User Preferences for Interdependent Privacy Preservation Strategies in Social Media

AARON NECAISE, University of Central Florida, USA
TANGILA ISLAM TANNI, University of Central Florida, USA
ANEKA WILLIAMS, University of Central Florida, USA
YAN SOLIHIN, University of Central Florida, USA
APU KAPADIA, Indiana University Bloomington, USA
MARY JEAN AMON, University of Central Florida, USA

Interdependent privacy (IDP) violations occur when users share personal information about others without permission, resulting in potential embarrassment, reputation loss, or harassment. There are several strategies that can be applied to protect IDP, but little is known regarding how social media users perceive IDP threats or how they prefer to respond to them. We utilized a mixed-method approach with a replication study to examine user beliefs about various government-, platform-, and user-level strategies for managing IDP violations. Participants reported that IDP represented a 'serious' online threat, and identified themselves as primarily responsible for responding to violations. IDP strategies that felt more familiar and provided greater perceived control over violations (e.g., flagging, blocking, unfriending) were rated as more effective than platform or government driven interventions. Furthermore, we found users were more willing to share on social media if they perceived their interactions as protected. Findings are discussed in relation to control paradox theory.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy**; **Social aspects of security and privacy**; • **Human-centered computing → User studies**; **User studies**; • **Social and professional topics → User characteristics**; **User characteristics**.

Additional Key Words and Phrases: interdependent privacy, content moderation, social media, usable security and privacy

ACM Reference Format:
Aaron Necaise, Tangila Islam Tanni, Aneka Williams, Yan Solihin, Apu Kapadia, and Mary Jean Amon. 2023. User Preferences for Interdependent Privacy Preservation Strategies in Social Media. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 271 (October 2023), 30 pages. https://doi.org/10.1145/3610062

1 INTRODUCTION

With the immense growth of social media and online photo sharing, the protection of private information has become critically important for the safety of internet users. Although the discourse surrounding online privacy often focuses on self-disclosure [56] and the mismanagement of data by social media platforms [79], users themselves can violate the privacy of others by re-sharing content without permission or consideration for the original context [62]. As such, personal privacy on

Authors' addresses: Aaron Necaise, aaron.necaise@ucf.edu, University of Central Florida, Orlando, Florida, USA; Tangila Islam Tanni, tanni@knights.ucf.edu, University of Central Florida, Orlando, Florida, USA; Aneka Williams, aneka_williams@knights.ucf.edu, University of Central Florida, Orlando, Florida, USA; Yan Solihin, yan.solihin@ucf.edu, University of Central Florida, Orlando, Florida, USA; Apu Kapadia, kapadia@indiana.edu, Indiana University Bloomington, Bloomington, Indiana, USA; Mary Jean Amon, mjamon@ucf.edu, University of Central Florida, Orlando, Florida, USA.

2573-0142/2023/10-ART271 $15.00
https://doi.org/10.1145/3610062

social media is dependent on *interpersonal* sharing decisions, an idea referred to as ‘interdependent privacy’ [68]. Multimedia interdependent privacy (IDP) violations pose a major risk to all users, as millions of photos and videos are circulated online every day [87]. Victims may be entirely unaware that their personal photos have been shared or re-shared [65], often with the intention of using the photo for entertainment or even shaming the subject’s physical appearance [74, 103]. Notably, one of the most common sources of IDP violations are well-intended posts made by family and friends without the permission of the original user [14]. Thus, IDP violations can be difficult to prevent and require a level of coordination between users to avoid harmful information-sharing practices.

Considering how rapidly information spreads on social media, multimedia-based IDP violations can lead to severe consequences, including loss of employment [136], emotional distress [29], and harassment [136]. Photos that are re-shared are often re-captioned so that the original content of the photo is misrepresented and misinformation is spread about the subject of the photo [45]. Additionally, information shared outside of its original source can be difficult to remove, meaning the impact of online privacy violations is long-lasting and can occur many years after the initial incident. Users require a means of protecting their information from non-consensual re-sharing by other users, but, thus far, social media companies have demonstrated a limited ability to protect users from the consequences of IDP violations [121, 124].

Social media users have several options available to them after an IDP violation has occurred, one of which includes relying on social media companies for prompt and efficient content moderation. Content moderation is the systematic process of reviewing user-generated content (UGC) posted on social media, blogs, and other online forums to determine whether it is appropriate for the site, region, or legal jurisdiction in question [101, 119]. For example, a common approach involves providing users with a button to report inappropriate content so that the platform can review it for removal at a later date [40]. Since restrictions regarding what UGC is permitted reflect that platform’s brand, its tolerance for risk, and the type of user involvement it desires to attract, approaches to moderation differ from platform to platform [119]. Furthermore, content moderation can be done by volunteers, professionals, or third-party companies [119], leading to inconsistencies that can contribute to consumer dissatisfaction [132].

In addition to platform-controlled content moderation, there are several other strategies for IDP protection applied at the user- and government-levels. At the user-level, people can respond to IDP violations by leveraging social pressure, publicly calling-out others who share inappropriate content, self-censoring, or using platform safety tools [93, 117]. These types of strategies provide users with more input over their privacy, but their effectiveness in the context of IDP is unclear. For example, even an individual who abstains from social media can have their photos shared by family members without consent [14]. At a higher level of influence, governments play a critical role in IDP by determining online privacy laws and offering legal protections to victims [83]. Thus, there are a variety of strategies for managing IDP violations, which we refer to collectively as *IDP strategies* from hereon. We take the perspective that IDP is a complex privacy issue that requires increased collaboration between users, governments, and social media technologies to appropriately identify and remove harmful content. This is consistent with the view that "Everyone has a role to play in safeguarding the future of the web," including citizens, corporations, and government stakeholders [49]. Despite the number of IDP strategies available, a major challenge remains to understand and respond to the nuanced social practices surrounding information sharing on social media that lead to complex multiparty privacy issues.

Appropriately addressing IDP concerns requires an understanding of what types of practices users themselves find acceptable. Examining user beliefs about IDP can point to promising new research directions, as users are the ones experiencing and perpetuating violations. Understanding user preferences is also necessary for avoiding false positives (e.g., moderating socially acceptable content) and false negatives (e.g., failing to moderate privacy violating content). Finally, given that there is a large diversity of privacy attitudes, some IDP strategies may be perceived more positively by different types of users. For example, some users, including those from marginalized and vulnerable populations, may have a higher sensitivity to privacy, such that a "one size fits all" approach to moderation neglects their best interests [43]. Others may be prone to distrusting IDP strategies depending on their personal beliefs, such as perceiving automated detection systems or government regulations as biased [132]. Despite the complexities of IDP, research has not yet thoroughly examined how users perceive these more nuanced privacy issues or the actions taken to remedy offenses. Thus, our work addressed four primary research questions (RQs):

RQ1: *How do users perceive the threat of multimedia IDP violations on social media?*
RQ2: *Who do users consider primarily responsible for managing IDP violations on social media?*
RQ3: *Which IDP strategies are perceived as the most effective response to IDP violations?*
RQ4: *How do user demographics and social media usage relate to IDP beliefs?*

To answer these research questions, we conducted an online survey using a Qualtrics participant panel. Participants from the United States read a description of IDP and rated the level of effectiveness and familiarity they had with a variety of IDP strategies applied at the government-, platform-, and user-level. We used a mixed-method approach, with a replication study, collecting real-world (i.e., gathered from real social media accounts), self-report, and qualitative data to measure user traits, social media usage, privacy beliefs, and sharing tendencies. The qualitative results complement the quantitative component by providing users with a voice in noting the perceived harms of IDP violations and their opinions on IDP strategies. Unlike previous research that focused on privacy preferences more broadly, we narrowed the scope of our work to focus on *interdependent* privacy on social media, which involves complex interactions between many interconnected users. Moreover, we compared several common IDP strategies, including audience management strategies, moderation techniques, government regulations, and interpersonal methods, to determine which approaches were perceived as the most effective by users themselves. Across studies, we observe a consistent preference for IDP strategies that involve a component of user control, with users qualitatively describing their desire to provide more input on privacy issues and moderation decisions. These findings support the need for greater collaboration between social media platforms and their users as a means of improving trust and privacy. The implications of these findings are discussed in the context of collaborative moderation and privacy systems, which may, in the future, enhance user privacy through increased participation.

2 BACKGROUND

2.1 Interdependent privacy

People find happiness through connection with others, and communication is the key to this social interaction [71]. People communicate by sharing information in both offline and online settings [82, 126]. The effectiveness of this communication depends on shared knowledge; the more knowledge people share during the conversation, the more effective the communication [147]. It follows that communication often includes sharing information about one's own personal details and what is known about others [71]. In the context of social media, which contain a permanent record of information [75], privacy concerns arise from this sharing. Biczok and Chai introduced the term "interdependent privacy" by describing the fact that individual privacy depends not only on their own sharing decisions, but also the decisions of others [15].

The interdependent aspect of privacy means that an individual who has information about another person can compromise their privacy without even realizing it [71]. Although every individual knows something about others [111], protecting the privacy of peers in the real world is not as complicated as we experience it in the virtual world. Privacy is like a social contract [90, 92], where respecting peer privacy is about “understanding the implicit social norms about what, why, and to whom information is shared within a specific relationship” [89]. However, maintaining implicit social contracts with others becomes difficult within large-scale networks where sharing is ubiquitous. Members of social media communities can develop their own privacy norms, which is difficult for outsiders to recognize and understand [91, 139]. In online settings, these privacy norms are reshaped as informal contracts that can be respected or violated [90]. For example, Alice can take a funny photo of Bob and share it on a social media platform without Bob’s consent. By tagging Bob in his image, both Alice and Bob’s friend have access to this photo in the default settings [15]. Therefore, people may unknowingly reveal information to each other, commercial entities, and the government. In this interconnected digital world, social media has made information sharing more accessible and increased the scope at which IDP violations can occur.

For people in online settings, privacy norms and expectations are more like "rules of conduct" that are highly context dependent [88, 91, 105]. If people perceive lower risks through an understanding of “rules of transaction conduct” [51], they are more willing to transact with greater frequency [51, 90]. Therefore, meeting or violating privacy expectations can be viewed as an antecedent to trust [90]. Previous research indicates that respecting informal contracts is positively associated with trust, while violating privacy expectations can negatively impact people’s trust in online sites [112]. Researchers have stated that privacy violations can cause physical and psychological damage [3, 66]. Reflecting on the story of “Dog Poop Girl”, who refused to clean up her dog’s feces on an underground train with her image taken by a fellow train passenger, sharing her photo caused her to drop out of her university due to feelings of humiliation. These types of incidents are becoming increasingly common and easy for almost everyone to participate in [24].

Technological advancements have made the collection and use of personal data invisible. For example, smartphone photography makes it possible for people to have their images captured at virtually any time in public and even sometimes in the privacy of their own homes. Thus, even self-censorship is not an effective method for preventing IDP violations due to dependence on other people. Very few people know what information other people, including commercial organizations, firms, and governments, have about them, how they use it, and its potential consequences [1]. Moreover, individuals may exhibit drastically different privacy preferences, ranging from extreme concern to apathy toward privacy, depending on the context [1, 48]. Privacy management is a continuous process that is learned over time and is influenced by a variety of motivational, cultural, and other individual factors [1], ranging from country laws, socioeconomic status, and online experience [12, 35, 118]. For example, a study on political disclosure on social media by Hispanic and non-Hispanic White populations in the United States found greater self-disclosure among Hispanic people and a higher likelihood of unfriending people with opposite political beliefs [109]. Furthermore, age and social media usage patterns also influence privacy concerns. Young people who have grown up with the Internet tend to spend more time on social media platforms and use techniques to obscure their personal details [80]. Taken together, research suggests that users exhibit diverse attitudes toward social media sharing, highlighting the need for diverse IDP strategies.

2.2 Strategies for managing IDP violations

Content moderation is one of several strategies used to protect IDP, in addition to limiting the spread of hate speech, nudity, violence, copyright violations, and harassment. Social media platforms are expected to take quick action against particularly harmful IDP violations, including attempts at doxxing and revenge pornography [53]. However, social media companies are highly variable in how they moderate, which can make it difficult to control the spread of violations when content is re-shared across platforms. Companies rely to different degrees on automatic moderation to remove offensive content [54], which is a broad term for non-human review methods (e.g., keyword filtering, AI moderation). Facebook takes a highly centralized approach to moderation, claiming to have automatic moderation tools that detect violations faster than users or human reviewers [95]. Although they work more efficiently than human moderators, these algorithms are unable to understand context as well as humans [44]. In comparison, Reddit takes a more decentralized approach to moderation by allowing users to create their own self-policed communities with individualized standards of conduct [69]. This decentralization means that Reddit relies more heavily on community volunteers to enforce platform-level policies [27]. In addition to variations in how content is moderated, standards regarding what is ’acceptable’ are established at the intersection of user interests, financial incentives, and compliance with local governments [61, 119, 120]. Thus, the forces influencing UGC (and IDP by extension) extend beyond the control of a single social media company.

At the government-level, there are various policies concerning privacy and hate speech, and the types of content that social media companies are required to moderate depend on local laws [32, 144]. As an example, France has stricter standards regarding online privacy than the United States, including legislation on the sharing of images of children by parents [64]. France has adopted laws that require children to consent to have their image shared in monetized online videos (e.g., Youtube videos generating ad revenue) [17], and the French Supervisory Authority has published specific recommendations for protecting the privacy of minors on the Internet [42]. There has been a growing demand in the United States for government oversight of social media [26], however, specific laws regarding online privacy vary [108]. Even the legality of severe IDP violations, such as revenge pornography, differ by state and contribute to difficulties in punishing offenders [123]. Although government regulation of social media is a contentious issue in the United States [59], it is clear that the government plays a role in protecting online privacy.

In addition to the efforts of governments and social media companies, there are several user-driven IDP strategies that can help prevent the spread of IDP violations. One example would be public call-outs, which refers to instances in which users leverage social dynamics to call-out those who go against societal norms. Public call-outs effectively shame people for inappropriate behavior, and, in extreme instances, can de-platform offenders by pressuring them to close their accounts [104, 122]. In the context of IDP, a public call-out could involve commenting with disapproval on a family member’s post that contains an image of one’s self re-shared without consent. Alternatively, users could take advantage of user-interface (UI) tools to manage their audience and privacy risks. Common UI settings allow users to block or unfollow others as a less public form of shaming and to prevent future violations from these individuals [50]. Users can also alert the platform of a potential community-standard violation by reporting (or flagging) a post.

It is important to note that many of the IDP strategies described above tend to overlap and are not necessarily mutually exclusive. Posts flagged by users as inappropriate may be reviewed by a combination of human moderators and automatic systems [6]. Social media UI tools can be co-opted by users to apply social leverage, including through mass reporting with the goal of ’gaming’ moderation systems [37]. The UI can even prompt users with safety reminders before a privacy violating photo is re-shared (e.g., asking them if they are sure they want to share sensitive information), and this subtle ’nudging’ [137] can slow the spread of secondary IDP violations (i.e., when they are not the source of the violation). Despite the practical overlap, understanding how users perceive these different strategies will provide insight about the features that users perceive as effective or trustworthy.

2.3 User beliefs about IDP

The current work takes the perspective that users are not homogeneous, and the strategies applied to limit IDP violations must take into account differences in user backgrounds and privacy preferences in order to be most effective. Users undoubtedly differ in their opinions about which IDP strategies are most effective, equitable, or trustworthy. However, little research has evaluated user privacy beliefs in general, let alone their beliefs about more complex IDP issues. A recent paper by Amon et al. [5] explored user beliefs about IDP by asking participants to rate a series of memes depicting possible IDP violations according to the degree to which they perceived them as being 'too private to share' on social media. The researchers found that higher education and increased experience with photo-sharing was associated with increased sensitivity to IDP issues [5]. Additionally, when participants perceived memes as funny or entertaining, that entertainment value decreased their sensitivity to potential privacy concerns [5]. This work suggests that IDP preferences are closely tied to a user's experiences on social media as well as their motivations for sharing images online. Related qualitative research by Hargittai and Marwick [59] used focus group interviews to assess participants' perceptions of networked privacy on social media, reporting that participants were generally aware and educated about the IDP risks associated with self-disclosing to social media. However, participants could not reach a consensus on who should be responsible for protecting IDP (i.e., companies, individuals, or governments) as they viewed violations as an inevitability of self-disclosure [59].

The negative public perception of social media companies [72] may also impact user beliefs about how privacy should be managed on social media. Social media companies have been criticized for their lack of transparency [11, 41, 84], contributing to user confusion about disciplinary actions [120] and the perception that moderation unfairly limits free speech [101]. Transparency involves providing those who violate community standards with clear information about the actions that were taken against their accounts. However, many social media users report not being notified if their content was removed and not receiving specific information on why a post was flagged in the first place [132]. Suzor et al. [132] argue this lack of communication leads users to adopt conspiratorial beliefs about the intentions of social media platforms, such as the belief that specific actions taken against them are politically motivated. Thus, transparency appears to be a barrier to user trust and acceptance of the actions of social media companies [22]. The increased reliance on opaque automatic moderation methods, such as those relying on deep-learning algorithms, may instead frustrate users who do not understand the mechanisms driving the platform's decision making.

An important theoretical consideration is the influence of the 'control paradox' [19] on privacy attitudes and sharing behaviors. Previous research highlights how perceiving greater amounts of control over personal information contributes to a diminished sense of risk and subsequent increases in self-disclosure [58, 113]. In other words, when users think they have more control over personal information, they are more confident in the outcomes of their sharing decisions. Despite believing that they can effectively manage their own data, users are often unable to exert as much control as they perceive, and this overconfidence leads to greater amounts of risk-taking and lower security (i.e., the control paradox [19]). IDP strategies vary in terms of the amount of control afforded to users, and strategies that do not incorporate user input are often viewed as less transparent [132]. It is likely that perceived control would be an important factor contributing to user perceptions about the various IDP strategies available to them.

2.4 Study overview

The current study explores user beliefs about IDP and the strategies employed in response to violations. We surveyed 204 social media users from the United States on Qualtrics who were highly active on Twitter to evaluate their beliefs about government-, platform-, and user-level IDP strategies. Users rated the effectiveness, familiarity, and trustworthiness of automatic moderation, UI reporting tools, government regulation, safety prompts, and social mechanics (i.e., call-outs, blocking, or unfriending) as it relates specifically to IDP. Although these strategies often overlap in practice, we include them separately to investigate *user perceptions* of their effectiveness. That is, a user may perceive manual reporting tools as more effective than automatic moderation despite these tools being utilized in parallel. Moreover, we do not include self-censorship (a common privacy strategy) because IDP violations occur in response to the sharing decisions of *other* users, such that self-censorship is not a method for responding to existing privacy concerns. In addition to analyzing participant survey ratings, we also used thematic qualitative analysis to examine user responses to open-ended questions in which they described how they perceived the threat of IDP violations on social media (RQ1) and the IDP strategies they believed to be the most effective (RQ3). Finally, we collected a mixture of real-world, self-report, and task-based data regarding user demographics and social media usage.

Data quality and generalizability are common concerns about online research using convenience sampling [33, 46, 55, 77, 94]. To address these concerns, we present the results of a replication study in which we recruited participants through an alternative recruitment source and successfully reproduced our primary findings. Recent scientific discourse has highlighted a substantial lack of reproducibility in multiple fields of research [8, 34, 36], generating alarm about the validity of many previously published articles. Thus, replication is a critically important process necessary for ensuring scientific integrity and improving confidence in research findings [8, 23, 36, 96]. The current study included a mixture of real-world social media behavior, collected from users with a verified amount of social media activity, and reproduced across multiple recruitment sources.

Our first research question focused on determining whether users perceived IDP risks as a substantial threat to their personal security on social media. We hypothesized that users would report being only moderately concerned with IDP violations given prior research highlighting relatively permissive sharing attitudes [4]. Next, we were interested in identifying *who* users considered to be primarily responsible for responding to IDP violations and which IDP strategies were perceived as the most effective. Previous research on the control paradox [19] demonstrates that having a greater sense of control over personal information can contribute to an inflated sense of security [58, 113] and subsequent self-disclosures [19]. Based on this theory, we predicted participants would perceive strategies carried out at the user-level (i.e., public shaming and blocking), as more effective than those carried out by governments or social media corporations. For our final research question, we wanted to determine how user demographics and social media usage were related to IDP beliefs. We predicted users who more frequently ’tweeted’ and shared photos would perceive the IDP strategies as more effective overall, as these users are more willing to self-disclose. In other words, we believed that users who were more comfortable sharing personal content would have more confidence in the ability to protect private content on social media.

3 METHOD

3.1 Participants

This study was approved by an Institutional Review Board located in the southeast United States. Participants were recruited and compensated via Qualtrics’ online participant panel. To be eligible, they had to be living in the United States, fluent in English, at least 18 years old, and have normal or corrected-to-normal vision. The decision to exclude participants based on uncorrected vision was made to ensure all participants were able to adequately view the image-based memes presented in the image-rating task. In addition, because we were interested in assessing IDP beliefs in the context of real-world social media activity, participants were required to have public Twitter accounts that were at least two years old with a minimum of 100 published tweets during that period. The final sample included 204 participants with an average age of 40.97 (*SD* = 14.84). A total of 113 participants identified as female (55.39%), while 90 participants identified as male (44.12%), and one participant identified as non-binary (.59%). The most common degree held by participants was a Bachelor's degree (31.86%), followed by a High school diploma or equivalent (31.37%), a Graduate-level degree (19.61%), an Associate's degree (13.73%), and participants who reported none of the above (3.43%). The majority of participants identified as White (66.18%), followed by 11.27% identifying as biracial, 10.29% as Black or African American, 7.35% as Hispanic, 3.92% as Asian, and .98% as Native American.

3.2 Questionnaires

3.2.1 Perception of IDP Questionnaire. A 38-item questionnaire focusing on user beliefs about IDP on social media was created for the purposes of this study (see Appendix A for the complete questionnaire). Specifically, we were interested in learning which IDP strategies participants were most familiar with, as well as which strategies were perceived as the most effective. The questionnaire was divided into three main sections. In the first section, participants were provided brief descriptions of various strategies used to handle IDP violations, and they rated how effective, familiar, and trustworthy they perceived each strategy on a scale of 1 (*not at all*) to 5 (*extremely high*). For example, participants responded to the following question: "Do you believe automatic moderation is an effective strategy for preventing or removing inappropriate content?" Participants also rated how often they experienced (or observed) each strategy on a scale of 1 (*never*) to 5 (*very frequently*). See Table 1 for an overview of the IDP strategies included. The second section of the questionnaire contained a set of Likert-type questions in which participants rated how serious they perceived privacy violations to be on social media, who should be held accountable for IDP violations, and indicated their level of personal experience (e.g., "Has your content on social media ever been removed by the social media platform?").

Finally, an open-ended response section allowed us to further investigate our key research questions through the lens of the participants' individual experiences with IDP. While quantitative analysis focuses on numerical data for convergent reasoning, qualitative analysis seeks more in-depth and free-form solutions to problems by focusing on how and why behavior occurs [7, 30, 31, 107]. To further investigate RQ1 regarding user perceptions of IDP threats on social media, we asked participants the open-ended question: "In your opinion, why or why not is it a serious problem when people post private or embarrassing information (or photos) about others on social media without permission?" Next, to investigate RQ3 regarding which IDP strategies were perceived as the most effective response to IDP violations, participants responded to the open-ended question: "What do you believe are the best methods for limiting the amount of private or embarrassing information posted online without permission? Please briefly explain why?" The qualitative and quantitative analyses associated with RQs 1 and 3 were performed independently, and the findings were integrated in the sections below. This is consistent with a convergent mixed-method design [57]) that has been described by Clark [30] as offering "more holistic and comprehensive conclusions" to complex research topics.

An inductive thematic analysis based on the procedure described by Clarke and Braun [20, 31] was applied to analyze participants' open-ended responses. Using NVivo [114], three coders independently reviewed the open-ended responses to familiarize themselves with the data, generate an initial set of codes, and organize those codes into themes relevant to RQs 1 and 3. The coders then worked together to compare their thematic mappings and develop a set of agreed upon themes, theme labels, and definitions. There were two additional rounds of coding in which each team member recursively reviewed the participant responses in lieu of their updated frameworks. This was an opportunity for them to re-examine how the responses were coded and refine the themes to ensure they were appropriately scoped and representative of the data. In the last step, the coders collaboratively reviewed their codebooks to converge on a final coding solution, and they selected exemplar quotes for inclusion in the paper [20, 25]. Considering that our approach to the analysis was collaborative and recursive, we did not calculate intercoder reliability [20].

Table 1. IDP strategies evaluated by the Perception of IDP Questionnaire

IDP Strategy	Description
User-interface (UI) tools	UI tools available to users to report inappropriate content
Safety prompts	Preemptive warnings about potential risks to sharing
Blocking or Unfriending	Manually ’blocking’ users who act inappropriately
Calling-out	Publicly confronting other users who share inappropriately
Automatic moderation	The use of AI and automated services to remove content
Government enforced	Moderation driven and regulated by government entities

3.2.2 Social Media Usage Questionnaire. The Social Media Usage Questionnaire was a seven-item survey developed to collect basic information about participants’ online photo sharing activity. Participants rated how often they shared or re-shared photos on social media, including the frequency with which they shared photos taken by themselves, their friends, or their family versus the frequency with which they shared photos found on the Internet on a scale from 1 (*never*) to 8 (*multiple times a day*). Participants also identified the intended audience for the photos they shared online (i.e., friends/connections, general viewers/public, or both), which social media platforms they used, and whether they more frequently shared their own photos versus the photos of others.

3.2.3 Social Media Disorder Scale (SMD). The SMD is a nine-item questionnaire used to measure maladaptive social media usage[140]. In the original survey, participants were asked whether they participated in potentially problematic social media behavior and responded with either "Yes" or "No." For example, one of the items asked participants if they "often used social media to escape from negative feelings?" The SMD was slightly modified for the current study so that participants rated each on a scale of 1 (*never*) to 5 (*always*) instead of a binary response. This change was made to maintain consistency with the other questionnaires in the survey, with higher average scores indicating more maladaptive social media attitudes and behaviors.

3.2.4 Privacy Preference Question. We assessed participants’ privacy preferences using a single question measure: “Are you a private person who keeps to yourself, or an open person who enjoys sharing with others?"[67]. Participants responded to this question using a 7-point Likert scale ranging from 1 (*very private*) to 7 (*very open*). Lower scores reflected stronger privacy preferences.

3.3 Image-rating task

To examine user sharing preferences, participants completed an image-rating task in which they viewed 68 memes collected from social media and rated the likelihood of sharing those memes on their own profiles from 1 (*extremely unlikely*) to 5 (*extremely likely*). Each meme contained at least one person and a text caption of 50 words or less for context, which is a common format for memes found on social media. Each meme contained potentially sensitive information about strangers who may or may not have been aware that their information was being shared. Thus, the memes represented one type of potential IDP violation occurring online, and higher average ratings represented a greater willingness to share privacy-violating content on social media. To help ensure participants' ratings were not influenced by extraneous factors unrelated to IDP, we excluded memes that involved sexist, racist, or bigoted themes. We also excluded images containing celebrities, because the privacy of celebrity photos may have been perceived differently from the privacy of strangers. We attempted to diversify the memes so that they depicted people of diverse ages, genders, and racial and ethnic backgrounds. Furthermore, each meme in the task depicted private information belonging to a specific category of information. For example, some of the memes displayed private medical information, drug usage, or revealed the sexual history of the photo subject. These categories were generated by evaluating a large data set of images using a thematic approach, and the final stimuli were selected from that larger data set so that each category contained five to six exemplar memes. The process used to generate these stimuli mirrored that of Amon et al. [4]. Ratings of the 68 images were averaged, with higher scores indicating a greater likelihood of sharing potentially privacy-violating memes.

3.4 Procedure

Participants from the United States were recruited via Qualtrics' online participant panel to complete our online survey. After providing informed consent, participants responded to a brief checklist to help ensure they were eligible for the study. As part of this screening process, participants provided their Twitter profile name and were instructed to send a direct message from that profile to our research Twitter account so that we could verify they owned a Twitter account with the minimum required activity for the study. This prevented participants from providing a fake profile name and helped to ensure that our sample consisted of active Twitter users. To help minimize privacy concerns related to the collection of Twitter data, participants were informed prior to consent about what data would be downloaded from Twitter and how it would be used in the study.

Next, participants completed the Perception of IDP Questionnaire and image-rating task described above in counterbalanced order. It was possible that viewing the privacy-violating memes presented in the image-rating task would influence participants' perceptions of IDP. Therefore, to control for order effects, half of the participants completed the image-rating task first and the other half completed the Perception of IDP Questionnaire first. The remaining questionnaires were presented in randomized order. Finally, participants completed demographic questions about their level of education, marriage status, age, race, typical grade in school (i.e., on a 4.0 scale), and employment status. After participants successfully completed the study, we utilized Twitter's API to download their complete history of public activity on Twitter. This data included any tweets, retweets, or replies to other users' tweets, as well as aggregate information about the total number of followers, total number of friends (people the participant is following), and likes given to others' tweets.

3.5 Replication study

Although some research has suggested that participants recruited through Qualtrics participant panels may be more representative of the general population than other survey panels [16], reliability and data quality are frequent concerns of online convenience sampling. To further support our findings, we replicated the study by recruiting participants from Amazon's Mechanical Turk (MTurk) platform. A total of 197 participants were recruited on MTurk, and these participants completed the same questionnaires and survey procedures as those recruited through Qualtrics.

However, Amazon's terms of service restricted the collection of personally identifiable data, preventing us from analyzing MTurk participants' Twitter histories. Thus, we were unable to verify MTurk participants' self-reported Twitter activity nor collect public Twitter data. Given this limitation, we focus on the results from the Qualtrics sample below. However, we discuss the similarities and differences between the two samples and have provided the full set of results on our OSF page.[1]

4 RESULTS

In what follows, we present a series of descriptive statistics and linear models to investigate user beliefs about IDP violations. For all models, a variance inflation factor (VIF) of 4 was utilized as a cutoff when evaluating fit to minimize issues of multicollinearity [2]. Additionally, the order participants completed the survey (i.e., whether they completed the image-rating task or the Perception of IDP Questionnaire first) was included as a covariate to control for order effects.

4.1 Descriptive statistics

The final sample from Qualtrics (n = 204) was highly active on Twitter, having used the service for an average of 8.85 years (SD = 3.56) and published an average of 15,564.64 (SD = 45,605.52) Tweets. Participants were typically active on several social media sites (M = 2.61, SD = 1.52), and most indicated that they shared photos multiple times per week. Furthermore, they indicated sharing photos of strangers (M = 3.92, SD = 2.04) at a rate similar to that of sharing photos of themselves, family, or friends (M = 3.72, SD = 1.80) on a scale from 1 (*never*) to 5 (*very frequently*).

In terms of their perceptions of the IDP strategies, participants expressed a relatively low to moderate level of confidence in their overall effectiveness. On a scale of 1 (*not at all*) to 5 (*extremely high*), the average perceived effectiveness of the strategies described in the Perception of IDP Questionnaire was 2.64 (SD = 0.81), while the average perceived trustworthiness was slightly lower at 2.37 (SD = 0.87). Despite having moderate familiarity with the strategies (M = 3.35, SD = 0.78), the participants did not often notice them in practice (M = 2.30, SD = 0.57). See Table 2 for correlation coefficients and reliability. Finally, participants had rarely experienced an IDP violation themselves (M = 1.89, SD = .91) or any associated consequences (M = 1.73, SD = .94) according to frequency ratings on a scale from 1 (*never*) to 5 (*very frequently*).

Table 2. Correlations between average perceived effectiveness, trustworthiness, familiarity, and frequency of first-hand experience with IDP prevention strategies

Variable	M	SD	α	1.	2.	3.
1. Effectiveness	2.64	0.81	0.83			
2. Trustworthiness	2.37	0.87	0.86	.88**		
3. Familiarity	3.35	0.78	0.78	.44**	.38**	
4. Frequency experienced	2.30	0.59	0.67	.44**	.36**	.62**

Pearson coefficients. Note. $^{*}p$ < .05; $^{**}p$<.001

4.2 How did participants perceive the threat of IDP?

4.2.1 Quantitative findings. As a baseline consideration, we were interested in understanding how participants perceived the threat of IDP on social media. Participants rated whether they believed IDP violations were a serious online threat on a scale from 1 (*not serious at all*) to 5 (*extremely serious*) and responded with an average rating of 3.82 (SD = 0.99; see Figure 1). Over

[1] https://osf.io/wema3/?view_only=97dfa7b6b3e74cec84dd87bd2417ca33

half (51.47%) of respondents believed IDP violations were a "very" or "extremely" serious threat, while substantially fewer rated them as "slightly" (15.68%) or "not at all" serious (2.45%). Similarly, participants rated how prevalent they believed IDP violations were on social media from 1 (*not at all) to 5 (extremely common)*, with an average rating of 3.50 (*SD* = 1.04). Ratings of seriousness were positively correlated with ratings of prevalence ($r(202) = .38, p < .001$). In other words, the majority of participants perceived IDP violations as a serious threat to their online safety, and this belief was positively associated with perceptions of prevalence. These findings were consistent with those from MTurk[1]. We explore this research question further in the qualitative analysis presented below.

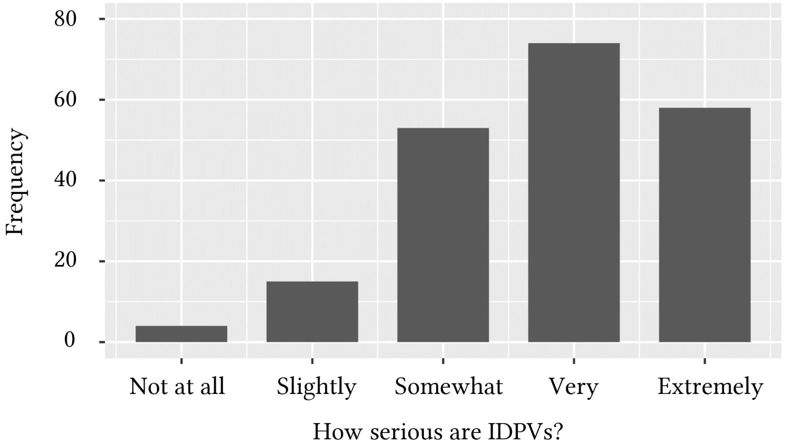

Fig. 1. Histogram demonstrating the distribution of perceived seriousness of IDP violations.

4.2.2 Qualitative findings. To compliment the descriptive analysis, participants were asked why (or why not) they thought IDP violations were a serious online issue (see Table 3). Only 5% of participants stated that IDP violations were *not* a serious issue. The majority of responses expressed concerns about the most dangerous potential outcomes of IDP violations. For example, 21% of the participants were concerned about online harassment. As described by one of these participants, "Online bullying is a major epidemic and can be fatal, so it's very serious." On the topic of harassment, participants provided specific examples of cyberbullying, blackmail, doxxing (i.e., releasing someone's real-world information to the online public), revenge pornography, and stalking. In a personal anecdote, a situation was described in which a participant was doxxed and forced to "move, change all of her contact information, change jobs, and pull [her] boys out of daycare" because of IDP-related harassment. Overall, participants appeared to understand the threat of IDP well, which should not be surprising given participants in our sample had a large amount of first-hand experience with social media (§4.1).

When providing their opinions about IDP threats, participants tended to emphasize the *types* of harm inflicted by violations, including economic, emotional, general, and social harm. Most commonly, participants believed IDP violations could lead to emotional harm (18%) due to embarrassment. There was a belief that humiliation caused by IDP violations could lead to depression, self-esteem issues, or (in extreme cases) suicide. For example, "...With social media platforms, the content you post faces a chance to reach the eyes of hundreds and even thousands of people. Having that many people laugh at and make fun, or even just be aware of your private or embarrassing information can take a serious toll on somebody's mental health." In terms of economic harm (8%), participants were more focused on how employers may view embarrassing online information and thought IDP violations could affect victims' careers and financial security. As such, one user stated, "people have been known to lose jobs over photos or easily photoshopped photos that have appeared on social media." A related concern was that IDP violations could cause social harm (11%) by damaging victims' personal relationships. It was stressed that "certain pictures and videos can become viral making the persons in said post become notorious for that one embarrassing thing. This can damage reputations with work or acquaintances or close friends and family."

Table 3. Themes identified for the open-ended question: "In your opinion, why or why not is it a serious problem when people post private or embarrassing information (or photos) about others on social media without permission?"

Theme	% of users	Description	Example
Harassment	25%	Information is weaponized and used intentionally to damage victim	"People blackmail others all the time! It's important to have control over your own image and likeness online."
Privacy Infringement	24%	Violated right to privacy	"...people should have the right to decide what information about them is public."
Emotional Harm	18%	Leads to psychological distress	"It can lead to anxiety and depression and suicidal thoughts for those involved."
Social Harm	11%	Damages reputation and personal relationships	"Some of the content can be extremely embarrassing and can ruin a persons reputation"
General Harm	9%	Described as harmful but not specified	"You could be posting something that may cause people to get hurt in a very serious way"
Economic Harm	8%	Job loss and career damage	"Because a prospective employer could come across your photo or post"
Long-lasting	8%	Damage cannot be undone for many years	"Anything that's posted without permission is made available to the whole world and can be saved. It's impossible to take it back"
Unethical	6%	Against formal or social standards	"if this behavior was engaged in offline, through the mail, it would be considered criminal..."
Not serious	5%	IDP violations not a serious problem	"this kind of content usually only endangers one person, so i do not particularly consider it a serious problem."
Situational	3%	Viewed as a serious problem is content was highly sensitive	"It's serious if it's like actually important information like addresses or personal nudes"

4.3 Whom ought to be responsible for managing IDP?

When asked whether more precautions should be taken regarding IDP violations on a scale of 1 (*no more precautions*) to 5 (*many more precautions*), participants generally believed that individuals (M = 3.99, SD = .83), social media companies (M = 3.96, SD = 0.94), and governments (M = 3.54, SD = 1.04) should all take more precautions with an overall average of 3.83 (SD = 0.78, α = 0.77). However, when asked who ought to be responsible for managing IDP violations, most users indicated that the original person who posted the private content should be responsible (39%), followed by social media companies (35%), other users (15%), and government (11%). Although exploratory in nature, these trends suggest that participants placed very low emphasis on government regulation and interpersonal methods (i.e., call-outs) for managing privacy-violating content. Replicating these findings, MTurk respondents also placed the most responsibility on the original poster (37.99%) followed by social media companies (37.15%).

4.4 Which IDP strategies were perceived as the most effective?

4.4.1 Quantitative findings. To evaluate which strategies were preferred by participants, we fit a linear mixed-effects model with participants' ratings of effectiveness as the outcome and participant ID as the random effect (see Table 4). IDP strategy was included as a categorical predictor with the UI category as the reference level. UI was selected as the reference because it was the most ubiquitous and forward-facing approach to removing content among the options provided to participants. We also entered a number of covariates in the model to control for differences in familiarity with social media privacy issues. This included the degree participants felt negatively impacted by IDP violations, how frequently they had been criticized for sharing private information on social media, how frequently they had their own content moderated, and how frequently they were the victim of an IDP violation. These four items were single-item survey questions. Finally, a Bonferonni correction was applied to the resulting p-values to control for multiple comparisons against the same baseline.

Beliefs about IDP strategy effectiveness differed significantly depending on strategy (See Figure 2). On average, participants believed blocking was more effective than UI tools (β = .33, p < .001). However, UI tools were perceived as more effective than publicly calling-out other users for their behavior (β = -.42, p < .001), automatic moderation (β = -.41, p < .001), or government regulation (β = -.52, p < .001). Participants who reported being negatively affected by IDP violations also perceived the strategies as more effective on average (β = .18, p = .02). There were no significant order effects or relationships with other covariates.

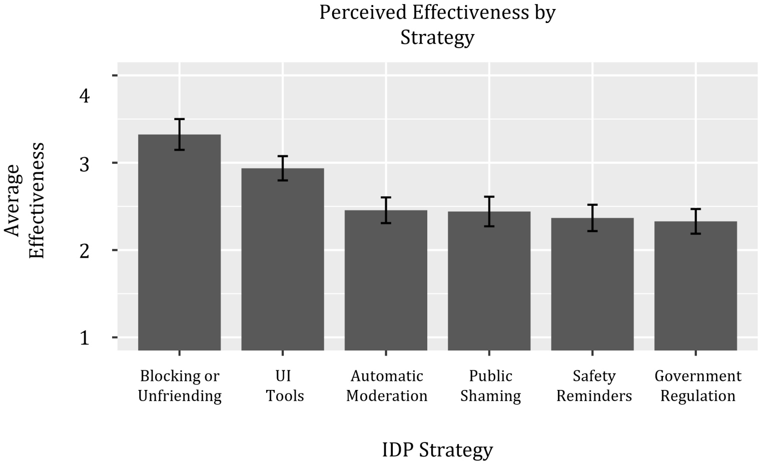

Fig. 2. Average perceived effectiveness by strategy with 95% confidence intervals

In our follow-up MTurk replication study, we replicated the primary findings of this model. As with the results from Qualtrics, participants from MTurk indicated that blocking was the most effective strategy followed by the use of UI tools, and they perceived both strategies as more effective than automatic moderation, call-outs, government regulation, and safety prompts. Furthermore, there were no significant order effects or associations with the included covariates in the MTurk sample.

Table 4. Mixed-effects model predicting user perceptions of effectiveness based on strategy and personal experience

	Perceived Effectiveness		
Predictors	*Std. Estimates*	*Std. CI*	*p*
(Intercept)	.25	.12 - .38	**<.001**
Survey order	.10	.01 - .19	.32
Strategy: UI Tools	Reference		
Strategy: Blocking	.33	.19 - .47	**<.001**
Strategy: Safety prompts	-.48	-.62 - -.34	**<.001**
Strategy: Call-outs	-.42	-.56 - -.28	**<.001**
Strategy: Automatic moderation	-.41	-.55 - -.27	**<.001**
Strategy: Government regulation	-.52	-.66 - -.38	**<.001**
Personally experienced privacy violations	-.05	-.16 - .06	1.00
Negatively affected by privacy violations	.18	.06 - .29	**.02**
Criticized for sharing	.14	.03 - .24	.11
Posted content that was moderated	-.11	-.32 - .01	.034
Random Effects			
σ^2	.71		
τ_{00}	$.48_{\text{Response Id}}$		
ICC	.40		
N	$204_{\text{Response Id}}$		
Observations	1224		
Marginal R^2 / Conditional R^2	.15 / .49		

4.4.2 Qualitative findings. Although the analysis presented in §4.4.1 compared participant ratings of several common IDP prevention strategies, the survey did not include all possible solutions. Thus, to provide additional context to the survey results and the opportunity to express preferences for IDP strategies not included in the survey, participants were asked what they believed was the best method for preventing IDP violations and *why*. Most users (44%) endorsed a type of user-driven IDP strategy (see Table 5), such as UI tools, blocking, or self-censorship. A common theme among these individuals was the belief that social media participants were better at determining offenses, e.g., "honestly, users tend to get more done when it comes to flagging that type of content. AI don't have a good handle on what's embarrassing or inappropriate quite yet, and companies will never see the tweets as fast as the users will." While UI tools were the most frequently mentioned type of user-driven IDP strategy, a common sub-theme emerged revolving around the idea of individual responsibility and self-censorship. These participants emphasized that social media participants must practice better judgment when sharing embarrassing or damaging personal information. As one person described, "don't share it [private information] with people you don't trust completely." Another said, "I don't post things on social media that I don't want shared. I assume that others do the same." The idea of self-censorship is aligned with a preference for having more direct control over social media processes.

A substantially smaller percentage of participants endorsed strategies that did not incorporate user input. Only 12% mentioned AI and even fewer cited a form of government intervention (8%). The large discrepancy between support for user-driven approaches and other strategies may be related to the distrust or political opinions voiced by several when describing their preference: "The government should not get involved because that raises the issue of freedom of expression (First Amendment rights.)..." This discrepancy may also have been related to a lack of confidence in social media companies to effectively moderate. Notably, among participants who preferred platform-driven moderation, nearly half of them expressed that social media companies were not doing enough to prevent IDP violations. Moreover, participants who wanted more platform-driven moderation were sometimes frustrated by the perceived lack of action: "We need more enforcement of reported posts by the social media platform. Countless times I have reported something only for the social media platform to say they found no issue with it. This is especially true for bullying and harassment."

Table 5. Themes identified for the open-ended question: "What do you believe are the best methods for limiting the amount of private or embarrassing information posted online without permission? Please briefly explain why?"

Theme	% of users	Description	Example
User-driven	44%	Strategies that gives user direct input over inappropriate content	"I think people should censor themselves"
Platform-driven	20%	Direct action taken by social media company	"human moderation, as sometimes automated systems make errors"
Interpersonal Methods	13%	Social methods used to influence others' content	"I think real-life social pressure is the most effective. People who are shunned in real life tend to modify their behavior."
Automatic Moderation	12%	Tools that automatically remove offensive content	"Automatic moderation would be able to be the fastest way to detect things and be more effective rather than a manual review"
Government Regulation	8%	Government policies and legal punishments for IDP violations	"Government policies because they come with real world consequences."
Education	8%	Teaching users about social media privacy and best practices	"Children should be educated in school about the potential dangers of social media and posting information about others."
No Solution	7%	Did not believe moderation was possible or helpful	"i dont think there should be a way, it's a slippery slope"

4.5 How do user attributes relate to IDP beliefs?

4.5.1 Social media usage and sharing preferences. Next, we investigated how social media usage influenced attitudes about strategy effectiveness. We collected self-reported and real-world data from Twitter to measure how often participants posted on social media, how many accounts they maintained, and their photo-sharing preferences. For self-reported image-sharing frequency, participants rated how often they shared 1) images in general, 2) photos of friends, and 3) photos of strangers, and we utilized their average rating across the three items (M = 3.18, SD = 1.72, α = 0.73). We also collected data on the likelihood of sharing privacy-violating memes during the image-rating task by averaging their ratings of the 68 memes. To assess how these social media variables were related to beliefs about IDP protection, we used a linear regression model with average effectiveness rating (i.e., averaged across all strategy types) as the outcome variable. Average scores on the SMD (M = 1.71, SD = 0.68, α =0.88), total number of tweets (log-scale), number of social media accounts, self-reported photo sharing frequency, and average meme sharing likelihood (M = 2.12 out of 5, SD = 0.81, α = 0.98) were entered as predictors. Additionally, we included whether participants preferred sharing photos of themselves (versus others) and whether participants intended on sharing photos with their friends and family (versus strangers) as categorical factors.

Participants who reported more frequently sharing photos on social media (β = .31, p < .001) and those who were more likely to share the privacy-violating memes presented in the image-rating task (β = .31, p < .001) believed IDP prevention was more effective on average. In other words, participants were more willing to share photos, even if those photos contained potentially private information about other people, as long as they were confident in the ability to effectively manage offensive material. In contrast, participants who made more Tweets on Twitter tended to have lower perceptions of effectiveness (β = -.13, p = .04). There was also a significant order effect, such that participants who completed the imaging rating task first had lower perceptions of effectiveness (β = 0.19, p = .004). The lower ratings may be explained by the fact that the imaging rating task had participants view a number of potentially offensive internet memes, reducing their confidence in the ability to prevent that offensive content from spreading. None of the other social media usage predictors were significantly related to beliefs about effectiveness.

When attempting to replicate these findings on Mturk, we were unable to include the number of tweets made by participants in our analysis because the MTurk Terms of Service did not allow the collection of Twitter data. However, we repeated the analysis presented above excluding the number of tweets, and we found identical results. Photo sharing frequency and willingness to share privacy-violating memes were both positively related to beliefs about IDP strategy effectiveness, but none of the other social media usage variables or survey order contributed significantly. Therefore, we found reliable evidence that increased photo sharing was associated with a heightened perception of effectiveness.

4.5.2 Demographic information. Finally, we used linear regression to test if participant age, education, gender, or race was associated with differences in the perceived effectiveness of the IDP strategies. We found that age was negatively associated with perceived effectiveness (β = -.35, p < .001), and participants who had earned a Graduate level degree believed the IDP strategies were more effective than those whose highest degree was from High school (β = .44, p = .03). However, none of the other education or demographic variables were significantly associated with perceptions of effectiveness. There were also no significant order effects. It is worth noting that we were unable to replicate the findings regarding age or education level, as none of the participant demographics variables were significantly related to perceived effectiveness in the MTurk sample.

5 DISCUSSION

Despite attempts to limit the spread of privacy violations, social media companies have faced major challenges when it comes to correctly identifying potential IDP violations. Adding to this challenge is the fact that users themselves differ in their opinions about what types of content should be considered ‘public’ [1, 48]. Given the wide range of individual differences expressed by users about social media privacy, it is essential that the strategies used in response to IDP violations align with user attitudes and beliefs in order to effectively address the concerns of social media users. Although substantial work has explored user privacy and sharing preferences, less work has empirically investigated user preferences regarding *IDP* preservation. The present work addresses these challenges by employing a mixed-methods and replication approach to investigate user beliefs about a variety of strategies employed to protect IDP. Uniquely, we directly compare user perceptions about the effectiveness of content moderation, audience management, government regulation, and various interpersonal methods as a means of protecting against IDP violations.

5.1 Preference for user-controlled privacy tools

Across studies, we found participants assigned themselves primary responsibility for preserving IDP on social media, despite also being the source of those violations. We speculate that dissatisfaction with social media companies and government regulations may be the basis for the preference for user-controlled IDP strategies, such as opting for audience management. This conclusion is consistent with prior research demonstrating that users are concerned about the intentions of social media companies [9, 101], perceiving them as untrustworthy [73, 130] or generally opaque [39, 70, 132]. Due to a lack of transparency, inaccuracy [98, 133], and difficulty understanding context, it follows that users would lose their trust in social media companies and perceive them as less effective at protecting IDP [54]. Similarly, social media users also report low levels of confidence in governments when it comes to privacy preservation [86]. It is not surprising that users would prefer to take more control over identifying and responding to potential IDP violations given their limited confidence.

Previous literature on the control paradox highlights how a greater sense of control over personal information can reduce the assessed risk by social media users during self-disclosure, such that users are more willing to self-disclose if they believe they have control over their information [19, 58, 113]. Our results suggest this inflated sense of security related to perceived control extends into the realm of content management, as IDP strategies involving user input were rated as the most effective. In reality, these strategies are likely not as effective on a large scale as automatic moderation systems that are capable of responding to massive amounts of content or government regulations mandating greater privacy protections. Thus, there appears to be a general need to improve consumer confidence through transparency and communication on privacy issues, including providing users with more educational resources so they are familiar with the mechanisms underlying content moderation. Given the strong desire to provide input on content management by social media users, there is also a promising opportunity for collaborative systems that users can interact with to derive a greater sense of involvement.

Initially, we predicted IDP strategies carried out at the user-level, such as blocking and publicly calling out other users, would be perceived as the most effective IDP strategies because these methods rely primarily on community input. This aligned with research examining the use of public shaming on social media [135]. However, we only found partial support for our hypothesis, as participants generally did not believe public call-outs were a particularly effective response to IDP violations. A common issue with call-outs is that social media users can get carried away, turning the targets into victims themselves. What might begin as appropriate criticism can degenerate into brutish displays of virtual tar-and-feathering [93]. It is possible that users differentiate between the shaming of large public figures versus less popular social media accounts due to these extreme responses. For example, people may perceive public shaming targeting celebrities as a form of accountability while perceiving the same behaviors as bullying when targeting less visible community members. Another explanation is that public call-outs are viewed politically due to the extensive attention given to ‘cancel culture’ in media [106], thus, opinions regarding public call-outs may depend on political orientation or age of participants [38, 100]. However, more work is needed to explore the use of public shaming on social media.

The results of our qualitative analysis agree with and expand on our quantitative findings. One notable finding was that a large number of participants believed IDP violations could be completely avoided had victims made better sharing decisions or avoided sharing sensitive content. Undoubtedly, social media requires that users assume some degree of risk by the nature of online disclosure [10], and users often do not take enough precautionary action when it comes to protecting their privacy according to personal preferences [28, 59, 134, 142]. That being said, the occurrence of

IDP violations on social media is unique in that they arise from the sharing decisions of networks of users and not because of a single user [68]. Thus, it is unclear whether the participants in our sample fully appreciated the number of circumstances in which an IDP violation could occur without the victim's knowledge [65] or in response to apparently 'safe' disclosures [14]. Moreover, even someone completely abstaining from social media (or practicing self-censorship) can experience an IDP violation when their friends re-share group photos or if they are photographed by strangers in public. Our findings highlight that social media users may need education when it comes to the complexities and connectivity of online privacy, focusing on interpersonal responsibility in addition to individual responsibility.

5.2 Perceived seriousness of interdependent privacy violations

Given the connection between perceived prevalence and perceived seriousness of IDP violations, education about the prevalence of privacy issues may help convince users that online privacy is a serious problem requiring their action. Social media users generally indicate a desire for personal privacy [85], but this preference is not reflected in their real-world behaviors [52]. This phenomenon is known as the 'privacy paradox' [52] and is theorized to be driven by a 'privacy calculus' [78] whereby users weigh out the benefits of disclosing information against the cost to privacy. In the context of IDP, research shows users are more willing to violate the privacy of others when the content they are re-sharing is perceived as humorous [62]. If educating users about the prevalence of IDP violations increases the perceived threat, users may be more reluctant to violate those standards for the sake of sharing entertaining content.

When describing why they thought IDP violations were a serious risk to online security, the majority of participants expressed their concerns depending on the types of harm that could occur after a violation. For example, they thought having embarrassing photos circulated on the internet could lead to secondary harm in the form of economic, emotional, or social damage. Previous research on privacy risks has described harm along similar dimensions of emotional, financial, reputation-based, and physical threats [110, 143, 146]. Privacy violations are an inevitable occurrence on social media, so it is important that social media users have a means of reducing the harm they experience after infringement [138, 149]. Given that participants were concerned that their personal information would be used against them, new initiatives aimed at limiting public shaming and targeted harassment [18, 102, 131] may function as a type of harm reduction by helping to reduce the secondary harm experienced by victims after an IDP violation.

5.3 User characteristics predicting IDP beliefs

When assessing participants' social media usage, we found that participants with higher photo sharing frequency believed the IDP strategies could more effectively protect them from violations. It is reasonable to assume that users who do not share as many photos may limit the type of content they share because of their privacy concerns. Social media users express a desire to preserve privacy [115], so it follows they would not want to share content perceived as more sensitive (e.g., personal photos) if they believed they could not adequately protect that content from misuse. Likewise, individuals who have more confidence in different IDP strategies may feel comfortable sharing photos. These findings suggest that improving user confidence in IDP on social media can promote participation and connectivity. There is evidence to suggest that privacy concerns can be a barrier to participating on social media [21, 116, 129], and adopting privacy settings that are aligned to user preferences can improve the user experience by increasing social connection [145]. Alternatively, these findings may reflect a level of cognitive dissonance. Cognitive dissonance refers to the discomfort that people feel when there is a conflict between their beliefs and behaviors, leading them to change their attitudes to minimize dissonance [47, 60]. Under this alternative explanation, it is possible that participants who were more open to self-disclosure were also more inclined to describe the IDP strategies as being effective as a way of downplaying the associated risks. Furthermore, considering we were not able to find replicable associations of age, education, or demographics with IDP preferences, it appears that a user's experiences on social media may be the primary driver of their beliefs.

5.4 Design implications

Even though participants prefer control over IDP strategies, their preferred strategies also require a degree of collaboration between users and social media technologies. For example, the preferred method of flagging requires a UI tool and a message sent to the social media corporation for further handling. This work suggests that tools supporting collaborative decision-making are a promising direction for content moderation developments. As an example, social media platforms can try to develop moderation methods that include users in the decision-making process as a means of increasing trust in company actions. One possibility is the development of collaborative human-AI moderation systems that involve the input of human moderators or users themselves [76]. Alternatively, previous research has highlighted various obfuscation tools that support privacy by providing users with the option to blur aspects of photos. Obfuscation tools are desirable as they provide users control over privacy considerations. However, their application remains limited [141]. Using a third-party application to edit photos may not be convenient for users due to individual technological skills, usability, and time constraints. Thus, social media platforms should further develop these tools so that users can easily edit sensitive information while sharing it on social media. Additional obfuscation features can incorporate a user awareness component, for example, automatically suggesting blurring of bystander faces before the user is allowed to post [63].

Taking into account the qualitative responses of participants regarding self-censorship, these obfuscation tools can be extended to allow individuals who violate the terms of service the opportunity to censor their own content after it has been flagged. For example, if a user accidentally posts a picture that contains an image with a bullying message, they could receive a notification that informs them of why their post was flagged and the opportunity to remedy the situation themselves. Instead of removing the entirety of their content without a clear description [132], this approach provides the user with greater agency and educates them about privacy and moderation.

Furthermore, it is possible that increasing education about IDP violations may improve user responses to potential violations. To this end, social media platforms can leverage nudges that influence the way information is presented. Previous research has indicated that various psychological nudges can influence an individual's decision-making [125]. Moreover, these decisions are highly context-dependent and influenced by the choice environment [137]. The concept of nudging arises from the knowledge that the choice environment can effect the likelihood of certain decisions and their associated behaviors [13, 97]. Nudging as a design feature allows users to control sharing decisions while also aiming to enhance privacy-oriented behaviors. Nudges are minor changes to the design environment that often aim to influence specific psychological effects to guide users toward predefined options [97]. Previous researchers have found that nudges can be effective as privacy notices to elicit awareness in users [99] but require careful consideration as they can backfire in some contexts [4]. One option is to provide visual cues for users to consider the content of their photos before posting on social media. Multi-step digital nudges can even be used to guide users. First, users can be warned to check the audience of a post, reminding users that they are potentially giving access to third-party commercial platforms. Second, nudges can offer user guidance that will introduce the available technological affordances to the users for editing/removing unwanted content from the photo. Third, users can be informed about the potential consequences of information disclosure and how it can harm users unknowingly. Although a comprehensive overview of potential UI tools is outside of the scope of this paper, the present work suggests that HCI researchers advancing user-centered and convenient UI tools [81, 127, 128, 148] to enhance privacy are well aligned with user preferences for privacy control.

6 LIMITATIONS

A notable limitation of our survey design was the use of single-item survey measures, which can be unreliable and limit the applicability of linear models. To address this concern, we provided reliability measures whenever possible and limited our use of these single items in the statistical modeling portions of the analysis. It should be noted that we were able to successfully replicate virtually all of our findings, including the more exploratory data observations, across two separate data collection platforms. Thus, we do not believe single-item reliability issues were a significant factor impacting our results, but it is important to take these potential effects into consideration. Additionally, although we attempted to include a large variety of common IDP strategies that are used in the real world, the list of included strategies was not exhaustive. An area for continued research would be to explore user perceptions of additional IDP strategies, such as interactive obfuscation systems or educational interventions.

7 CONCLUSIONS

We examined people's beliefs about the effectiveness of common strategies used on social media in response to interdependent privacy violations, where one user violates the privacy of another. Participants identified interdependent privacy violations as a serious issue, highlighting substantial concerns about numerous associated harms. Despite users being the source of interdependent privacy violations on social media, we provide mixed-methods evidence across multiple studies that social-media users assign themselves primary responsibility for interdependent privacy preservation. Users overwhelmingly prefer IDP strategies they feel familiar with and that provide them with more direct control over potential interdependent privacy violations. In addition, we identified a number of social media behaviors and user demographics associated with IDP beliefs. Those who more frequently shared photos on social media believed that the IDP strategies were more effective on average, while age was negatively associated with the perception of effectiveness. Our findings highlight the importance of content moderation transparency, education about IDP, and user involvement as a means of increasing consumer confidence in privacy on social media.

8 ACKNOWLEDGMENTS

This material is based upon work supported in part by the National Science Foundation under grant CNS-2053152. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.

[2] Paul Allison. 1999. *Multiple regression: A primer.* Pine Forge Press.

[3] Irwin Altman. 1975. The environment and social behavior: privacy, personal space, territory, and crowding. (1975).

[4] Mary Jean Amon, Rakibul Hasan, Kurt Hugenberg, Bennett I Bertenthal, and Apu Kapadia. 2020. Influencing photo sharing decisions on social media: A case of paradoxical findings. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1350–1366.

[5] Mary Jean Amon, Aaron Necaise, Nika Kartvelishvili, Aneka Williams, Yan Solihin, and Apu Kapadia. 2023. Modeling User Characteristics Associated with Interdependent Privacy Perceptions on Social Media. *ACM Transactions on Computer-Human Interaction* (2023).

[6] María-Florencia Amorelli and Isabel-María García-Sánchez. 2020. Critical mass of female directors, human capital, and stakeholder engagement by corporate social reporting. *Corporate Social Responsibility and Environmental Management* 27, 1 (2020), 204–221.

[7] Attest. 2017. Quantitative vs qualitative research—what's the difference? https://www.askattest.com/blog/articles/quantitative-vs-qualitative-research-and-how-to-use-each

[8] Monya Baker. 2016. Reproducibility crisis. *Nature* 533, 26 (2016), 353–66.

[9] Stephanie Alice Baker, Matthew Wade, and Michael James Walsh. 2020. ? covid19? The challenges of responding to misinformation during a pandemic: Content moderation and the limitations of the concept of harm. *Media International Australia* 177, 1 (2020), 103–107.

[10] Nadine Barrett-Maitland and Jenice Lynch. 2020. Social media, ethics and the privacy paradox. *Security and privacy from a legal, ethical, and technical perspective* (2020).

[11] Rebecca Bellan. 2020. Americans want transparency in content moderation decisions on social media. https://www.forbes.com/sites/rebeccabellan/2020/06/19/americans-want-transparency-in-content-moderation-decisions-on-social-media/?sh=51b4362e70ae

[12] Steven Bellman, Eric J Johnson, Stephen J Kobrin, and Gerald L Lohse. 2004. International differences in information privacy concerns: A global survey of consumers. *The Information Society* 20, 5 (2004), 313–324.

[13] Kristoffer Bergram, Valéry Bezençon, Paul Maingot, Tony Gjerlufsen, and Adrian Holzer. 2020. Digital Nudges for Privacy Awareness: From consent to informed consent?. In *ECIS*.

[14] Andrew Besmer and Heather Richter Lipford. 2010. Moving beyond untagging: photo privacy in a tagged world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1563–1572.

[15] Gergely Biczók and Pern Hui Chia. 2013. Interdependent privacy: Let me share your data. In *International conference on financial cryptography and data security*. Springer, 338–353.

[16] Taylor C. Boas, Dino P. Christenson, and David M. Glick. 2020. Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research and Methods* 8, 2 (2020), 232–250. https://doi.org/10.1017/psrm.2018.28

[17] Nicolas Boring. 2020. *France: Parliament Adopts Law to Protect Child Influencers on Social Media*. Library of Congress. https://www.loc.gov/item/global-legal-monitor/2020-10-30/france-parliament-adopts-law-to-protect-child-influencers-on-social-media/

[18] Catherine P. Bradshaw. 2013. Preventing bullying through positive behavioral interventions and supports (PBIS): A multitiered approach to prevention and integration. *Theory Into Practice* 52, 4 (2013), 288–295.

[19] Laura Brandimarte, Alessandro Acquisti, and George Loewenstein. 2013. Misplaced confidences: Privacy and the control paradox. *Social psychological and personality science* 4, 3 (2013), 340–347.

[20] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[21] Laura F. Bright, Hayoung Sally Lim, and Kelty Logan. 2021. "Should I Post or Ghost?": Examining how privacy concerns impact social media engagement in US consumers. *Psychology & Marketing* 38, 10 (2021), 1712–1722.

[22] Jens Brunk, Jana Mattern, and Dennis M Riehle. 2019. Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, Vol. 1. IEEE, 429–435.

[23] Leonard E. Burman, Robert Reed, and James Alm. 2010. A call for replication studies. *Public Finance Review* 38, 6 (2010), 787–793.

[24] John Buschman. 2019. The Future of Reputation: Gossip, Rumor, and Privacy on the Internet. *Journal of Information Ethics* 28, 1 (2019), 157–159.

[25] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & quantity* 56, 3 (2022), 1391–1412.

[26] Ashley Capoot. 2023. More social media regulation is coming in 2023, members of Congress say. https://www.cnbc.com/2023/01/01/more-social-media-regulation-is-coming-in-2023-members-of-congress-say.html

[27] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! examining the effects of a community-wide moderation intervention on reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.

[28] Hsuan-Ting Chen and Wenhong Chen. 2015. Couldn't or wouldn't? The influence of privacy concerns and self-efficacy in privacy management on privacy protection. *Cyberpsychology, Behavior, and Social Networking* 18, 1 (2015), 13–19.

[29] Danielle Keats Citron and Daniel J Solove. 2021. Privacy harms. *Available at SSRN* (2021).

[30] Vicki Clark. 2019. Meaningful integration within mixed methods studies: Identifying why, what, when, and how. *Contemporary Educational Psychology* 57 (2019), 106–111.

[31] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.

[32] Victor Claussen. 2018. Fighting hate speech and fake news. The Network Enforcement Act (NetzDG) in Germany in the context of European legislation. *Rivista di diritto dei media* 3 (2018), 1–27.

[33] Scott Clifford and Jennifer Jerit. 2014. Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies. *Journal of Experimental Political Science* 1, 2 (2014), 120–131. https://doi.org/10.1017/xps.2014.5

[34] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a replication crisis in empirical computer science. *Commun. ACM* 63, 8 (2020), 70–79.

[35] Sophie Cockcroft and Saphira Rekker. 2016. The relationship between culture and information privacy policy. *Electronic Markets* 26, 1 (2016), 55–72.

[36] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015).

[37] Brian Contreras. 2021. 'I need my girlfriend off tiktok': How hackers game abuse-reporting systems. https://www.latimes.com/business/technology/story/2021-12-03/inside-tiktoks-mass-reporting-problem

[38] Christine L Cook, Aashka Patel, Meciel Guisihan, and Donghee Yvette Wohn. 2021. Whose agenda is it anyway: an exploration of cancel culture and political affiliation in the United States. *SN Social Sciences* 1, 9 (2021), 1–28.

[39] Christine L Cook, Aashka Patel, and Donghee Yvette Wohn. 2021. Commercial versus volunteer: Comparing user perceptions of toxicity and transparency in content moderation across social media platforms. *Frontiers in Human Dynamics* 3 (2021), 3.

[40] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.

[41] Giovanni De Gregorio. 2020. Democratising online content moderation: A constitutional framework. *Computer Law & Security Review* 36 (2020), 105374.

[42] Anna Oberschelp de Meneses, Nicholas Shepherd, and Dan Cooper. 2021. *French CNIL Publishes Recommendations for Protecting Minors Online*. https://www.insideprivacy.com/childrens-privacy/french-cnil-publishes-recommendations-for-protecting-minors-online/

[43] Bryan Dosono, Yasmeen Rashidi, Taslima Akter, Bryan Semaan, and Apu Kapadia. 2017. Challenges in Transitioning from Civil to Military Culture: Hyper-Selective Disclosure through ICTs. *Proceedings of the ACM Journal: Human-Computer Interaction: Computer Supported Cooperative Work and Social Computing (CSCW '18)* 1, CSCW (Nov. 2017), 41:1–41:23. https://doi.org/10.1145/3134676

[44] Natasha Duarte, Emma Llanso, and Anna Loup. 2017. Mixed messages? The limits of automated social media content analysis. (2017).

[45] Marc J Dupuis and Andrew Williams. 2019. The spread of disinformation on the web: An examination of memes on social networking. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 1412–1418.

[46] Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* (2021), 1–20.

[47] Leon Festinger. 1957. *A theory of cognitive dissonance*. Vol. 2. Stanford university press.

[48] Piers Fleming, Andrew P Bayliss, S Gareth Edwards, and Charles R Seger. 2021. The role of personal data value, culture and self-construal in online privacy behaviour. *PloS one* 16, 7 (2021), e0253568.

[49] World Wide Web Foundation. 2019. Contract for the web. https://contractfortheweb.org/

[50] Liridona Gashi and Kathrin Knautz. 2016. Unfriending, hiding and blocking on facebook. In *3rd European Conference on Social Media Research*. 513–520.

[51] David Gefen and Paul A Pavlou. 2012. The boundaries of trust and risk: The quadratic moderating role of institutional structures. *Information Systems Research* 23, 3-part-2 (2012), 940–959.

[52] Nina Gerber, Paul Gerber, and Melanie Volkamer. 2018. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & security* 77 (2018), 226–261.

[53] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[54] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234.

[55] Robert Greszki, Marco Meyer, and Harald Schoen. 2014. The impact of speeding on data quality in nonprobability and freshly recruited probability-based online panels. *Online Panel Research: Data Quality Perspective, A* (2014), 238–262.

[56] Anatoliy Gruzd and Ángel Hernández-García. 2018. Privacy concerns and self-disclosure in private and public uses of social media. *Cyberpsychology, Behavior, and Social Networking* 21, 7 (2018), 418–428.

[57] Timothy C Guetterman, Michael D Fetters, and John W Creswell. 2015. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. *The Annals of Family Medicine* 13, 6 (2015), 554–561.

[58] Nick Hajli and Xiaolin Lin. 2016. Exploring the security of information sharing on social networking sites: The role of perceived control of information. *Journal of Business Ethics* 133, 1 (2016), 111–123.

[59] Eszter Hargittai and Alice Marwick. 2016. “What can I really do?” Explaining the privacy paradox with online apathy. *International journal of communication* 10 (2016), 21.

[60] Eddie Harmon-Jones and Cindy Harmon-Jones. 2012. Cognitive dissonance theory. *Handbook of motivation science* 71 (2012).

[61] Ivar A Hartmann. 2020. A new framework for online content moderation. *Computer Law & Security Review* 36 (2020), 105376.

[62] Rakibul Hasan, Bennett I Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2021. Your Photo is so Funny that I don’t Mind Violating Your Privacy by Sharing it: Effects of Individual Humor Styles on Online Photo-sharing Behaviors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[63] Rakibul Hasan, David Crandall, Mario Fritz, and Apu Kapadia. 2020. Automatically detecting bystanders in photos to reduce privacy risks. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 318–335.

[64] Fitria Fauziah Hasanah and Mela Ermawati. 2022. Sharenting of Young Mothers in Yogyakarta: A Phenomenological Study. *JOYCED: Journal of Early Childhood Education* 2, 2 (2022), 133–146.

[65] Benjamin Henne and Matthew Smith. 2013. Awareness about photos on the web and how privacy-privacy-tradeoffs could help. In *International Conference on Financial Cryptography and Data Security*. Springer, 131–148.

[66] David J Houghton and Adam N Joinson. 2010. Privacy, social network sites, and social relations. *Journal of technology in human services* 28, 1-2 (2010), 74–94.

[67] Roberto Hoyle, Luke Stark, Qatrunnada Ismail, David Crandall, Apu Kapadia, and Denise Anthony. 2020. Privacy Norms and Preferences for Photos Posted Online. *ACM Trans. Comput.-Hum. Interact.* 27, 4, Article 30 (aug 2020), 27 pages. https://doi.org/10.1145/3380960

[68] Mathias Humbert, Benjamin Trubert, and Kévin Huguenin. 2019. A survey on interdependent privacy. *ACM Computing Surveys (CSUR)* 52, 6 (2019), 1–40.

[69] Reddit Inc. 2022. Reddit Content policy. https://www.redditinc.com/policies/content-policy

[70] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. " Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.

[71] Bernadette Kamleitner and Vince Mitchell. 2019. Your data is my data: a framework for addressing interdependent privacy infringements. *Journal of Public Policy & Marketing* 38, 4 (2019), 433–450.

[72] Heather Kelly and Emily Guskin. 2021. Americans widely distrust Facebook, TikTok and Instagram with their data, poll finds. https://www.washingtonpost.com/technology/2021/12/22/tech-trust-survey/

[73] David Kemp and Emily Ekins. 2021. Poll: 75% Don’t Trust Social Media to Make Fair Content Moderation Decisions, 60% Want More Control over Posts They See.

[74] Vivian Kim. 2009. Suicide and “Dog Poop Girl” Lead to Clash Between Google and South Korean Government. http://www.allgov.com/news/us-and-the-world/suicide-and-dog-poop-girl-lead-to-clash-between-google-and-south-korean-government?news=838651

[75] Molly Knefel. 2015. Permanent Records. https://thenewinquiry.com/permanent-records/

[76] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[77] Richard N Landers and Tara S Behrend. 2015. An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology* 8, 2 (2015), 142–164.

[78] Namyeon Lee and Ohbyung Kwon. 2015. A privacy-aware feature selection method for solving the personalization–privacy paradox in mobile wellness healthcare services. *Expert systems with applications* 42, 5 (2015), 2764–2771.

[79] Kalev Leetaru. 2018. Social Media Platforms Are Still Powerless To Stop Data Misuse. https://www.forbes.com/sites/kalevleetaru/2018/12/05/social-media-platforms-are-still-powerless-to-stop-data-misuse/?sh=7b743da4742b

[80] Amanda Lenhart and Mary Madden. 2007. Social networking websites and teens: An overview. (2007).

[81] Fenghua Li, Zhe Sun, Ang Li, Ben Niu, Hui Li, and Guohong Cao. 2019. HideMe: privacy-preserving photo sharing on social networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 154–162.

[82] Chang Liu, Hsiao-Ying Huang, Dolores Albarracin, and Masooda Bashir. 2018. Who shares what with whom? Information sharing preferences in the online and offline worlds. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 149–158.

[83] Natasha Lomas. 2020. Germany: Flawed Social Media Law. https://techcrunch.com/2020/06/19/germany-tightens-online-hate-speech-rules-to-make-platforms-send-reports-straight-to-the-feds/

[84] Mark MacCarthy. 2022. How online platform transparency can improve content moderation and algorithmic performance. https://www.brookings.edu/blog/techtank/2021/02/17/how-online-platform-transparency-can-improve-

content-moderation-and-algorithmic-performance/

[85] Mary Madden. 2012. Privacy management on social media sites. *Pew Internet Report* 24 (2012), 1–20.

[86] Mary Madden. 2014. Most Would Like to Do More to Protect their Personal Information Online. https://www.pewresearch.org/internet/2014/11/12/most-would-like-to-do-more-to-protect-their-personal-information-online/

[87] Bernard Marr. 2018. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=693eeefb60ba

[88] Kirsten Martin. 2015. Privacy notices as tabula rasa: An empirical investigation into how complying with a privacy notice is related to meeting privacy expectations online. *Journal of Public Policy & Marketing* 34, 2 (2015), 210–227.

[89] Kirsten Martin. 2016. Understanding privacy online: Development of a social contract approach to privacy. *Journal of business ethics* 137, 3 (2016), 551–569.

[90] Kirsten Martin. 2018. The penalty for privacy violations: How privacy violations impact trust online. *Journal of Business Research* 82 (2018), 103–116.

[91] Kirsten Martin and Katie Shilton. 2016. Why experience matters to privacy: How context-based experience moderates consumer privacy expectations for mobile applications. *Journal of the Association for Information Science and Technology* 67, 8 (2016), 1871–1882.

[92] Kirsten E Martin. 2012. Diminished or just different? A factorial vignette study of privacy as a social contract. *Journal of Business Ethics* 111, 4 (2012), 519–539.

[93] Adrienne Matei. 2019. Call-out culture: how to get it right (and wrong). https://www.theguardian.com/lifeandstyle/2019/nov/01/call-out-culture-obama-social-media

[94] Lauren B. McInroy. 2016. Pitfalls, Potentials, and Ethics of Online Survey Research: LGBTQ and Other Marginalized and Hard-to-Access Youths. *Social Work Research* 40, 2 (04 2016), 83–94. https://doi.org/10.1093/swr/svw005 arXiv:https://academic.oup.com/swr/article-pdf/40/2/83/7281773/svw005.pdf

[95] Meta. 2022. How enforcement technology works. https://transparency.fb.com/enforcement/detecting-violations/how-enforcement-technology-works/

[96] Marcin Miłkowski, Witold M Hensel, and Mateusz Hohol. 2018. Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of computational neuroscience* 45, 3 (2018), 163–172.

[97] Tobias Mirsch, Christiane Lehrer, and Reinhard Jung. 2017. Digital nudging: Altering user behavior in digital environments. *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)* (2017), 634–648.

[98] Maria D Molina and Shyam Sundar. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication* 27, 4 (2022).

[99] Shara Monteleone, René van Bavel, Nuria Rodríguez-Priego, and Gabriele Esposito. 2015. Nudges to privacy behaviour: Exploring an alternative approach to privacy notices. *JRC Science and Policy Report. Luxembourg, Luxembourg: Publications Office of the European Union* (2015).

[100] Thomas S Mueller. 2021. Blame, then shame? Psychological predictors in cancel culture behavior. *The Social Science Journal* (2021), 1–14.

[101] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.

[102] Jess Nerren. 2021. Preventing Cyberbullying and Online Harassment. In *Handbook of Research on Cyberbullying and Online Harassment in the Workplace*. IGI Global, 468–492.

[103] BBC Newsbeat. 2016. Viral support for meme victim hitting back at body-shamers. https://www.bbc.com/news/newsbeat-3830302

[104] Eve Ng. 2020. No grand pronouncements here...: Reflections on cancel culture and digital media participation. *Television & New Media* 21, 6 (2020), 621–627.

[105] Helen Nissenbaum. 2009. Privacy in context. In *Privacy in Context*. Stanford University Press.

[106] Pippa Norris. 2021. Cancel culture: Myth or reality? *Political Studies* (2021), 00323217211037023.

[107] University of Southern California Libraries. 2023. Organizing Your Social Sciences Research Paper. https://libguides.usc.edu/writingguide/quantitative

[108] National Conference of State Legislatures. 2022. *State Laws Related to Digital Privacy*. National Conference of State Legislatures. https://www.ncsl.org/technology-and-communication/state-laws-related-to-digital-privacy

[109] Yong Jin Park. 2018. Social antecedents and consequences of political privacy. *New Media & Society* 20, 7 (2018), 2352–2369.

[110] Martin Pekárek and Ronald Leenes. 2009. Privacy and social network sites: Follow the money. In *W3C Workshop on the future of social networking*. 15–16.

[111] Sandra Petronio and S Petronio. 2000. The boundaries of privacy: Praxis of everyday life. *Balancing the secrets of private disclosures* (2000), 37–49.

[112] Laura Poppo and Todd Zenger. 2002. Do formal contracts and relational governance function as substitutes or complements? *Strategic management journal* 23, 8 (2002), 707–725.

[113] Evgenia Princi and Nicole C. Kramer. 2020. Out of control–privacy calculus and the effect of perceived control and moral considerations on the usage of IoT healthcare devices. *Frontiers in psychology* 11 (2020).

[114] QSR International Pty Ltd. 2010. *NVivo*. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

[115] Lee Rainie. 2020. Americans' complicated feelings about social media in an era of privacy concerns. https://www.pewresearch.org/fact-tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/

[116] Lee Rainie, Aaron Smith, and Maeve Duggan. 2020. Coming and going on Facebook. https://www.pewresearch.org/internet/2013/02/05/coming-and-going-on-facebook/

[117] Yasmeen Rashidi, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. 2020. "It's easier than causing confrontation": Sanctioning Strategies to Maintain Social Norms of Content Sharing and Privacy on Social Media. *Proceedings of the ACM Journal: Human-Computer Interaction: Computer Supported Cooperative Work and Social Computing (CSCW '20)* 4, CSCW1 (May 2020), 23:1–23:25. https://doi.org/10.1145/3392827

[118] Philip J Reed, Emma S Spiro, and Carter T Butts. 2016. Thumbs up for privacy?: Differences in online self-disclosure behavior across national cultures. *Social science research* 59 (2016), 155–170.

[119] Sarah T Roberts. 2017. *Content moderation*.

[120] Sarah T Roberts. 2018. Digital detritus:'Error'and the logic of opacity in social media content moderation. *First Monday* (2018).

[121] Sarah T Roberts. 2019. Behind the screen. In *Behind the Screen*. Yale University Press.

[122] Aja Romano. 2019. Why we can't stop fighting about cancel culture. https://www.vox.com/culture/2019/12/30/20879720/what-is-cancel-culture-explained-history-debate

[123] Roni M Rosenberg and Hadar Dancig-Rosenberg. 2022. Revenge Porn in the shadow of the first amendment. *University of Pennsylvania Journal of Constitutional Law* 24 (2022).

[124] Minna Ruckenstein and Linda Lisa Maria Turunen. 2020. Re-humanizing the platform: Content moderators and the logic of care. *new media & society* 22, 6 (2020), 1026–1042.

[125] Eldar Shafir. 2013. *The behavioral foundations of public policy*. Princeton University Press.

[126] Tara J Sinclair and Rachel Grieve. 2017. Facebook as a source of social connectedness in older adults. *Computers in Human Behavior* 66 (2017), 363–369.

[127] Anna Cinzia Squicciarini, Dan Lin, Smitha Sundareswaran, and Joshua Wede. 2014. Privacy policy inference of user-uploaded images on content sharing sites. *IEEE transactions on knowledge and data engineering* 27, 1 (2014), 193–206.

[128] Anna C. Squicciarini, Heng Xu, and Xiaolong Zhang. 2011. CoPE: Enabling collaborative privacy management in online social networks. *Journal of the American Society for Information Science and Technology* 62, 3 (2011), 521–534.

[129] Wouter Steijn, Alexander P. Schouten, and Anton H. Vedder. 2016. Why concern regarding privacy differs: The influence of age and (non-) participation on Facebook. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 10, 1 (2016).

[130] Elizabeth Stewart. 2021. Detecting fake news: Two problems for content moderation. *Philosophy & technology* 34, 4 (2021), 923–940.

[131] Stopbullying.gov. 2017. What is cyberbullying. https://www.stopbullying.gov/cyberbullying/what-is-it

[132] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication* 13 (2019), 18.

[133] Matt Swayne. 2016. Users trust AI as much as humans for flagging problematic content. https://www.sciencedaily.com/releases/2022/09/220916112424.htm

[134] Monika Taddicken. 2014. The 'privacy paradox'in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure. *Journal of computer-mediated communication* 19, 2 (2014), 248–273.

[135] Edson C Tandoc Jr, Beverly Tan Hui Ru, Gabrielle Lee Huei, Ng Min Qi Charlyn, Rachel Angeline Chua, and Zhang Hao Goh. 2022. # CancelCulture: Examining definitions and motivations. *New Media & Society* (2022), 14614448221077977.

[136] Chutikulrungsee Tharntip Tawnie and Burmeister Oliver Kisalay. 2017. Interdependent privacy. *The ORBIT Journal* 1, 2 (2017), 1–14.

[137] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

[138] Eran Toch, Yang Wang, and Lorrie Faith Cranor. 2012. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction* 22, 1 (2012), 203–220.

[139] Joseph Turow, Jennifer King, Chris Jay Hoofnagle, Amy Bleakley, and Michael Hennessy. 2009. Americans reject tailored advertising and three activities that enable it. *Available at SSRN 1478214* (2009).

[140] Regina JJM Van den Eijnden, Jeroen S Lemmens, and Patti M Valkenburg. 2016. The Social Media Disorder Scale. *Computers in Human Behavior* 61 (2016), 478–487. https://doi.org/10.1016/j.chb.2016.03.038

[141] Nishant Vishwamitra, Bart Knijnenburg, Hongxin Hu, Yifang P Kelly Caine, et al. 2017. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 39–47.

[142] Arun Vishwanath, Weiai Xu, and Zed Ngoh. 2018. How people protect their privacy on Facebook: A cost-benefit view. *Journal of the Association for Information Science and Technology* 69, 5 (2018), 700–709.

[143] Isabel Wagner and Eerke Boiten. 2018. Privacy risk assessment: from art to science, by metrics. In *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Springer, 225–241.

[144] Thomas Wischmeyer. 2020. What is illegal offline is also illegal online: the German Network Enforcement Act 2017. In *Fundamental Rights Protection Online*. Edward Elgar Publishing.

[145] Pamela Wisniewski, AKM Najmul Islam, Bart P Knijnenburg, and Sameer Patil. 2015. Give social network users the privacy they want. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1427–1441.

[146] David Wright and Charles Raab. 2014. Privacy principles, risks and harms. *International Review of Law, Computers & Technology* 28, 3 (2014), 277–298.

[147] Shali Wu and Boaz Keysar. 2007. The effect of information overlap on communication effectiveness. *Cognitive Science* 31, 1 (2007), 169–181.

[148] Kaihe Xu, Yuanxiong Guo, Linke Guo, Yuguang Fang, and Xiaolin Li. 2015. My privacy my decision: Control of photo sharing on online social networks. *IEEE Transactions on Dependable and Secure Computing* 14, 2 (2015), 199–210.

[149] Rachel Young, Volha Kananovich, and Brett G Johnson. 2021. Young adults' folk theories of how social media harms its users. *Mass Communication and Society* (2021), 1–24.

A PERCEPTION OF IDP QUESTIONNAIRE

Directions: People share lots of information about themselves and about other people online. Sometimes social media users post other peoples' embarrassing or private photos and information without their permission. We are interested in learning about strategies that can be used to minimize online privacy violations and feelings of embarrassment.Next, we will ask you about your opinions and experiences with different types of these strategies for managing inappropriate posts you encounter online. Please carefully read about each of the strategies on the following pages before answering each question.

User-interface tools: Some social media companies provide the option to 'flag' or 'report' a post as inappropriate, so those posts can be reviewed by the company for possible removal.

- Are you familiar with the option to 'flag' or 'report' posts as inappropriate on social media? 1. Not at all familiar; 2. Slightly familiar; 3. Somewhat familiar; 4. Very familiar; 5. Extremely familiar
- How often do you 'flag' or 'report' posts as inappropriate on social media? 1. Never; 2. Rarely; 3. Occasionally; 4. Frequently; 5. Very frequently
- Do you believe 'flagging' or 'reporting' a post is an effective strategy for preventing or removing inappropriate content? 1. Not at all effective; 2. Slightly effective 3. Somewhat effective; 4.Very effective; 5. Extremely effective
- How much do you trust social media platforms to remove or prevent inappropriate content that has been flagged by users? 1. Not at all; 2. Slightly trust; 3. Somewhat trust; 4. High level trust; 5. Extremely high level of trust

Safety Reminders: Social media companies can provide safety reminders that prompt users not to upload inappropriate content or private information.For example, social media companies will ask you to check your posts for dangerous or harmful content before you are able to share.

- How familiar are you with the use of safety reminders on social media? 1. Not at all familiar; 2. Slightly familiar; 3. Somewhat familiar; 4. Very familiar; 5. Extremely familiar
- How often have you received safety reminders when posting on social media? 1. Never; 2. Rarely; 3. Occasionally; 4. Frequently; 5. Very frequently
- Do you think the use of safety reminders is an effective strategy for preventing inappropriate content on social media? 1. Not at all effective; 2. Slightly effective 3. Somewhat effective; 4.Very effective; 5. Extremely effective
- How much do you trust the use of safety reminders provided by social media companies to prevent and remove inappropriate content? 1. Not at all; 2. Slightly trust; 3. Somewhat trust; 4. High level trust; 5. Extremely high level of trust

Interpersonal - blocking: Users can block or unfollow someone who posts inappropriate content to avoid and discourage that behavior in the future.

- How familiar are you with the option of blocking or unfollowing users who post inappropriate content? 1. Not at all familiar; 2. Slightly familiar; 3. Somewhat familiar; 4. Very familiar; 5. Extremely familiar
- How often do you block or unfollow someone who posts inappropriate content? 1. Never; 2. Rarely; 3. Occasionally; 4. Frequently; 5. Very frequently
- Do you believe blocking or unfollowing someone is an effective strategy for preventing or removing inappropriate content? 1. Not at all effective; 2. Slightly effective 3. Somewhat effective; 4.Very effective; 5. Extremely effective

Interpersonal - call-outs: Social media users can call out other users that posts inappropriate content by commenting and describing why their behavior is not appropriate. For example, users might comment on a post by noting it is inappropriate to discourage the content from being spread.

- How familiar are you with user attempts to call out people that post inappropriate content? 1. Not at all familiar; 2. Slightly familiar; 3. Somewhat familiar; 4. Very familiar; 5. Extremely familiar
- How often do you call out people that post inappropriate content? 1. Never; 2. Rarely; 3. Occasionally; 4. Frequently; 5. Very frequently
- Do you believe calling out other users is an effective strategy for preventing or removing inappropriate content? 1. Not at all effective; 2. Slightly effective 3. Somewhat effective; 4.Very effective; 5. Extremely effective
- How much do you trust the use of interpersonal moderation on social media (e.g., blocking or calling out other users) to prevent and remove inappropriate content? 1. Not at all; 2. Slightly trust; 3. Somewhat trust; 4. High level trust; 5. Extremely high level of trust

Automatic Moderation: Social media companies use computer algorithms to automatically detect and remove inappropriate content, for example, pictures with nudity or bullying comments.

- How much do you know about the technology used for automatic moderation? 1. Nothing at all; 2. Know a small amount; 3. Somewhat know; 4. Know a large amount; 5. Know a very large amount
- How often have you noticed posts on social media being automatically moderated? 1. Never; 2. Rarely; 3. Occasionally; 4. Frequently; 5. Very frequently
- Do you believe automatic moderation is an effective strategy for preventing or removing inappropriate content? 1. Not at all effective; 2. Slightly effective 3. Somewhat effective; 4.Very effective; 5. Extremely effective

- How much do you trust the use of automatic moderation on social media to prevent and remove inappropriate content? 1. Not at all; 2. Slightly trust; 3. Somewhat trust; 4. High level trust; 5. Extremely high level of trust

Government policies: Government policies are sometimes put in place to support user safety and privacy. For example, several governments around the world have outlawed "revenge pornography" and hate speech, and require social media companies to remove such content from their websites.

- How much do you know about your government's policies for preventing or removing inappropriate content? 1. Nothing at all; 2. Know a small amount; 3. Somewhat know; 4. Know a large amount; 5. Know a very large amount
- Do you believe government policy is effective in preventing or removing inappropriate content? 1. Not at all effective; 2. Slightly effective 3. Somewhat effective; 4.Very effective; 5. Extremely effective
- How often have you heard of government interventions protecting people's privacy (e.g., in the news, through word-of-mouth, etc.)? 1. Never; 2. Rarely; 3. Occasionally; 4. Frequently; 5. Very frequently
- How much do you trust the use of government policy on social media to prevent or remove inappropriate content? 1. Not at all; 2. Slightly trust; 3. Somewhat trust; 4. High level trust; 5. Extremely high level of trust

Additional Questions:

- How common do you think it is for people to post private or embarrassing information (or photos) about others on social media without permission? 1. Not common at all; 2. Slightly common; 3. Somewhat common; 4. Very common; 5. Extremely common
- How serious of a problem is it when people post private or embarrassing information (or photos) about others on social media without permission? 1. Not serious at all; 2. Slightly serious; 3. Somewhat serious; 4. Very serious; 5. Extremely serious
- Do you think that social media users should take more or less precautions to reduce the number of inappropriate posts on social media (e.g., posts that include embarrassing or private information about others? 1. No precautions; 2. Fewer precautions; 3. Same amount of precautions; 4. More precautions; 5. Many more precautions
- Do you think that social media corporations should take more or less precautions to reduce the number of inappropriate posts on social media (e.g., posts that include embarrassing or private information about others)? 1. No precautions; 2. Fewer precautions; 3. Same amount of precautions; 4. More precautions; 5. Many more precautions
- Do you think that the government (e.g., laws and law enforcement) should take more or less precautions to reduce the number of inappropriate posts on social media (e.g., posts that include embarrassing or private information about others)? 1. No precautions; 2. Fewer precautions; 3. Same amount of precautions; 4. More precautions; 5. Many more precautions
- Who you believe should be responsible for managing inappropriate private content on social media that is posted without permission? Select all that apply. 1. The original person who posted; 2. Other social media users; 3. Social media platforms; 4. Governments
- Who ought to be responsible for educating social media users about the dangers of posting embarrassing or private photos and information? Select all that apply. 1. Parents/guardians; 2. Peers or friends; 3. Schools; 4. Social media platforms; 5. Governments (e.g., social programs)?

  Open-ended responses:

  - Open-ended: What do you believe are the best methods for limiting the amount of private or embarrassing information posted online without permission? Please briefly explain why.

- Open-ended: In your opinion, why or why not is it a serious problem when people post private or embarrassing information (or photos) about others on social media without permission?
- Open ended: What do you believe should be the consequences for users who post other people's private information on social media without their permission?

Received July 2022; revised January 2023; accepted March 2023