*, 32-38. https://doi.org/https://doi.org/10.1016/j.chb.2013.12.025

Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, & Jed R. Brubaker. (2021). A Framework of Severity for Harmful Content Online. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2), Article 368. https://doi.org/10.1145/3479512

Sarita Schoenebeck, Oliver L. Haimson, & Lisa Nakamura. (2021). Drawing from justice theories to support targets of online harassment. *New Media & Society*. https://doi.org/10.1177/1461444820913122

Kelsea Schulenberg, Lingyuan Li, Guo Freeman, Samaneh Zamanifard, & Nathan J. McNeese. (2023). *Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality* Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany. https://doi.org/10.1145/3544548.3581090

Isabella Seeber, Eva Bittner, Robert O. Briggs, Triparna de Vreede, Gert-Jan de Vreede, Aaron Elkins, Ronald Maier, Alexander B. Merz, Sarah Oeste-Reiß, Nils Randrup, Gerhard Schwabe, & Matthias Söllner. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, *57*(2), 103174. https://doi.org/https://doi.org/10.1016/j.im.2019.103174

Joseph Seering, Tony Wang, Jina Yoon, & Geoff Kaufman. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, 1461444818821316-1461444818821316.

Shaina J. Sowles, Monique McLeary, Allison Optican, Elizabeth Cahn, Melissa J. Krauss, Ellen E. Fitzsimmons-Craft, Denise E. Wilfley, & Patricia A. Cavazos-Rehg. (2018). A content analysis of an online pro-eating disorder community on Reddit. *Body Image*, *24*, 137-144. https://doi.org/https://doi.org/10.1016/j.bodyim.2018.01.001

Elisabeth Staksrud, Kjartan Ólafsson, & Sonia Livingstone. (2013). Does the use of social networking sites increase children's risk of harm? *Computers in Human Behavior*, *29*(1), 40-50. https://doi.org/https://doi.org/10.1016/j.chb.2012.05.026

Sharifa Sultana, Mitrasree Deb, Ananya Bhattacharjee, Shaid Hasan, S.M.Raihanul Alam, Trishna Chakraborty, Prianka Roy, Samira Fairuz Ahmed, Aparna Moitra, M Ashraful Amin, A.K.M. Najmul Islam, & Syed Ishtiaque Ahmed. (2021). *'Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger* Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, https://doi.org/10.1145/3411764.3445154

Sue Tait. (2008). Pornographies of Violence? Internet Spectatorship on Body Horror. *Critical Studies in Media Communication*, *25*(1), 91-111. https://doi.org/10.1080/15295030701851148

Shalini Talwar, Amandeep Dhir, Puneet Kaur, Nida Zafar, & Melfi Alrasheedy. (2019). Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services*, *51*, 72-82. https://doi.org/10.1016/j.jretconser.2019.05.026

Héloïse Théro, & Emmanuel M. Vincent. (2022). Investigating Facebook's interventions against accounts that repeatedly share misinformation. *Information Processing & Management*, *59*(2), 102804. https://doi.org/https://doi.org/10.1016/j.ipm.2021.102804

Alex Thompson, Lindsay Stringfellow, Mairi Maclean, & Amal Nazzal. (2021). Ethical considerations and challenges for using digital ethnography to research vulnerable

populations. *Journal of Business Research*, *124*, 676-683. https://doi.org/https://doi.org/10.1016/j.jbusres.2020.02.025

Ufuk Tugtekin, Esra Barut Tugtekin, Adile Aşkım Kurt, & Kadir Demir. (2020). Associations Between Fear of Missing Out, Problematic Smartphone Use, and Social Networking Services Fatigue Among Young Adults. *Social Media + Society*, *6*(4), 2056305120963760. https://doi.org/10.1177/2056305120963760

Fiona Westin, & Sonia Chiasson. (2021). "It's So Difficult to Sever that Connection": The Role of FoMO in Users' Reluctant Privacy Behaviours. CHI '21: CHI Conference on Human Factors in Computing Systems,

Sijia Xiao, Shagun Jhaver, & Niloufar Salehi. (2023). Addressing interpersonal harm in online gaming communities: The opportunities and challenges for a restorative justice approach. *ACM Trans. Comput.-Hum. Interact.*

7. Appendix

Survey Appendix:

Start of Block: Block: Facebook Use

Have you used the social media site Facebook over the past year?

- ○ Yes (1)
- ○ No (2)

End of Block: Block: Facebook Use

Start of Block: Block: FoMo

Below is a collection of statements about your everyday experience. Using the scale provided, please indicate how true each statement is of your general experiences. Please answer according to what really reflects your experiences rather than what you think your experiences should be. Please treat each item separately from every other item

I fear others have more rewarding experiences than me.

- ○ Not at all true of me (1)
- ○ Slightly true of me (2)
- ○ Moderately true of me (3)
- ○ Very true of me (4)
- ○ Extremely true of me (5)

I fear my friends have more rewarding experiences than me.

- ○ Not at all true of me (1)
- ○ Slightly true of me (2)
- ○ Moderately true of me (3)
- ○ Very true of me (4)
- ○ Extremely true of me (5)

I get worried when I find out my friends are having fun without me.

- ○ Not at all true of me (1)
- ○ Slightly true of me (2)
- ○ Moderately true of me (3)
- ○ Very true of me (4)
- ○ Extremely true of me (5)

I get anxious when I don’t know what my friends are up to.

- ○ Not at all true of me (1)
- ○ Slightly true of me (2)
- ○ Moderately true of me (3)
- ○ Very true of me (4)
- ○ Extremely true of me (5)

It is important that I understand my friends' ‘‘in jokes’’.

- ○ Not at all true of me (1)
- ○ Slightly true of me (2)
- ○ Moderately true of me (3)
- ○ Very true of me (4)
- ○ Extremely true of me (5)

Sometimes, I wonder if I spend too much time keeping up with what is going on.

- ○ Not at all true of me (1)
- ○ Slightly true of me (2)
- ○ Moderately true of me (3)
- ○ Very true of me (4)
- ○ Extremely true of me (5)

It bothers me when I miss an opportunity to meet up with friends.

- Not at all true of me (1)
- Slightly true of me (2)
- Moderately true of me (3)
- Very true of me (4)
- Extremely true of me (5)

When I have a good time it is important for me to share the details online (e.g., updating status).

- Not at all true of me (1)
- Slightly true of me (2)
- Moderately true of me (3)
- Very true of me (4)
- Extremely true of me (5)

When I miss out on a planned get-together, it bothers me.

- ○ Not at all true of me (1)
- ○ Slightly true of me (2)
- ○ Moderately true of me (3)
- ○ Very true of me (4)
- ○ Extremely true of me (5)

When I go on vacation, I continue to keep tabs on what my friends are doing.

- ○ Not at all true of me (1)
- ○ Slightly true of me (2)
- ○ Moderately true of me (3)
- ○ Very true of me (4)
- ○ Extremely true of me (5)

End of Block: Block: FoMo

Start of Block: Block: Social Media Addiction

Please answer the following questions with regard to your Facebook use over the past year.

Have you spent a lot of time thinking about Facebook or planned use of Facebook?

- ○ Very rarely (1)
- ○ Rarely (2)
- ○ Sometimes (3)
- ○ Often (4)
- ○ Very often (5)

Have you felt an urge to use Facebook more and more?

- ○ Very rarely (1)
- ○ Rarely (2)
- ○ Sometimes (3)
- ○ Often (4)
- ○ Very often (5)

Have you used Facebook in order to forget about personal problems?

- ○ Very rarely (1)
- ○ Rarely (2)
- ○ Sometimes (3)
- ○ Often (4)
- ○ Very often (5)

Have you tried to cut down on the use of Facebook without success?

- ○ Very rarely (1)
- ○ Rarely (2)
- ○ Sometimes (3)
- ○ Often (4)
- ○ Very often (5)

Have you become restless or troubled if you have been prohibited from using Facebook?

- ○ Very rarely (1)
- ○ Rarely (2)
- ○ Sometimes (3)
- ○ Often (4)
- ○ Very often (5)

Have you used Facebook so much that it has had a negative impact on your job/studies?

- ○ Very rarely (1)
- ○ Rarely (2)
- ○ Sometimes (3)
- ○ Often (4)
- ○ Very often (5)

End of Block: Block: Social Media Addiction

Start of Block: Block: Subjective Norms

Please rate the following statements regarding your preferences of Facebook use

Suppose Facebook offers a moderation setting that limits the number of upsetting or offensive posts appearing on my news feed. Most people who are important to me would expect me to turn on this setting.

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)

Most people who are important to me would expect me to mute a Facebook friend's account if it frequently posts upsetting or offensive content. Note: Muting an account hides its posts from your news feed.

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)

Suppose Facebook offers a moderation setting that limits the number of upsetting or offensive posts appearing on the news feed. Most people who are important to me would turn on this setting in their Facebook profile.

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)

Most people who are important to me would mute a Facebook friend's account if it frequently posted upsetting or offensive content.

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)

End of Block: Block: Subjective Norms

Start of Block: Block: Trust in the Moderation Process

Please rate the following statements regarding your views of Facebook's content moderation process. This process determines which posts will be allowed or removed on the site.

Facebook's content moderation process can be trusted to judge how upsetting or offensive each post is

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)

Facebook's content moderation process is fair and impartial in determining how upsetting or offensive each post is

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)

If Facebook's content moderation process determines that a post is upsetting or offensive, I am likely to find it upsetting or offensive.

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)

I feel confident that if I mute an account on Facebook, its posts will no longer appear on my news feed.

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)

End of Block: Block: Trust in the Moderation Process

Start of Block: Block: Use of Personal Moderation Tools

Please answer the following questions regarding your preferences for using Facebook's moderation settings.

Suppose that Facebook provides a setting that lets you control the volume of posts that you may find upsetting or offensive on your news feed. How would you configure this setting?

- ○ **Allow:** You may see more posts that could be upsetting or offensive. (1)
- ○ **Limit:** You may see some posts that could be upsetting or offensive. (2)
- ○ **Limit even more:** You may see fewer posts that could be upsetting or offensive. (3)

Suppose that you encounter a Facebook friend frequently posting upsetting or offensive content that appears on your news feed. How likely are you to mute this friend's account? Note: Muting an account hides their posts from your news feed.

- ○ Extremely unlikely (1)
- ○ Moderately unlikely (2)
- ○ Slightly unlikely (3)
- ○ Neither likely nor unlikely (4)
- ○ Slightly likely (5)
- ○ Moderately likely (6)
- ○ Extremely likely (7)

Table 1: *Demographic Details of Survey Participants*

	This study, US Survey Jan 2024 (%)	**American Community Survey**, US sample 2021 (%)
Age		
18-29	13.3	17.4
30-49	39.6	29.5
50-64	24.9	25.6

65+	22.2	27.3
Gender		
Male	48.3	48.6
Female	51.7	51.4
Race/Ethnicity		
White	74.8	68.3
Black	12.6	9.3
Other	12.6	22.4
Hispanic		
Yes	11.3	13.7
Education		
High school or less	27.5	33.5
Some college	34.9	33.3
College+	37.6	33.1

Table 2*: Hierarchical Multiple Regression Analyses Predicting Participants' Preferences for Setting up Account- and Content-based Personal Moderation Tools (N = 1,061).*

Independent Variable	**Support for setting stricter sensitivity (*β*)**	**Likelihood of muting offensive account (*β*)**
Model #	Model 1	Model 2
Step 1		

Age	.122***	.081**
Gender (Female)	.156***	.104***
Race (White)	-.097**	.000
Education[a]	-.006	.065*
Political affiliation[b]	.012	-.038
R^2	.062***	.044***
Step 2		
Fear of missing out (FoMO)	-.053	-.095**
Facebook addiction	-.081*	.063
Injunctive norms	.174***	.158***
Descriptive norms	.207***	.282***
Trust in Facebook moderation	.145***	.134***
R^2 change	.160***	.222***
Total R^2	.222***	.266***

*p < .05, **p < .01, ***p < .001 (t test for β, two-tailed; F test for R^2, two-tailed).

[a]0= Less than secondary education; 1= Secondary education or more.

[b]1= Strong Democrat, 7= Strong Republican.

β = Standardized beta from the full model (final beta controlling for all variables in the model).

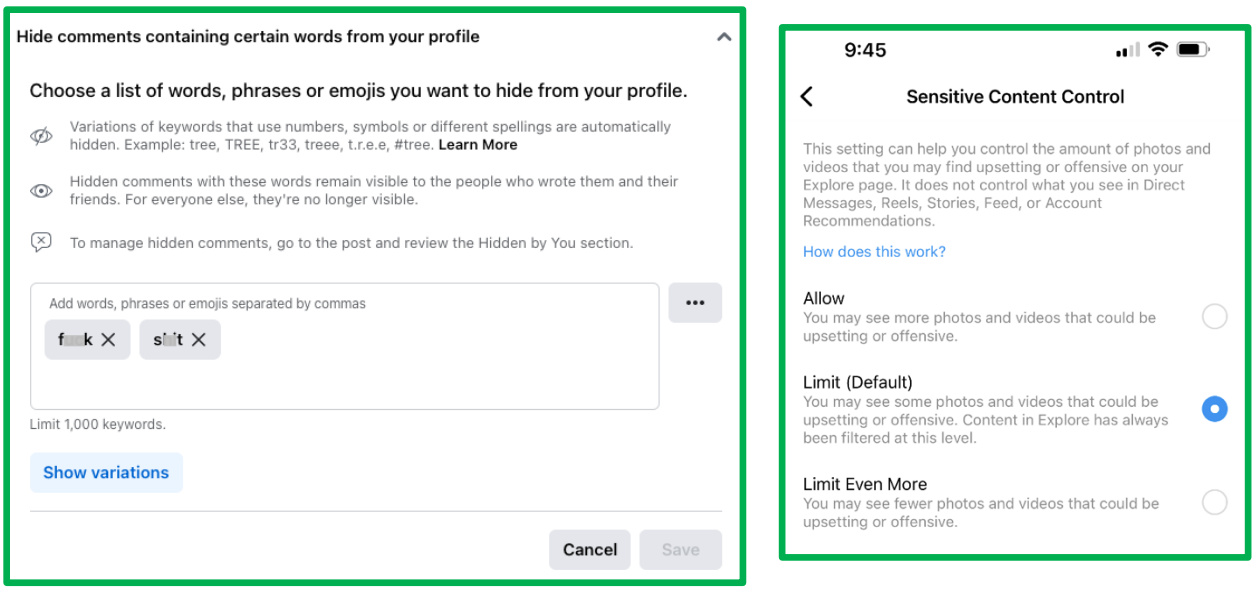

Figure 1: Examples of Personal Content-based Moderation Tools on Facebook (left) and Instagram (right).

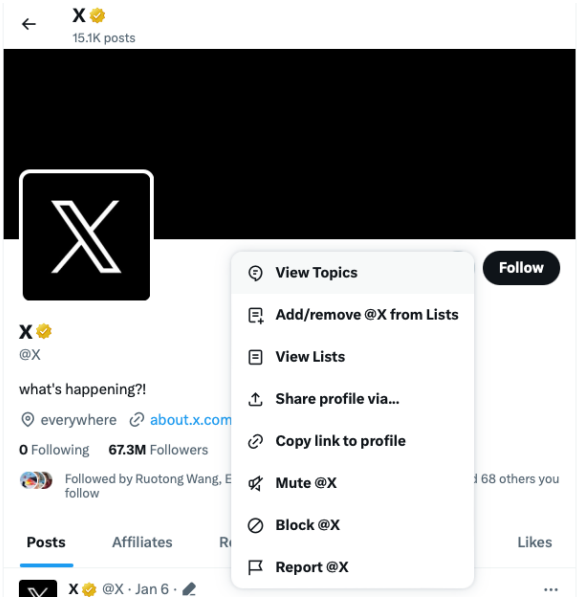

Figure 2: Examples of Account-based Moderation Actions Available on Twitter. Users can Choose to 'Mute' or 'Block' any Account.

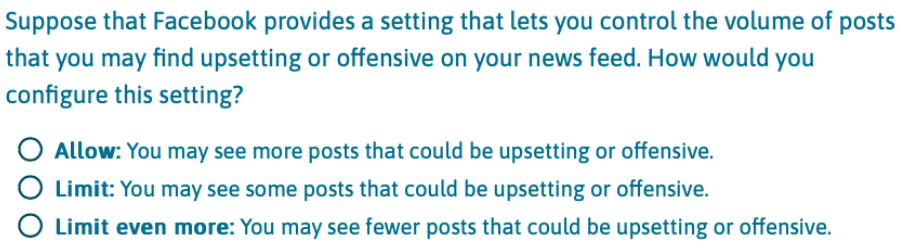

Figure 3: Survey Question Asking Participants to Configure Their Preference for a Hypothetical Content-based Moderation Tool on Facebook.

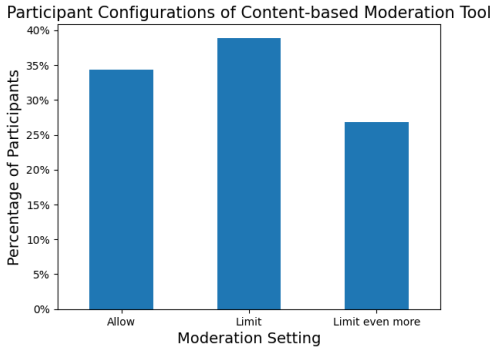

Figure 4: Participants' Configurations of Sensitivity Controls.

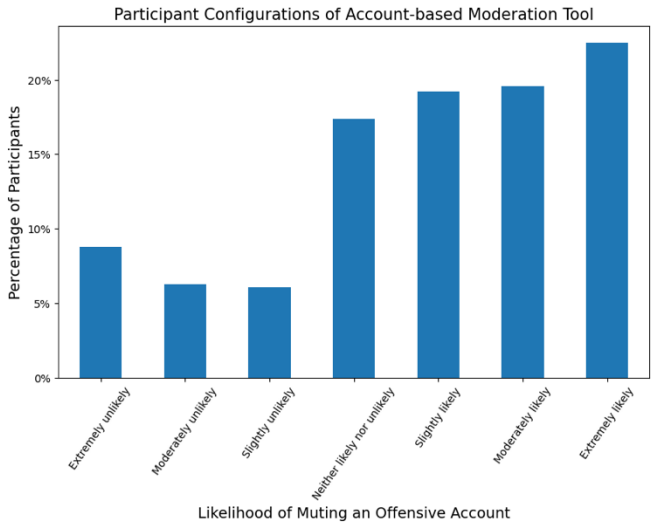

Figure 5: Participants' Configurations of the Muting Function.

Authors' contributor roles: Shagun Jhaver is the sole author of this paper and has solely contributed to all aspects of this study, including its conceptualization, data curation, formal analysis, visualization, and writing.

Disclosure of Interests: The author does not have any conflicts of interest to report.

Disclosure of AI usage: The author declares he has not used any AI services to generate or edit any part of the manuscript or data.

Biographical note:

Shagun Jhaver is a social computing scholar whose research focuses on improving content moderation on digital platforms. He is currently studying how internet platform design and moderation policies can address societal issues such as online harassment, misinformation, and the rise of hate groups. Jhaver is an assistant professor at Rutgers University's School of Communication & Information in the Library & Information Science Department.

Acknowledgments:

I thank Kiran Garimella and Koustuv Saha for their feedback on this project. This work was supported by the National Science Foundation Award 2329394.