Understanding Social Media Users' Perceptions of Trigger and Content Warnings

Muskan Gupta

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Sang Won Lee, Chair
Ha Rim Rho
Donald Scott McCrickard

August 31, 2023
Blacksburg, Virginia

Keywords: Social Media, Triggers, Trigger Warnings, Content Warnings, Content Moderation, Trauma, Sensitive Content, Trauma-Informed Computing

Understanding Social Media Users' Perceptions of Trigger and Content Warnings

Muskan Gupta

(ABSTRACT)

The prevalence of distressing content on social media raises concerns about users' mental well-being, prompting the use of trigger warnings (TW) and content warnings (CW). However, varying practices across platforms indicate a lack of clarity among users regarding these warnings. To gain insight into how users experience and use these warnings, we conducted interviews with 15 regular social media users. Our findings show that users generally have a positive view of warnings, but there are differences in how they understand and use them. Challenges related to using TW/CW on social media emerged, making it a complex decision when dealing with such content. These challenges include determining which topics require warnings, navigating logistical complexities related to usage norms, and considering the impact of warnings on social media engagement. We also found that external factors, such as how the warning and content are presented, and internal factors, such as the viewer's mindset, tolerance, and level of interest, play a significant role in the user's decision-making process when interacting with content that has TW/CW. Participants emphasized the need for better education on warnings and triggers in social media and offered suggestions for improving warning systems. They also recommended post-trigger support measures. The implications and future directions include promoting author accountability, introducing nudges and interventions, and improving post-trigger support to create a more trauma-informed social media environment.

Understanding Social Media Users' Perceptions of Trigger and Content Warnings

Muskan Gupta

(GENERAL AUDIENCE ABSTRACT)

In today's world of social media, you often come across distressing content that can affect your mental well-being. To address this concern, platforms and content authors use something called trigger warnings (TW) and content warnings (CW) to alert users about potentially upsetting content. However, different platforms have different ways of using these warnings, which can be confusing for users. To better understand how people like you experience and use these warnings, we conducted interviews with 15 regular social media users. What we found is that, in general, users have a positive view of these warnings, but there are variations in how they understand and use them. Using TW/CW on social media can be challenging because it involves deciding which topics should have warnings, dealing with the different rules on each platform, and thinking about how warnings affect people's engagement with content. We also discovered that various factors influence how people decide whether to engage with warned content. These factors include how the warning and content are presented and the person's own mindset, tolerance for certain topics, and level of interest. Our study participants highlighted the need for better education about warnings and triggers on social media. They also had suggestions for improving how these warnings are used and recommended providing support to users after they encounter distressing content. Looking ahead, our findings suggest the importance of holding content creators accountable, introducing helpful tools and strategies, and providing better support to make social media a more empathetic and supportive place for all users.

Dedication

To my cherished family, especially Mom, Dad, and my sibling - your unwavering support means the world to me.

To my loving grandmothers, I'm forever grateful for their hard work in providing us with a better life.

To Siddharth, your close friendship and invaluable advice has been my rock.

To my roommates and fellow Powerpuff girls in HCI, Jaitun and Swetha, you've been my home away from home.

To my dear friends, both new ones made and old ones sustained, who eased my journey to the US and stood by me through thick and thin.

And to my past self, who once felt scared and lonely amidst deteriorating mental health but summoned the strength to push through.

This is for each of you, with heartfelt gratitude.

Acknowledgments

I had the incredible fortune of being guided by three amazing mentors during my journey at Virginia Tech: my esteemed committee members, Dr. Sang Won Lee, Dr. Eugenia Rho, and Dr. Scott McCrickard. They have played pivotal roles in my growth as a researcher, scientist, and professional, and their influence has exceeded my initial expectations when I embarked on my master's degree.

Sang, you are the epitome of an exemplary advisor. You introduced me to the world of HCI research, and your CSCW class was the catalyst that ignited my passion for this field. From conceiving the idea during a CSCW class assignment to its submission to CHI LBW 2023, you've been with me every step of the way. Thank you for your unwavering support, even during those last-minute Slack messages and meetings. Thank you for being a guardian and letting me take a break from research when I was crippled with personal trauma. Your motivation was my guiding light during moments of doubt, and I'll forever associate your name with every professional milestone I achieve from hereon.

Dr. Rho, your Computational Social Science class instilled in me the importance of research quality and a critical approach to academic papers. Your invaluable feedback on our "Trans people in sports" project and class readings has shaped me into a better researcher. Witnessing a female researcher like you tackling significant issues at the intersection of Computer Science and Social Science was truly inspiring. Thank you for serving on my committee and checking in on me and my progress whenever we crossed paths.

Dr. McCrickard, your consistent presence throughout my time at the Center of HCI, our engaging conversations about Jaitun, and our chance encounters in the Usability Lab brought brightness to my master's journey. My only regret is not enrolling in your "Models and Theories of HCI" class. Nevertheless, thank you for continuing to treat me as one of your students and for your valuable guidance, encouragement, and positivity.

I extend my gratitude to the Center of Human-Computer Interaction for their financial support, which enabled me to compensate the research participants. I want to express my sincere thanks to every participant, including those involved in pilot studies, for generously offering their time, and invaluable insights, and helping me discover that I enjoy interviewing.

I also want to acknowledge a group of friends and professors who have been instrumental in this journey: Daniel Manesh, Swetha Annavarapu, Jaitun Patel, Siddharth Ahuja, Simran Kalera, Donghan, Marx, Joanna, Liz, Dr. Shaddi Hasan, Dr. Joe Gabbard, Dr. Denis Gracanin, and, especially, Emily Altland. Emily, our collaboration since day one in the CSCW class, conducting initial interviews, analyzing preliminary themes, and co-authoring this work with Dr. Lee, has been an inspiration, and you are truly the best team member I've ever had in a group project.

I owe immense gratitude to my sibling, Momo, whose mere presence and authenticity have fueled my commitment to creating inclusive safe spaces. I'm grateful for my therapist, Ankita Keswani, whose guidance and support were instrumental in my personal growth and development during the challenges I encountered in my master's journey. Last but certainly not least, I want to express my deepest appreciation to my parents, Dr. Mohini Saini and Dr. Praveen Kumar Gupta. Your extraordinary support has made my master's education possible, and your upbringing instilled in me the belief that I can achieve anything I set my

mind to. Thank you from the bottom of my heart.

Contents

List of Figures xii

List of Tables xiii

1 Introduction 1

1.1 Motivation 2

1.2 Research Questions and Contribution 3

2 Review of Literature 7

2.1 Background On TW/CW Usage 7

2.2 HCI Discourse Around Social Media Warnings Usage 8

2.2.1 Sensitive Triggering Content: A Gray-Area 8

2.2.2 Personalized Filtering 9

2.2.3 Specific Photosensitive Warnings More Effective 10

2.3 Social Computing Systems to Keep Users from Triggering Content: What has been done so far? 11

2.4 Knowledge Gap 12

3 Study Design 14

3.1 Recruitment 14

3.2 Participants and Eligibility . . . 15
3.2.1 Participant Demographics . . . 15
3.2.2 Participant Background . . . 15
3.3 Interviews . . . 17
3.4 Analysis . . . 22
3.5 Researcher Positionality . . . 23

4 Results **24**

4.1 Perceived Challenges in using TW/CW for social media content . . . 26
4.1.1 Challenges In Identifying What *Topic* Needs Warning . . . 26
4.1.2 Logistical Challenges with Adding a Warning . . . 37
4.1.3 Tension of TW/CW with Engagement . . . 50
4.2 RQ2: How do people decide to view social media content with TW/CW? . . . 55
4.2.1 External Factors . . . 55
4.2.2 Internal Factors . . . 69
4.3 RQ3: How can the platform improve users' experience with TW/CW? . . . 76
4.3.1 Beforemath: Education of Triggers and TW/CW . . . 76
4.3.2 Design Recommendations for Warning Systems . . . 81
4.3.3 Post-TW/CW Measures . . . 92

5 Discussion **98**

5.1 Impact of Warnings on User Engagement . . . 99
5.2 Specificity in Warnings: A Social Translucence Phenomenon . . . 100
5.3 Balancing Personalizing Warnings and Filtering with Filter Bubbles and Privacy Concerns . . . 101
5.3.1 Privacy Considerations . . . 101
5.4 Author-Based Focus on Features . . . 102
5.4.1 Cultivating Author Accountability . . . 103
5.4.2 Nudging and Intervention Strategies . . . 103
5.5 Enhancing Post-Trigger Support . . . 104
5.6 Emotional Impact on Researchers - Personal Reflections . . . 105
5.7 Limitations and Future Work . . . 107

6 Conclusions **108**

7 Summary **110**

Bibliography **112**

Appendices **118**

Appendix A Eligibility Survey **119**

Appendix B Interview Questions **126**

B.1 Introduction to the Interview . . . 126

B.2 Participant's use of social media 127

B.3 Social media user as a viewer/consumer with current UI 127

B.4 Social media user as a poster/creator with current UI 128

B.5 Social media user and their ideal UI 129

B.5.1 Short version 129

B.5.2 Longer version if time permits 130

B.6 Wrap-Up 130

Appendix C TW/CW Example Slides Shown in Interviews **131**

List of Figures

1.1 (A) TW for a text post on Instagram, (B) CW for an image on Facebook, (C) CW for a page of images like Instagram, (D) CW for a video on Facebook 5

1.2 Examples of TW/CW Usage on Different Platforms (Reddit, Twitter, and TikTok respectively) and Different Modalities (Text, Photo, Video) 6

3.1 Frequency of Posting with Trigger/Content Warning Added 16

3.2 Frequency of Viewing a Post with Trigger or Content Warning 16

5.1 The Cycle of Working as a Researcher on Warnings While Going Through a Traumatic Personal Event 106

List of Tables

3.1 Participant Demographics 18

3.2 Participant Background on Social Media Usage 19

3.3 Participant Background on Social Media Usage (Contd.) 20

3.4 Participant Background on Social Media Usage (Contd.) 21

List of Abbreviations

CW Content Warning

TW Trigger Warning

Chapter 1

Introduction

Katie, a compassionate person, still carried the weight of a miscarriage that shattered her dreams of motherhood, even after months of healing and therapy. One sunny afternoon, an Instagram post from a friend showcasing a decorated nursery triggered her, reopening old wounds. Tears welled up as she quickly scrolled past, feeling the trauma linger. Later that day, she went to a job interview, carrying the weight of her grief with her. It affected her performance, making it challenging to focus and leaving her feeling isolated and overwhelmed.

The story above, although hypothetical, exemplifies the complexity of warning online users with trigger warning (TW) and content warning (CW) labels. On the surface, Katie's friend's post is something that makes it difficult for the poster to imagine it will trigger someone, therefore not considering adding TW/CW. In practice, it can be triggering to someone who has recently had a miscarriage.

Trauma is the experience and resulting aftermath of an extremely distressing event or series of events [11] like violence, abuse, neglect, loss, disaster, war, and other emotionally harmful experiences [22]. With different forms of content stimulating our senses, social media platforms can be risky for those who have experienced a traumatic event(s), which is 70 percent of people [5]. For example, someone who has experienced sexual assault may not want to read about another person sharing their story of surviving sexual assault while on social media.

To protect the vulnerable in online space, social media users use trigger warnings (TW) and content warnings (CW) as ways to shield their followers from being exposed to such content. These warnings are widespread across multiple sectors like health, education, media, arts, and literature [10]. With the rise in the use of social media, especially after the pandemic [37], trauma and traumatic stress reactions can increasingly impact—and be impacted by—people's experiences with technology [11]. TW/CW are used on social media to warn those who may get triggered by sensitive content [10]. There are various ways to add TW/CW *(See Figure 1.1 (A-D) that shows some ways trigger and content warnings can be given depending on the mode of content — text, image, video).*

In addition, other related concepts like the NSFW (Not Safe For Work) tag are used as warnings for media that contain inappropriate content, such as nudity, intense sexuality, violence, gore, or other potentially disturbing subjects matter [13].

1.1 Motivation

Exposure to distressing or triggering content on social media can have adverse effects on people's mental health [29, 41]. Given the recent emergence of TW/CW on social media, the information regarding their usage is relatively unknown. As seen in some of the examples of warnings usage on social media in Figure 1.2, the diverse norms and modalities of warnings across different platforms suggest that *there may be a lack of clear and shared understanding among social media users on the usage of warnings.*

To foster trauma-informed social media interactions, we need detailed insights into the social media users' thoughts and preferences around warnings, which is currently limited [20]. Given the ubiquitous nature of such warnings on social media, we need insights into their usage and viewership by the general user population. Understanding how people utilize TW/CW

on social media is essential, and that is why we aimed to uncover the challenges perceived by average users in contrast to specialists. This study seeks to provide qualitative insights into user's varied experiences and perceptions of TW/CW within the realm of social media.

1.2 Research Questions and Contribution

To better understand the challenges associated with warnings usage and the intricacies of viewing content with such warnings on social media, we conducted a qualitative interview study. We aimed to answer the following research questions:

- **RQ1:** What challenges do users perceive when using TW/CW for social media content?
- **RQ2:** What factors influence the user's decision-making process when it comes to viewing content with TW/CW?
- **RQ3:** What improvements can be made by the platform to enhance users' experiences with TW/CW?

While efforts have been made to address sensitive content through warnings, previous research has only focused on particular features or specific topics. A comprehensive understanding of warning usage on social media remains limited. Filling this knowledge gap is essential in the development of trauma-informed tools capable of efficiently incorporating TW/CW for social media posts, thereby enabling users to navigate sensitive content on these platforms. Our qualitative insights, drawn from users' experiences with warnings, aim to contribute to the development of more effective warning systems by providing more comprehensive challenges in their usage and the intricate decision-making processes regarding content viewing behind these warnings.

In summary, our study brings three key contributions to the understanding of TW/CW usage on social media:

- We investigate the perceived challenges users encounter when employing TW/CW for social media content.
- We explore the factors that shape users' decision-making processes when viewing content with TW/CW.
- We identify potential improvements that the platform can implement to enhance users' experiences with TW/CW on social media.

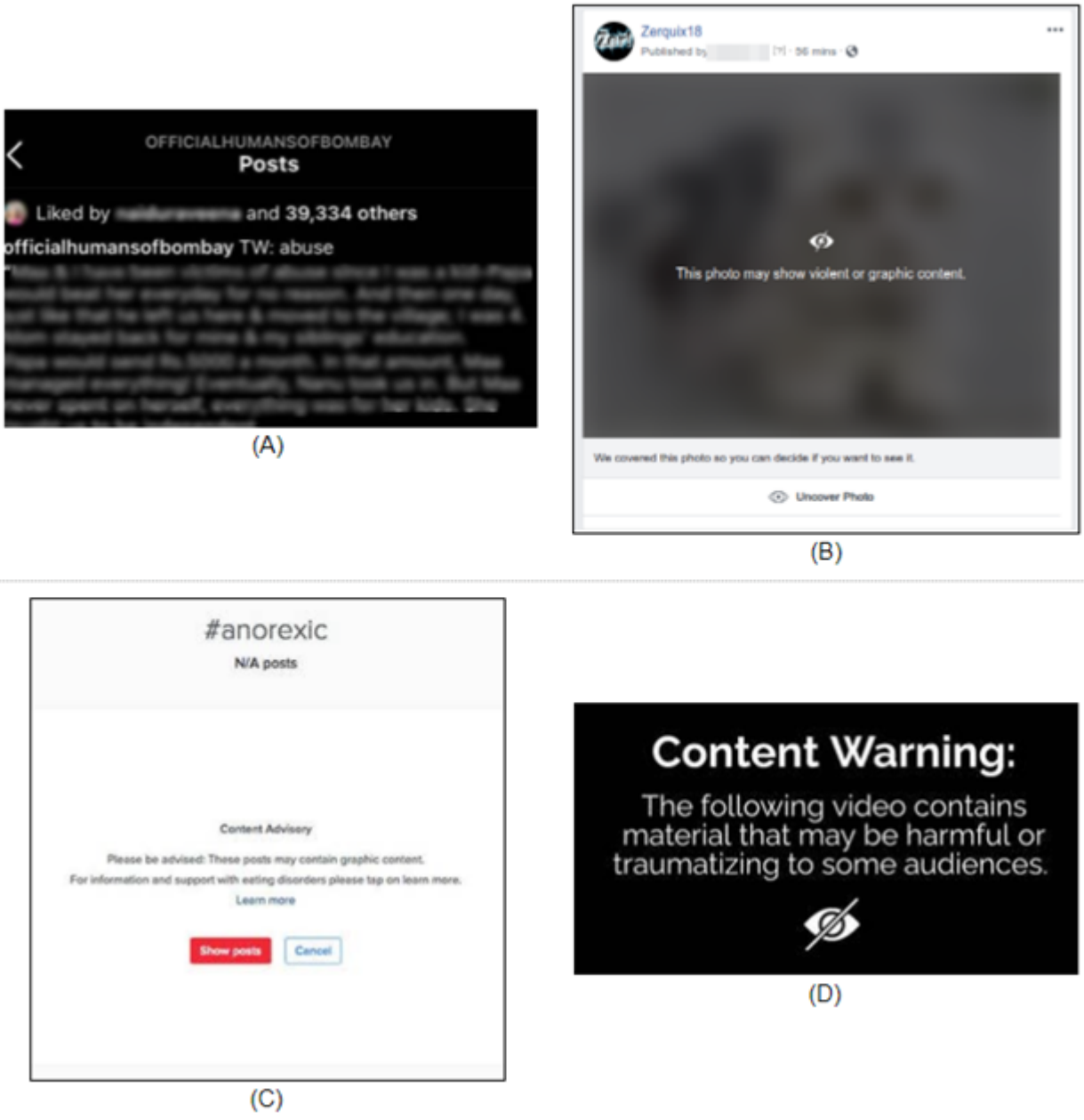


Figure 1.1: (A) TW for a text post on Instagram, (B) CW for an image on Facebook, (C) CW for a page of images like Instagram, (D) CW for a video on Facebook

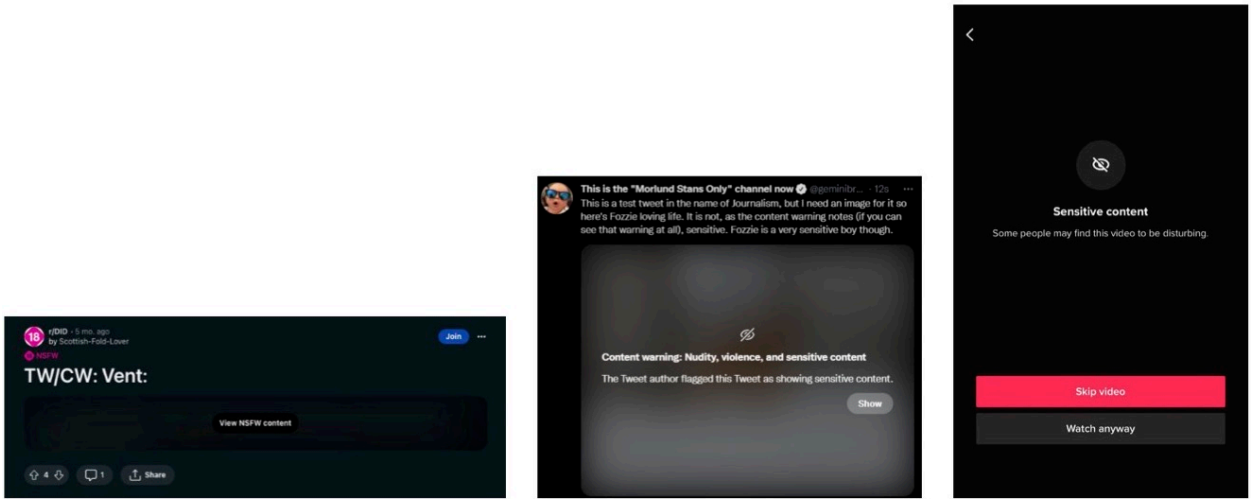


Figure 1.2: Examples of TW/CW Usage on Different Platforms (Reddit, Twitter, and TikTok respectively) and Different Modalities (Text, Photo, Video)

Chapter 2

Review of Literature

2.1 Background On TW/CW Usage

In academic literature, researchers frequently employ the terms 'trigger warning' and 'content warning' interchangeably, although some studies delineate trigger warnings as a distinct subset aimed at individuals who have experienced trauma or suffer from post-traumatic stress disorder (PTSD) [10]. In a recent systematic review, a comprehensive analysis categorized 2,209 warnings from diverse sectors into 14 distinct content warning categories [10].

There have been contrasting views in the literature regarding the utility of warnings. Advocates claim that warnings allow people to prepare themselves, reduce negative reactions toward content [8], and increase individual agency in making informed decisions about engaging with content [10]. On the other hand, critics suggest that warnings may lead to negative interpretations and encourage avoidance, potentially causing harm [8, 9]. Educational contexts have shown that warnings can increase anxiety and reinforce trauma's centrality in an individual's identity [10]. However, the effectiveness of trigger warnings may be influenced by our shared cultural understanding, suggesting that their benefits or drawbacks can go beyond their initial clinical purpose [9].

So far trigger warnings have been studied in clinical settings [23, 39]. However, their prevalence and impact among the general population in online spaces, such as social me-

dia platforms and online forums, have received comparatively little attention, despite their common utilization in these digital environments [17, 28].

To advance this discussion, there is a need to shift away from the oversimplified discourse on whether content warnings have a positive or negative impact. Instead, a nuanced understanding of how content warnings affect different individuals in various contexts and when used for different reasons is required [10].

Therefore, we employ a qualitative approach to gain a nuanced understanding of how users utilize TW/CW and engage with content featuring these warnings on social media through direct conversations with them.

2.2 HCI Discourse Around Social Media Warnings Usage

2.2.1 Sensitive Triggering Content: A Gray-Area

Social media contains diverse sensitive content, ranging from controversial discussions (politics, religion, race, gender) to age-restricted material (nudity, pornography) and graphic imagery (violence, gore). What may be acceptable to some can trigger traumatic memories in others due to personal triggers. Content around reflection systems is one such example that triggers social media users rather unexpectedly. Technology-mediated reflection (TMR) systems, such as Facebook's Memories, curate specific past content for you to reflect on [26]. But sometimes the content, even when the content itself is a positive memory (e.g., wedding photos), with grief attached to it in some way (divorce, death, severe illness), gets shown and evokes *"bittersweet"* feelings.

Dealing with explicit content may seem straightforward, as platforms can use content moderation to remove or hide posts. However, certain sensitive content gets removed that falls into a gray area concerning social media site policies and community norms even though sometimes they do not violate site policies [21]. For example, content related to the transition experiences of transgender individuals can be considered gray-area material. While crucial for LGBTQ+ visibility, education, and activism, it can inadvertently clash with nudity guidelines or be subject to reporting and blocking due to discrimination against personal identities [20]. Sensitive topics like self-harm and suicide present a dilemma between censorship to protect vulnerable viewers and the need to raise awareness for prevention [4].

Another form of gray area content is tied to past trauma, which is context-dependent and challenging to predict. Consider, for instance, individuals recovering from substance use disorder who return to social media after completing treatment. They could encounter images of people using those substances, a common occurrence on these platforms, which can be unexpectedly triggering [30]. This demonstrates that algorithmic knowledge can inadvertently harm users, as seen with targeted weight-loss advertisements negatively affecting people with histories of disordered eating [16].

TW/CWs are often associated with sensitive content [23]. However, as these examples show, moderating such content and navigating triggers online is a complex task. Our research aims to contribute to a deeper understanding of user risks and vulnerabilities when engaging with potentially triggering content, whether or not it includes TW/CW labels.

2.2.2 Personalized Filtering

Trauma-informed approaches are widely considered to be generally beneficial for all people, regardless of whether they are trauma survivors [22]. Trauma-informed computing recog-

nizes that digital technologies can both cause and exacerbate trauma and seeks out ways to avoid technology-related trauma and retraumatization [11]. Chen et al.'s trauma-informed computing framework suggests that engineers and designers could follow the principles of collaboration and enablement to incorporate people's conscious choices into their information feeds. A potential downside of content filtering they discussed is that *"filter bubbles"* limit the content people are exposed to, often without their awareness [11].

On the contrary, Randazzo et al.[31] argue that while filter bubbles may polarize, filtering algorithms can be beneficial for trauma survivors by empowering them to challenge societal filters imposed by institutions. For instance, people in substance use recovery may encounter subtle triggers in digital environments that can accumulate and jeopardize their progress if encountered at the wrong time [30]. To enhance user safety implementing advanced content filters [30, 31] and semi-automated trigger warnings [30] can be valuable. Despite the presence of basic content filters on social media, Phelan et al. [30] observed that participants frequently encountered triggering content, highlighting the limitations of these tools.

Our paper represents a significant step towards enhancing the design of content filtering and other warning-related features. We provide in-depth insights into the challenges linked to TW/CW utilization and the dynamics of users' interactions (viewing or skipping) with warned content on social media platforms.

2.2.3 Specific Photosensitive Warnings More Effective

Through thematic analysis of crowdsourced warnings, South et al. [35] explored the design space for photosensitive risk warnings. They found that providing scenic and temporal contexts rather than just general warnings was more effective in communicating to individuals with photosensitive epilepsy. Therefore, a comprehensive examination of factors influencing

warning effectiveness in aiding viewers' decision-making is essential.

2.3 Social Computing Systems to Keep Users from Triggering Content: What has been done so far?

Some social media platforms already have tools to filter out certain users easily. For example, on Twitter, people can use a blocklist (a list of preemptively blocked accounts from interacting with a subscriber) to avoid harassment/triggering content [24]. However, these features do not allow people to hide only certain posts from a user - it is either blocking the entire content from the specific user or nothing.

To overcome the nuances of content that users may want to be hidden from their view, *SquadBox* utilizes *friend-sourced moderation* as a possible solution to combat challenges with current moderation methods [27]. This method was used to block harassment emails from users who selected specific friends to help review the content. This worked well when only managing direct contact, but friend-sourced moderation will not apply to the overwhelming volume of content on social media platforms.

Warnings are a way not to remove sensitive content that does not violate site guidelines while warning users who can get triggered by seeing that content due to past trauma. A social media site for the transgender community, *TransTime*, dealt with sharing of triggering content around transition by having *Content Warning Tags*, which were similar to hashtags, and enabled viewers to see or not see posts based on the posts' tags [20].

Crowd-sourced approaches to labeling sensitive content seem to be the most successful so far. *Shinigami Eyes* is an extension developed to identify transphobic and trans-friendly social media users and content via community-feedback-driven color-coded labeling [2, 6, 33].

Another Chrome extension, *DeText*, automatically generates content warnings and identifies sensitive text relating to sexual violence by using keyword recognition and sentiment analysis [36]. However, these are both for very specific topics, trans-issues, and sexual violence, respectively. Another website, *DoesTheDogDie.com*, has over 100 categories users can use to pre-review movies, TV shows, and video games for certain triggers or even age-appropriate content [1]. The topics they cover vary from violence/abuse (domestic, sexual, child, elder) to amputation and broken bones, from a dragon dies to shaky cam is used, and a lot more, including hotlines as a resource for topics like self-harm [1]. While this is an exceptional resource for movies, TV shows, and video games, no social media equivalent is available so far.

In the existing literature, TW/CW has been explored within the context of social computing systems, often as a specific feature or in relation to particular cases or topics. However, there remains a gap in our understanding of the comprehensive usage of TW/CW by general users on social media platforms.

2.4 Knowledge Gap

In conclusion, the discourse surrounding TW/CW usage in social media highlights the intricate challenges associated with navigating sensitive content, with or without warnings. While researchers have delved into the design of TW/CW features for a specific case(s), examined content filtering, and underscored the necessity for nuanced approaches, there are persistent gaps in the existing solutions. Various topics, such as transphobia, photosensitivity, substance use disorder, and histories of disordered eating, have been examined individually. However, common challenges that transcend these specific contexts require further investigation. To comprehensively address these issues, holistic user insights are

essential for refining features that facilitate content viewing with TW/CW and assisting content posters in effectively incorporating TW/CW labels. This is where our study makes a valuable contribution to the existing literature.

Chapter 3

Study Design

Because trauma is complicated and we know comparatively less about trauma-informed interactions on social media, we designed our research to be qualitative, which *"is appropriate when the purpose of the research is to unravel complicated relationships"* [32]. We conducted 15 semi-structured interviews with general social media users to understand the experiences associated with content and trigger warnings on social media. All the interviews were conducted after getting approval from our university's Institutional Review Board.

3.1 Recruitment

We recruited participants through our university's graduate student mass email listserv and 'Computer Science Discord' server with a short description of our study, mentioning the eligibility criteria (specified in the following section) and a survey link. We additionally recruited participants by advertising the study on our personal Facebook, and Twitter accounts with the same information and the same eligibility survey questionnaire. On Reddit, we advertised it in the r/mentalheath subreddit. The screener survey asked potential participants for their informed consent to participate in the interview study. We received 181 responses and filtered spam out of those. Using a purposive sampling technique, we contacted 15 people for an interview based on demographics, social media use, and experiences.

3.2 Participants and Eligibility

The participants were eligible for this study if they were 18 years or older, English speaking, currently US citizens or legal residents, and regular social media users who have added or seen TW/CW on posts. Based on the responses to the eligibility survey, we chose participants to interview with diverse opinions and backgrounds. We focused on diversity, not only in race, ethnicity, and gender but also in their responses to how commonly they see the trigger and content warnings on social media and how often they add them to their posts.

3.2.1 Participant Demographics

All participants lived in the United States at the time of the interviews. They had ages ranging from 21 to 33 (M = 25.46, SD = 3.34). Seven participants identified as male, seven as female, and one as non-binary. The interviewee details are presented in Table 3.1.

3.2.2 Participant Background

Our participants were regular social media users, and we did not inquire about their mental health diagnoses. However, they had the option to discuss this during the interviews if they wished. In the pre-survey, we gathered information about their social media platform usage, posting, and viewing frequency, and their experiences with TW/CW. Detailed background information can be found in Table 3.1, 3.1, 3.4. In summary, all participants had encountered TW/CW in some form, and about half of them had posted content with TW/CW at some point (See Figure 3.1 and 3.2).

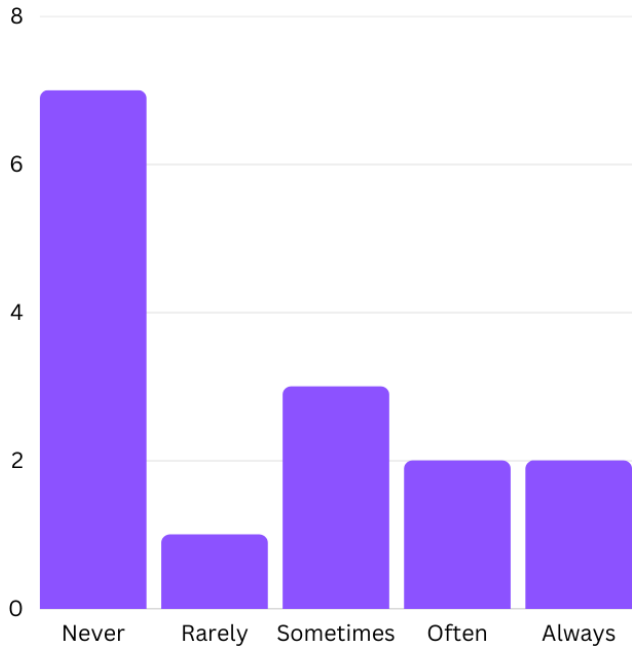


Figure 3.1: Frequency of Posting with Trigger/Content Warning Added

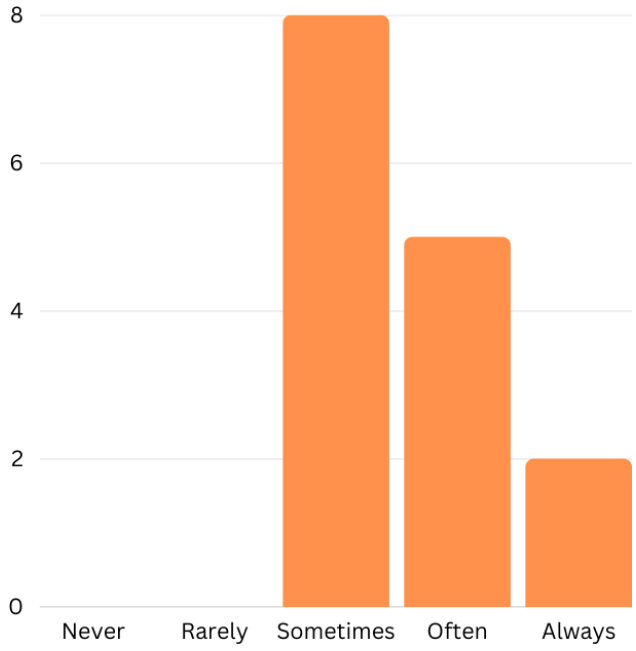


Figure 3.2: Frequency of Viewing a Post with Trigger or Content Warning

3.3 Interviews

At the start of the interview, we reminded participants that they could skip any question they did not want to answer, and if they felt uncomfortable, we would stop immediately. Additionally, we had resources available (e.g., campus hotline) if needed.

We met with these interview participants over Zoom, lasting in the range of 50 minutes to 1.5 hours per interviewee. We asked questions regarding (1) their typical social media use and how it differed by platform, (2) their use of social media and trigger/content warnings as a viewer/consumer of content, (3) their use of social media and trigger/content warnings as a poster/producer of content, and (4) their desirable user experience when it comes to how platforms should handle trigger and content warnings on social media. The questions are in Appendix Section B.

The section on the viewer's perspective from P07 to P15 also included slides with examples of common ways TW/CW are seen on different social media, separated by modality (text, image, video). These were added after reviewing the first six interviews (P01 to P06), incorporating feedback from our preliminary work-in-progress submission. We ensured the examples protected the identity of users who posted the content. We also made sure to use examples in a way that extremely distressing content is not visible, and the interviewee focuses on the way the warning is given. The slides presented during the interview are in Appendix Section C.

Participant ID	**Age**	**Race or Ethnicity**	**Gender**
P01	33	White or Caucasian	Non-binary
P02	27	Asian or Pacific Islander	Man/Male
P03	27	South Asian	Woman/Female
P04	26	Hispanic or Latino	Woman/Female
P05	28	White or Caucasian, North Africa and Middle East	Man/Male
P06	30	Black or African American	Woman/Female
P07	28	Asian or Pacific Islander	Woman/Female
P08	23	White or Caucasian	Man/Male
P09	25	Asian or Pacific Islander	Man/Male
P10	21	White or Caucasian	Woman/Female
P11	21	Hispanic or Latino	Woman/Female
P12	22	Asian or Pacific Islander	Man/Male
P13	25	Asian or Pacific Islander	Man/Male
P14	22	Black or African American	Man/Male
P15	24	Asian or Pacific Islander	Woman/Female

Table 3.1: Participant Demographics

Participant ID	**Social Media Platforms Used**	**Posting Frequency**	**TW/CW Posting Frequency**	**Viewing Frequency**	**TW/CW Viewing Frequency**
P01	Facebook, Discord, Twitter, YouTube, Tumblr, Imgur	Sometimes	Sometimes	Often	Often
P02	Facebook, Instagram, Reddit, TikTok, Pinterest, LinkedIn, BeReal	Sometimes	Never	Often	Often
P03	Instagram, Twitter, YouTube, LinkedIn	Often	Often	Often	Often
P04	Facebook, Instagram, Reddit, Twitter, YouTube, LinkedIn, Twitch	Sometimes	Never	Often	Often
P05	Facebook, Instagram, Reddit, Discord, Twitter, YouTube, LinkedIn, Quora, Twitch	Sometimes	Rarely	Often	Often

Table 3.2: Participant Background on Social Media Usage

Participant ID	Platforms Used	Posting Frequency	TW/CW Posting Frequency	Viewing Frequency	TW/CW Viewing Frequency
P06	Facebook, Instagram, Snapchat, Reddit, TikTok, Discord, Twitter, YouTube, LinkedIn, Twitch	Often	Always	Often	Always
P07	Facebook, Discord, YouTube, LinkedIn	Often	Always	Often	Sometimes
P08	Facebook, Snapchat, Reddit, Discord, Twitter, YouTube, LinkedIn, Twitch	Often	Sometimes	Often	Sometimes
P09	Facebook, Instagram, Snapchat, Discord, Twitter, YouTube, Medium, LinkedIn	Sometimes	Never	Often	Always
P10	TikTok, Discord, Twitter, YouTube, Twitch	Never	Never	Often	Sometimes

Table 3.3: Participant Background on Social Media Usage (Contd.)

Participant ID	Platforms Used	Posting Frequency	TW/CW Posting Frequency	Viewing Frequency	TW/CW Viewing Frequency
P11	Facebook, Instagram, Snapchat, Reddit, TikTok, Discord, Pinterest, Twitter, YouTube, LinkedIn	Often	Often	Often	Sometimes
P12	Facebook, Instagram, Reddit, YouTube, LinkedIn	Rarely	Never	Rarely	Sometimes
P13	Instagram, YouTube, LinkedIn, Quora	Sometimes	Never	Often	Sometimes
P14	Instagram, Snapchat, Discord, Twitter, YouTube, LinkedIn	Often	Never	Often	Sometimes
P15	Facebook, Reddit, Twitter, YouTube, Medium, LinkedIn, Quora	Sometimes	Sometimes	Often	Sometimes

Table 3.4: Participant Background on Social Media Usage (Contd.)

3.4 Analysis

We analyzed the interview data using thematic analysis [7]. We used inductive thematic analysis in two phases (analyzed 6 interviews in the first phase and the rest of the 9 interviews in the second phase) to generate knowledge from the interview data collected.

As the first step, we transcribed the interviews using Sonix.ai and manually corrected the incorrect transcription. We analyzed the transcripts using line-by-line inductive thematic open coding. We then performed axial coding, comparing open codes to find underlying themes. In addition, we used axial coding to examine relationships between the open codes. After a couple of interviews were analyzed, we looked for gaps in our understanding and iterated over interview questions and what sort of participants we needed to help fill those gaps. When we started the study, we focused more on the differences between TW and CW. However, after the first phase of interviews, we realized each participant has a different theory of the difference or does not think there is any difference. Based on the complexity of triggers and the nuances found, we then decided to explore the decision-making users face with TW/CW and the factors influencing the effectiveness of TW/CW. Thus, we performed this iteration of interviews and analysis in 2 cycles. The codes we developed could be grouped into six initial themes:

1. Overall Perceptions of Warnings on Social Media
2. Topics covered by TW/CW
3. Perceptions on adding a warning on sensitive content
4. Perceptions around viewing a warning over sensitive content
5. Factors that make a TW/CW effective

6. Design recommendations surrounding warnings

Here is the link to Miro Board with the Axial Coding- Initial Thematic analysis results: Axial Coding Analysis Link

We conducted another round of axial coding on the above results, organizing them into a tree structure and merging repetitive sections or paragraphs. Based on this structure, the final themes that emerged for each research question are:

1. Perceived challenges in using TW/CW for social media content.
2. Factors influencing the user's decision-making process when interacting with content containing TW/CW.
3. Recommendations for platform improvements to enhance users' experiences with TW/CW.

3.5 Researcher Positionality

Before I present my findings, I acknowledge my standpoint as a cis woman who deals with emotional abuse trauma, being raised in India and currently residing in the United States. I acknowledge that my positionality has somewhat influenced this project; regular feedback and resources from my teammates in this research helped me reduce bias in the results.

Chapter 4

Results

Comprehensively, participants' perceptions of TW/CW were overall positive in our study. The majority of the participants (12/15) mentioned that warnings on sensitive social media content are essential to protect viewers, *"even if it helps just one person" (P03).*

> *Even if we feel like one person could feel triggered, they should. It's not also their opinion; it's for the audience who might watch it or see it. (P09)*

They help viewers make informed decisions on whether to see potentially triggering content and safeguard themselves from the adverse outcomes of unexpected exposure to triggering content (4/15). For instance, three participants, one of whom had a history of suicidal thoughts, experienced situations where posts about suicide lacked warnings, causing them severe distress or concern. They also help viewers be prepared for the content coming forth (6/15) and take breaks from consuming a lot of potentially triggering content. Even if social media viewers want to view the sensitive content, through warnings, they can mentally prepare themselves *"going in, it's going to be a bad time" (P01).* For P07, warnings helped take breaks from watching too much sensitive content related to crime and gore.

> *So it's* [gore and crime content] *something that I know I enjoy. It's kind of my entertainment. So I don't mind trigger warnings. I don't really avoid them, but I know I need to actually have some breaks. ... And that's why the content or the warnings, really help. (P07)*

Participants (5/15) thought warnings facilitated challenging discussions and sharing heavy experiences rather than censoring them. P15 explained that it serves as a recognition of varying perspectives or the potential for differences in interpreting a topic that could be offensive to some individuals. They increased comfort, reduced hesitancy, and alleviated some of the anxiety or guilt associated with sharing potentially triggering content, as illustrated by the following comment from P08:

> *Let's say I was a survivor of sexual assault. And for me, being able to communicate that, just like to the universe is helpful for me, I would feel better about it if I could put something like a content warning over it so that somebody who might be triggered by it doesn't happen to stumble upon it, or at least they have to interact to be able to see it. So, I mean, I yeah, it serves a use I think it serves a, a use that is good. (P08)*

They also found (5/15) adding warnings a sign of thoughtfulness, politeness, and being considerate, as seen in P11's comment:

> *I really appreciate when they're added. … I just really, like, respect the time and attention they put to it, like. Even if it's something small, it really tells me a lot about that person. Like how considerate they are. (P11)*

In general, participants had a favorable view of warnings on social media. They felt that these warnings helped users make informed decisions, avoid the negative impacts of encountering upsetting content, mentally prepare themselves, take breaks when needed, and feel more at ease when sharing difficult experiences. Additionally, using these warnings was seen as a courteous and considerate gesture.

4.1 Perceived Challenges in using TW/CW for social media content

Although a small fraction of participants (2/15) found it relatively straightforward to determine which content required warnings, the majority, including folks who have not posted with warnings, faced challenges in this decision-making process. Identifying which topics need warnings, dealing with logistical issues, and navigating the balance between providing warnings and maintaining user engagement added complexity to the endeavor.

4.1.1 Challenges In Identifying What *Topic* Needs Warning

Topics Depending on Personal Context

There were certain topics on which the participants unanimously agreed that are expected to include warnings:

- Violence, abuse (animal abuse, sexual violence, rape, sexual abuse, child abuse, mental abuse, domestic violence, domestic abuse, murder, instigation of violence, police brutality, gun violence), and graphic content.
- Suicide and self-harm.
- Violence is committed against marginalized groups (people of color, women, children, LGBTQ+ people, religious minorities [Islamophobia], hate crimes)
- Flashing lights or other triggers for photosensitive epilepsy.
- Discrimination or hate speech against marginalized groups [Islamophobia, ableism, sexism, homophobia, transphobia, racism, harassment of women].

These matched the 14 types of content warning categories synthesized by Charles et al. [10].

However, one significant way the participants (8/15) felt adding TW/CW is complicated lies in the nuanced and diverse nature of individual triggers. These triggers vary according to a user's background, personal experiences, and community, rendering them non-universal.

> *What is a trigger for me might not necessarily be a trigger for you. (P03)*

Triggers are vast and contextually vary: Adding warnings becomes complex due to the impossibility of knowing all potential triggers, given the vastness of trauma experiences. Moreover, triggers can vary significantly among individuals and may evolve over time. P11 illustrated that for a person, a minor paper cut might be enough to trigger a reaction, while for someone else, a more graphic image might be needed to produce a similar level of triggering response.

> *P-11: Um, I think because it's so, like, subjective, like there's not really. Like a clear answer or clear guidelines to what a trigger is. You know, like a paper cut might be really triggering for someone, but like. You know, it might take like a really gory picture to be triggering for someone else.*
>
> *Interviewer: Right.*
>
> *P-11: So, yeah, I think it's just so subjective. And that's where the difficulty comes from.*

One participant (P02) also mentioned that the threshold or capacity of people to handle potentially triggering content may vary widely.

> *But someone's threshold could be very different from mine. So maybe something*

> *that I don't fathom that would be triggering to someone can be. So I don't put trigger warnings on that content. (P02)*

Moreover, triggers can be influenced by cultural norms within specific contexts. For example, when P07 was discussing child abuse, they highlighted that specific actions, such as parental discipline commonly accepted in their Asian cultural background, might be considered abusive in the US, emphasizing the contextual variability of triggers.

> *A lot of things that is considered trigger here* [in the US]. *Let's say we're both Asians, right? So being beaten up by our parents, they're not really abuse for us back home, but here. Oh, wow. It's. It's actually an abuse. You can be reported or something. (P07)*

Some topics mentioned by the participants that may potentially trigger a section of users who had undergone similar incidents, depending on their context were:

1. **Traumatic Incidents like Grief:** One participant (P05) mentioned that discussing topics related to grief, the passing of loved ones, or the illness of close family members could potentially be traumatic, depending on one's personal life experiences.

2. **Sharing Negative Experiences related to Social Issues:** Another category involved individuals opening up and discussing somber experiences related to traumatic events. For example, people share their encounters with racism, sexism, LGBTQ+ rights, or abortion. P15 said sharing stories about Roe V. Wade's overturn [3] on abortion is driven by the need to convey emotions and feelings, rather than just presenting facts.

   > *Like, abortion. It's more like, this is about, like, you know, how like, is more like that given the recent event, how it was overturned, I would have to share*

> *a story. This is sad and disturbing. And these my... Yeah, like it's more about feelings and emotions than just describing the fact. (P15)*

However, it might trigger people who have gone through similar experiences. P15 further explained that sharing such experiences and knowing *"I have experienced these"* or *"my friends experience that"* or *"I saw my friends going through these"* become relatable, even when the stories are from a third-party perspective.

3. **Political Content:** Another area mentioned by participants was political content like white supremacy, eugenics, and political beliefs since they might upset someone or cause anxiety. For instance, P02 highlighted their anxiety regarding political content on social media and emphasized the importance of warnings. They believed such content could be distressing and anxiety-inducing for themselves and many others on social media.

> *I think, for me personally, a lot of political stuff when it's posted on the Internet (debates are going on, people opposing views, they're hating on each other) that kind of induces anxiety in me. So I feel like that you can count that as triggering. And I'd rather see a post in front of that content saying, "Hey, this is related to political debates, unnecessary political debates that they're having". ... Not to take away anything from the moment or the [political] event itself...I think it can be triggering for a lot of people that don't want to see that on social media. (P02)*

4. **Immigration and Deportation:** Three participants mentioned that disturbing events around deportations, immigration, the border, and international students getting laid off should also include warnings. P04, an international student, shared their experience during the pandemic when the possibility of deportation loomed. They noted

that the absence of TW/CW on immigration-related topics significantly impacted their well-being and led to a temporary break from social media.

> *P-04: The one time I remember it happening was peak pandemic, I think. Um, and it wasn't really due to COVID, but as an international student, I was about to get deported.*
>
> *Interviewer: Oh, okay. I'm sorry.*
>
> *P-04: And there were a lot of people putting a lot of information and misinformation. And it affected me more than I thought it would. Just talking about immigration.*
>
> *Interviewer: Right.*
>
> *P-04: So. Like I mean, I guess not just general immigration, but, you know, like deportations, immigration, the border, the whole kids in cages and everything. I didn't see enough trigger warnings on that, and especially when I was in that situation myself. That really messed me up. I really had to take a break from social media, therefore.*
>
> *Interviewer: Yeah. I'm sorry about that.*

P09 mentioned another example, also an international student, where they found LinkedIn posts during layoff season sharing personal life details while looking for a job after getting laid off triggering and emotionally distressing.

> [Linkedin Posts about layoffs] *Those seeing those...(sigh)... Like we [international students] came here with a lot of hope. So, I mean, like, you lose that hope* when you see those kind of posts and all. I mean, like. If you don't have that hope and all, you cannot concentrate on your study and all. (P09)

5. **Additional Topics:** Some other topics briefly mentioned by participants that might trigger users with past traumatic experiences were eating disorders, food insecurity, homelessness, abortion, drug abuse, and drug usage.

Hence, deciding whether a topic should or should not include a warning is a complex undertaking due to the diversity and subjectivity of triggers. This makes it a challenging and constantly evolving process to address users' varying sensitivities and backgrounds.

Purpose of Sharing Triggering Content

Determining whether a warning is necessary can be challenging, as the intention behind sharing triggering content plays a significant role. For example, sharing a video of someone helping a skinny dog on the roadside might trigger some, even though the intention is to assist the animal (P13). In such cases, it becomes tricky if a warning should be used. P11 faced a dilemma about adding TW/CW for art that resembled gore or blood but was created using makeup.

> *I was, like, conflicted because the way they did their makeup made it seem like it was blood, but also, like, you couldn't really...it was like an in-between place. Like, I couldn't tell if it was the makeup or like, Yeah. Or, like, if they were intentionally trying to make it look like gore or things like that, I ended up putting the warning anyway, but like it did, I had to stop and think about it for a long time. I was like, "Should I?". (P11)*

Also, warnings may be included depending on how topics like mental health and race are covered. Educating about a mental health disorder (P03) or the concept of racism (P04)

would likely not have a warning, while a personal story about suicide ideation (P03) or human experience with racism (P04) would.

In summary, the necessity of issuing warnings for potentially triggering content also hinges on the intention behind sharing such material, making it a challenging decision.

Complex Decision-Making With Topics That Are Marginalized

Some participants (6/15) experienced inner conflict when deciding whether to include TW/CW in their posts. This conflict often arose when they wanted to raise awareness about a particular topic, especially one that disproportionately affected minority communities. On one hand, they believed it was crucial for such content to reach a broader audience, including those outside the affected community. However, they also acknowledged the potential distress that repeated exposure to traumatic events could inflict on the community. This conflict was particularly pronounced in discussions of social movements. For instance, one participant (P11) articulated the struggle by recognizing the importance of disseminating information about such movements while simultaneously acknowledging the potentially triggering nature of the content.

> [on social movements like BLM] *That's kind of hard to like, think about because, like, it is. You know, on one hand, you want people to know about it, but on the other hand, like. That in itself might be like. Really, really hard for someone to see. ... I have mixed feelings. (P11)*

Participants had varying opinions on whether social movements should come with TW/CW. One participant (P05) argued that political discussions surrounding issues like racism, while uncomfortable, did not require warnings. Regarding the Roe V. Wade verdict, another participant (P02) emphasized that the related stories shared during that period should not

come with warnings. They held the view that *"some things need to be thrown at your face; you should not be able to avoid them"*.

P03 cautioned that warnings could sometimes be used to restrict discussions of taboo topics, particularly those related to gender and sexual minorities.

> *So my friend does this art. She uses period cups like menstrual cups. And she uses... Like she stores the blood. And then she uses it to paint. So she has a lot of paintings put up on her Instagram. And of course, it talks about menstrual blood. And then she also has like, drawings of vaginas, like different vaginas and things like that. And people are like, "Oh my God, how can you do this? You have to give it content warning!". I follow the page, I think it's called the Vagina Project or something, which basically is demystifying the fact that vaginas are not supposed to look a certain way and they're not supposed to smell like flowers and they're not supposed to, you know, everyone has a different vagina. So there are a lot of vaginas. And it's sad that they have to put a content warning to it. (P03)*

They further mentioned that some topics, like menstruation, should not need a warning, asserting that certain aspects of life do not come with warnings in reality.

> *If you want to discuss menstruation, you're supposed to... People need a content warning for that. And I'm just like, "No, I'm sorry, I'm not going to give you a content warning." My body doesn't give me any warning, and I'm going to give you a content warning for this posting about menstruation. (P03)*

However, P14 emphasized the importance of prioritizing users' emotional well-being on social media, stating that *"you are only one person, and there is only so much you can take."*

Furthermore, P14 explained that if users have already been exposed to a triggering event, then adding TW/CW might not be as helpful since they already know what is to come. If it is first gaining momentum, then it might be crucial to add TW/CW.

In summary, participants faced complex decisions about when to use TW/CW in their content, balancing the need to raise awareness with concerns about potential distress to audiences.

Less Commonly Triggering Topics Do Not Have A Warning Present

Participants noted that they encountered content on certain topics that most people did not find triggering, so these topics were often overlooked when it came to adding warnings. The reason these topics were perceived as less commonly triggering by the participants was that they rarely saw any warnings associated with them. Additionally, they expressed a preference for warnings to be applied in such cases.

1. **Phobias:** One such category, mentioned by P11, was on different types of phobias, like a phobia for holes or needles. They also understood why their fear of needles is not considered a trigger.

   > *Anything with, like, needles, is like my main trigger. ...I feel like everyone has gotten, like, a vaccine or, like, has been around needles before, so, like, it's not really like a trigger or considered one. (P11)*

2. **Natural calamities and war:** Three participants mentioned that they wish posts around natural calamities, human error like plane crashes, or war-related content included TW/CW.

   > *Yes. Syria and Turkey earthquake recently happened. So you might just see*

> *a normal video or something. Also, Russia and Ukraine war, right? Like you see a missile that is being hit, a building. You don't see anyone dying, but someone might be dying in it. So it can be, or it cannot be, a trigger for someone like you specifically don't see someone dying. But if a missile has hit a building, at least there would be some casualties. (P13)*

3. **Mass shootings in the US:** P15 described the difficulty in deciding whether to include warnings for content related to gun violence and mass shootings. They recognized that due to the frequent occurrence of mass shootings in the United States, people have become desensitized to them. However, they wished for warnings on such content for individuals who have been directly affected by these events:

> *Maybe like gun violence and control should be one. But then I know that. And there are Americans out there believing that is essential thing to have. And they are like shooting cases happening every day. Why bother? ... But sometimes it's hard for me to imagine, like, you know, like parents seeing those news because actually, they have their children, like passing away because of those cases. So. Maybe. Yeah, like. But then I feel like it's really about public perception. Like, of you know what might be a trauma, right? Yeah. (P15)*

4. **Audio triggers:** In terms of modality, two participants mentioned that audio content and related triggers are currently being overlooked. There are instances where there is no visual-like trigger, but audio-wise, it exists like the scream of *"somebody when they get killed or murdered, or raped"* (P12). In those cases, P12 said there should be a warning.

> *Interviewer: And, like, has there ever been, like, an experience where you*

read or viewed something and you wish that had a trigger warning present? Like any sort of specific topic?

P-12: I mean, Oh, yeah. Like, there are some things that wasn't visually triggering but audio triggering. I don't think the filter can filter those kind of stuff.

Interviewer: Right. What were they like? Was it like just screaming and stuff?

P-12: Screaming Like it was a moaning sound. I think they got the sound from pornography. Like there was a waterfall of videos. Waterfall, which was like, totally PG.

Interviewer: That is so weird.

P-12: It was like a prank video.

Interviewer: Okay. That's that's interesting. Yeah, I guess our audio triggers could be something that.

P-12: Well, like if somebody was previously sexually assaulted it might trigger those kind of people.

In summary, one part of the complexity in deciding what topics should have a warning depends on how commonly associated it is with triggers or does it have lower awareness. Topics like phobias, natural disasters, war, gun violence, and audio-based triggers might be acceptable for the majority of users but according to some participants should also be considered for the inclusion of warnings, especially for the minority who may find them distressing.

4.1.2 Logistical Challenges with Adding a Warning

Participants also found it challenging to include warnings due to the absence of a consistent norm for TW/CW. Some social media platforms had dedicated warning features, while others did not, leading users to come up with their own solutions. Some users even repurposed existing platform features for adding warnings. Furthermore, within the same platform or communication method, there were various ways to add warnings, each with its own level of detail. These factors collectively made it complex for participants to manage the logistics of adding warnings.

Platform Difference

Different platforms have various methods to display warnings for the same triggering content. This variation is often a result of authors having access to different features based on the platform they are using.

For instance, platforms like Discord and Reddit offer *spoiler features* that allow creators to hide potentially triggering (text or multimedia) content until the viewer specifically clicks to reveal it. Although these features were not originally designed for this purpose, people have started using them to add TW/CW to triggering content. For instance, P10 mentioned that the spoiler feature on Reddit is commonly used for concealing triggering information. On Twitter, P10 also highlighted its potential to hide the images:

> *I know that Twitter, I know that Twitter has a spoiler option for images, and I don't know if I've seen it, but I feel like it would be a good thing you could utilize the spoiler feature, which should have like blur out the image until you click on it. And I would say I'd say certainly some people are using it for that purpose. (P10)*

Participants had mixed opinions about these features. P04 appreciated how Reddit blurs content, allowing users to decide whether or not they want to view potentially triggering material.

> *I personally like that Reddit blurs it because. ... Because if I don't feel like I can take any triggers today, like I can just skip right through if I'm like, "oh, maybe this is something", you know, like, it gives you even the option to see what the trigger is. (P04)*

However, P08 found this feature inaccessible, *"because you have to know how to do that"*.

Features not found in other platforms led to user workarounds: These spoiler features were not found on other platforms like Instagram, and Facebook. Four participants wished Facebook/Instagram had text and visual blocking features, like Reddit's text-based spoiler feature or Twitter's image-covering feature, for the author to add warnings effectively.

> *I think for Facebook, they would need... I think they should also blur the text that's there. They only blur the image. I think they should also blur the text or make it not visible because that kind of tells you what's going to be there. And like that can, for some people, it can draw that image in their head already. (P05)*
>
> *Yeah. I kind of wish that Instagram had like the thing that Reddit has where the... It has like the blocking. (P04)*

In platforms like Instagram, where authors lacked control over adding warnings through features like spoilers, new methods emerged. Participants (8/15) mentioned one of them as the *first-slide method*, where the first slide of the story or post gives a warning or background information about the sensitive visual content, and the next slide onwards is the actual

content. Participants noted that creators use this method as a way to effectively provide a warning, with one participant describing that *"It's your own way of blurring it out" (P04).* P11, who primarily reposted triggering content in their stories, employed stickers of various shapes to cover sensitive content in each story or slide, rather than relying solely on the first story/slide method.

Lack of consistency in adding warnings across platforms: Variations in adding warnings across different platforms signaled a lack of consistency in the format to add TW/CW. This led to some of the frustrations with them. *"It's just up to whoever is making the post."* (P10) One participant (P13) expressed a desire for consistent, cross-platform tools to make warnings more effective. They noted that widespread adoption would depend on the availability of such features across multiple social media platforms.

> *Like if a certain application or certain social media is posting it, and if everyone follows it, then I guess the user will also start incorporating it. But if there are like ten social media and only one or two have such features, I don't think there would be a mass adoption from everyone using that thing. (P13)*

While P05 suggested there should be more consistent ways to add TW/CW on a platform but not across all platforms.

> *I think that it's helpful to have different terminologies, but they definitely don't do a good job in, at least the ones I've seen on Reddit in making it like I think they should be consistent across the platform, but they shouldn't be the same across all platforms if that makes sense. Like across the same platform, they should be consistent. Across all platforms, I don't think so, because I think, for example, like I think having a video that is disturbing is a lot more, in my opinion at least,*

> *is a lot more impactful on a person than as text. ... So I think that YouTube has that three clicks [warning] that they have. I think they should keep that. However, [on] Reddit, I don't think you should need to confirm three times if you want to read a post in that regards. (P05)*

Overall, the varying approaches employed by different platforms for issuing warnings, along with users adopting different tools based on platform-specific features, presented logistical challenges when it came to adding warnings.

Modality Difference

Participants provided insights into the diverse approaches taken by creators to add TW/CW labels to sensitive posts based on the content's modality.

Text-based content: For text-based content, the majority of the participants (9/15) found that warnings include "TW" or "CW" or their full forms, followed by a descriptor of the content being warned about, such as "TW: Sexual Violence." However, there were some variations in how participants used these terms and descriptors. These approaches included using asterisks (e.g., *TW* abuse, P03), employing big, bold caps to emphasize warnings in text-based Reddit posts (P05), utilizing clean and complete sentences presented in all caps lock format (P07), adding warnings as \\CW or \\TW before elaborating on a situation (P10), using abbreviations like "SA" following the warning (P11), and crafting sentences that set the scene of the post while incorporating phrases that might be triggering (P15).

In some cases, posters may include disclaimers in their content but not explicitly label them as "Trigger Warnings" or "Content Warnings." For example, P01 mentioned that in the text, before talking about the content, they put: *"Hey, this is what's here beforehand"*.

Using the spoiler feature, some people blocked the entire content of the text, while some

blocked parts of it. For instance, P10 mentioned that the spoiler feature on Reddit is commonly used for concealing triggering information. They also noted that *"if you mark the link in spoiler tags on Discord, it will blur the thumbnail"*.

Location of text-based warning varied: Location of the text-based warning also had variations. Warning was found in the first sentence(s) of a text-based post by the participants, sometimes presented in all uppercase letters (P15). For instance, P04 explained that Reddit users often place the warning in the first few sentences of a post so that viewers can see it before expanding or opening the post. On platforms like Twitter where content can be posted in the form of threads, P10 elaborated that users often add TW/CW in the first tweet and then provide additional information or context in the same tweet or in subsequent tweets. Using the spoiler feature, the warning could also be inline when blocking parts of text. While P01 mentioned that warnings might sometimes be included in hashtags, they noted that this is not very effective because hashtags tend to be placed at the end of posts.

In summary, participants noted that the usage of warnings varied even within the text-based modality.

Visual/Multimedia content:

For visual/multimedia content, participants mentioned diverse methods for adding warnings, including the blur method, first-slide method, spoiler feature, using text fields within the visual content, and spoken warnings within the content.

1. **Blur method:** The blur method, as described by participants (12/15), involved adding a blurred or grayed-out overlay to visual content, often accompanied by descriptor text. Viewers must click or interact with the content to reveal it.

   > *At least on Twitter, where like you can it will grey out the image or the video or whatever, and then you have to choose to interact with it. ... [on Reddit]*

> *It will be grayed out too or like there will be an overlay of filter over the top and then you have to interact or click on it to view it. Same thing. It's pretty similar to Twitter, I guess. (P08)*

In certain circumstances, descriptor text is included along with the blurred content, providing viewers with an idea of what to expect. For instance, P15 mentioned that Facebook uses the blur method for videos, where the video initially appears blurry, and a warning message is displayed in uppercase letters. After a few seconds, the video begins to play.

> *[on Facebook video] When you click the play button, it's like blurry and then there's like the text in the middle in upper cases saying that, like "this video may contain like triggers or things that make you feel uncomfortable". ... I'm pretty familiar with the video, how it is blurry at first. The video on Facebook, how it was blurry at first. And then onto like. And then for like a few seconds before you actually start watching it. (P15)*

In some cases, the spoiler feature was used to conceal potentially triggering visual content. For instance, P13 mentioned encountering this feature on YouTube, where it is used for age-restricted content, requiring users to sign in before accessing it. P07 also noted that on YouTube, a pop-up warning is displayed for adult or graphic content, providing an added layer of protection.

Participants noted that both platforms and users employ the blur method. Seven participants mentioned that platforms automatically apply the blur to sensitive content, while two participants said that the blur is user-added.

2. **First-slide method:** Another method used for visual content was the first-slide method mentioned by the participants (8/15), where the author adds a warning in

the form of the written text in the first slide of the story or post. On one hand, P10 explained they like the first-slide method because it warns people first and does not show everything unless they scroll through the images. However, P11 noted that when posters use the first-slide method, they might assume that the user who clicks next wants to see the sensitive content.

3. **Using text fields with the visual content:** Participants (11/15) noted that warnings accompanying visual media often appear in one or more of the text fields, such as the caption, title, or description, of the visual post. This method was sometimes combined with other approaches- which meant that the visual content was covered by some means. For instance, P05 mentioned that when posting an image with a caption, they prefer to place a prominent warning at the beginning of the caption, often using uppercase letters for emphasis. When discussing YouTube, P08 mentioned encountering TW *"as shorthand for trigger warning"* in video titles, especially on political podcasts they follow. Moreover, on Twitter, P10 described users often incorporated a warning within their tweet, possibly accompanied by an emoji like an exclamation point, before sharing the sensitive video. This warning could be within the same tweet or provided in replies to the original tweet.

> *I definitely more often see like content warning, maybe like an exclamation point, exclamation point there, like an emoji with like exclamation point or something. And then it's just the video right there. ... I'm trying to think I've definitely seen some bad cases where it will just be like the image and then just say, like, content warning, whatever, and the image is still there. Yeah. But I have seen some where they will put the images or anything in, in the either replies or in a thread. (P10)*

In some cases, warnings are embedded as text within the video itself. P15 illustrated

that TikTok videos sometimes featured warnings in the same area where subtitles are typically placed, often superimposed over the person's forehead as they spoke.

4. **Spoken warning:** For issuing warnings specifically in videos, the authors provided a spoken warning. For example, P04 mentioned encountering spoken warnings in short videos on Instagram where people would verbally announce a trigger warning before discussing sensitive topics such as sexual abuse.

> *I've also seen it in reels. In Instagram's there are little short videos. Some people will be speaking to me like, "Oh, trigger warning, I'm going to talk about sexual abuse or I'm going to talk about...". And I think that's very helpful because it gives you time to like scroll and just avoid it. (P04)*

Spoken warning with timestamp: As a workaround for longer-form videos, individual YouTube channels placed the warning at the start of the video and also gave a timestamp, allowing viewers to skip past the sensitive content and proceed with the rest of the video.

> *So the poster, like you said, the poster actually make it like edit it in their video: "This is content warning". Some of them say "go to minute 4:19 if you if you're not into this" something like that or they can go as far as I'll click away. "Don't watch this video at all if you're not into this". So. Right. And I like that. (P07)*

Overall, the participants' accounts revealed significant variability in the usage of warnings for different content modalities and even within the same modality. This diversity further complicated the logistical challenges associated with adding warnings.

Specificity Difference

Different methods of adding warnings have different levels of specificity in terms of how much information is given to the viewer. Visual warnings, like blurred images or videos, can also vary in specificity, ranging from providing no context to hinting at the content's nature through colors or text descriptors. For instance, P12 admitted that encountering a blurred warning made them intrigued about what might be hidden behind it based on the visible colors in the blur.

> *I don't know if it's to me personally, it makes me more curious of what's underlying like behind the blurred picture. You know, if it's like skin color, maybe it's a nude; if it's a red color, maybe it's blood. (P12)*

P07, on the other hand, appreciated a case of image blur warning for its complete obscurity, leaving no identifiable information visible, including the post's author and caption. Because they *"don't want to have imagination sometimes about it."* (P07)

Adding descriptors to warnings enhances their specificity. The photosensitive warning method on TikTok, for instance, received positive feedback from four participants because it directly addressed photosensitivity as a specific trigger.

> [on showing the Photosensitive Warning example on TikTok] *I have not seen the photosensitive warning one before. It's cool that they have that, and it's specific. I think that's very helpful. Very useful. (P10)*

Some participants like P01 and P04 *"would rather over-label than under-label" (P01)* when deciding to add warnings. P04 believed that it was better to add a warning if there was any doubt about whether the content might be triggering, as the worst-case scenario was that no one would be negatively affected.

> *Interviewer: When you are deciding whether or not this should have a trigger warning... Or whether or not your post should have a trigger warning? Is there any hesitation on some topics?*
>
> *P-04: Not really. I mean, I think it's just it's better to be safe than sorry. Right? So if you think it might, then just put it. I mean, worst-case scenario, nobody cares about it, right?*

However, P10 noted that sometimes content does not match the warning because the creator *"might put more tags than maybe apply just kind of to be safe."* P05 expressed a similar sentiment, stating that they were careful not to add warnings unnecessarily as doing so could diminish the impact of the content:

> *... However, I wouldn't air too much on the safe side just because there are certain topics that like, "Oh, this might be triggering", but actually reality isn't. It just makes the... Like degrades off the importance of the post itself. If you add a trigger warning, it shouldn't be there. So I would air on the safe side, but I would also take care not to add it unnecessarily. (P05)*

Further, P11 noted that while specific warnings are effective for certain triggers like photosensitivity, they may not be as suitable for more subjective and challenging-to-regulate triggers. In situations where brevity is crucial, acronyms like "TW" are commonly used, as mentioned by P07, who highlighted the need for quick communication within the constraints of short video formats: *"you have only like 15 seconds video, you have to put everything quickly".*

In summary, the specificity of warnings varies depending on the method employed, with some participants favoring detailed descriptors, while others appreciating complete obscurity. These variations make it challenging to determine the most suitable warning approach in different circumstances.

Authorship Difference

Different platforms exhibit distinct parties for assigning warnings, involving the platform itself, content creators, or the platform's community.

- **Platform-Driven:** For visual content, seven participants believed the platform automatically adds overlay and warning.

  > *I think it's automatic. And you can't add trigger warnings because I have not seen an option to do that when I'm posting. So I think it's added by the platform. (P02)*

  Further, P02 goes on to say that *"when it's image and it's graphic in nature, the platform definitely puts a warning on it."* In P07's experience, Facebook does not give creators the option to add a blur to visual content, they automatically do it. In the case of text-based posts, very few participants (2/15) thought that sometimes the platform adds warnings.

- **Poster-centric:** Some participants mentioned that warnings, especially on text-based posts, are primarily added by the content author and are not platform-generated. They emphasized that these warnings tend to be specific to the content and its potential triggers.

  > *But like for the most part it's user-added, it's definitely someone they type this in because it's very specific as to what the trigger is about so. (P01)*

  For images, the poster has the option to flag the tweet on Twitter containing the photo as sharp content. P09 believed that *"technology is that advanced to go through the video and decide for itself"* to add the blur, so the creator has to flag it.

- **Community-Contributed:** The community, comprising the audience and other users, also had a role in adding warnings. If many people report a post as offensive or sensitive, the platform's content moderation team reviews it and may hide it with a warning.

> *So, so for example, like as far as I can see how this feature works, like if many people report like, you know, offensive content or something, then Instagram will automatically hide it like they are just reporting it. (P13)*

In addition, P04 stated that unless viewers report the Instagram content/video, the platform does not add any warning. Participants felt that if the poster does not put a TW, even after being prompted when making the post, the community feedback is important to keep them accountable and explain why they need to add a warning.

> *And for platforms like Reddit, which is a community-based thing, you'll see people rally behind it, and then it might be like... It can get to a point sometimes where people who don't put those on their posts are kind of like called out for it, and then that would like (force) more people to put it in. (P05)*

Further, on community-based social media like Reddit, based on how well the community is moderated, sometimes the moderators go in afterward to add TW/CW for the poster. Folks like P10 would also accept audience feedback if they missed out on putting a warning where it's needed:

> There's probably instances where I should have because I'm not the one that determines really if it needs a warning. It's more so you kind of have to guess and. I am certainly not going to be the type where if someone's like,

> "Hey, maybe add a warning on that next time or something like that", I'm not going to challenge them. (P10)

The only problem, in this case, is that before any action is taken by the platform on the feedback, *"before someone marked it or again, like 1000, 2000 people have already watched it."(P13)*

In summary, the responsibility for applying warnings to sensitive content fluctuates depending on both the nature of the content and the specific features of the platform. These variances in authorship roles further contribute to the logistical complexities associated with implementing warnings.

Language Difference

The language in which warnings are presented can vary depending on the audience and the context in which they are used. P15 said that when platforms add warnings, like videos on Facebook, they tend to use more formal language.

> *I noticed the videos that I watch on Facebook like those are probably more supposed to be for like the general public or for like educational purposes. So those would be like more official in a sense. Like "This video contains things that make them might make you feel uncomfortable, please be advised". (P15)*

In contrast, they found when users share personal stories or content on specific topics, such as abortion rights, the language used in warnings tends to be more informal. In some cases, content creators may include disclaimers in their content but not explicitly label them as "Trigger Warnings" or "Content Warnings." For example, P01 mentioned that in the text, before talking about the content, they put: *"Hey, this is what's here beforehand".*

The way TW/CW are expressed could also vary depending on the culture and language in which the content is presented. For example, P15 noted that in simplified Chinese, individuals may use a more advisory tone, informing viewers that the content might be disturbing while emphasizing the importance of respecting others' perspectives.

> *I mean, it's more like an advisory kind of thing, like "just so you know, like this is this might be disturbing", but then he's also depending on the culture. ... It's more like "Just so you know, you might not feel comfortable about this, but it's okay. We have to be respectful to what other people have to offer". Yeah. (P15)*

In short, warning language varies with audience, context, and culture, from formal to informal expressions, and might include subtle disclaimers rather than explicit labels like "Trigger Warnings" or "Content Warnings.", adding to the complexity of their use.

4.1.3 Tension of TW/CW with Engagement

Participants had varying opinions on how adding TW/CW affected engagement on a post. Most believed that adding a warning reduced views, while few felt that it either increased engagement or had no significant impact.

Adding A Warning Reduces Views:

Creators may not add warnings since it limits or negatively impacts how widespread their content will become. For instance, P02 suspected that Instagram's algorithm may rank sensitive content lower, resulting in reduced visibility, even for users they follow who share such content:

> *Yeah, and I don't think I've seen trigger warnings in the past two months on*

> *Instagram. And I mean, I do follow some pages that might post like triggering stuff, but I think they're algorithmically trying to reduce reach. (P02)*

Due to the belief that adding warnings may lead to reduced engagement, creators might opt not to include warnings in their content, as further explained by P02:

> *P-02: Yeah. Ah, but, that will have an unintended consequence that people will stop putting trigger warnings on their content because they'll get less views. So yeah, I mean.*
>
> *Interviewer: Yeah. That's why when you mentioned about like algorithm sort of like ranking that content lower, then I was like maybe people would just stop adding trigger warnings.*
>
> *P-02: Yeah, that's the actually direct consequence of it. That people will stop adding... Because...so of course there's like a new section of people that's, you know, posting genuinely but most pages are posting such content for shock value and effectively net engagement. Not all but a majority. So I think if they are forced to add trigger warnings, and the algorithm downgrades them, they will outright stop having them.*

Looking at the scenario from the perspective of a corporate social media account aiming for maximum views, another participant (P12) noted that they might avoid adding TW/CW because it could potentially reduce the number of views.

> *But if I'm in a corporate company where I want to have the most views possible, in their perspective, they won't like it because they will give less views if they have the trigger warning. (P12)*

Similarly, on YouTube, content creators often steer clear of certain words or content, including TW/CW, to avoid strikes or demonetization as mentioned by P07:

> *Interviewer: You think putting a warning affects views?*
>
> *P-07: Definitely. This is why I think it is difficult. I know for some, a lot of YouTube content that I watch, they try to avoid saying some specific words or putting specific content because they get strikes.*
>
> *Interviewer: Demonetized.*
>
> *P-07: Yeah, they will get demonetized. Some strikes. So.*

This situation can inadvertently give mainstream groups a stronger voice on social media compared to minority groups. Majority group posts tend to be more widely shared because they do not typically include TW/CW, while content from minority groups that do use TW/CW may have a limited reach. P03 provided an example from India, where the dominant Hindu culture often overshadows minority Muslim experiences due to differing use of TW/CW:

> *Like, again, coming back to Indian politics, I feel like the Hindutva politics is so out there that...And because you always… everything that happens to a Muslim person is under trigger warning. So those things don't go out there as much. But oh, my God, one thing happened to one Hindu man from a Muslim person, and you will see the rage. Because they are not careful as to giving trigger warnings. And then it doesn't... it automatically gain the religious momentum and not just as an individual case of violence, it gains a religious momentum and its widespread. And then it becomes a way to gather and garner support for the invalidation for the already messed up politics that we have. (P03)*

However, P11 emphasized the importance of TW/CW, acknowledging that while they might affect views, they are essential for protecting viewers from potentially distressing content.

Adding A Warning Increases Or Does Not Affect Engagement:

Some participants noted that creators might add warnings to encourage engagement with their posts. P01 shared an example where they observed spam bots using TW/CW as a strategy to grab attention, particularly on platforms like Tumblr and Reddit. These spam bots would employ specific warning tags to ensure their content appeared for users following those tags.

> *P-01: Well, it's using warnings like this is specifically when say like Tumblr, Reddit, or whatever. It's something where like if you follow or pay attention to a specific tag, then it'll pop up for you. So spam bots will use the specific tags so that you see their content.*
>
> *Interviewer: Oh! So they are they're basically making use of these warnings, I guess, to just...*
>
> *P-01: Yeah, it's really weird. I'm like, that's a CW tag. Why? Why is that? I don't want you to sell me sunglasses. What are you doing?*

Additionally, P12 mentioned that when users choose to open content with a warning, their satisfaction rate tends to be higher because they made an informed choice to view it. P10 further believed that warnings did not have a negative impact on awareness campaigns on platforms like Twitter. They emphasized that the ability to make multiple tweets allowed for effective information dissemination, even if some users chose not to engage with every tweet.

> *Um. At least on Twitter, like I said, which I use the most. I think the. Um. I don't think. I've... I see it ever negatively impacting like an awareness campaign. At least, that might just be because, you know, you can make more than one tweet or something like that. But. I think, even when reaching like... Even without engaging or reading the tweet. I think that. Enough people do. (P10)*

This dynamic suggests that there is tension between warnings and their influence on user engagement. This tension further complicates the use of warnings, with respect to motivations of content creators.

4.2 RQ2: How do people decide to view social media content with TW/CW?

Once a TW/CW is presented, the viewer gets to decide whether they want to engage with the content or skip it. As one participant aptly put it:

> *I feel like a big part of the trigger warning is to give the audience the choice to read something. (P04)*

This decision to view or skip content with TW/CW depends on internal factors, which are related to the viewer, and external factors, which are related to the warning and the content posted.

4.2.1 External Factors

Modality Of Content

Participants sometimes based their decision to view or skip content with warnings on the modality of the content. For example, P10 mentioned that they found it easier to process text-based content with TW/CW, and they usually engaged with such content without issues. However, they tend to skip videos with TW/CW, as these were more challenging for them to handle.

> *Whenever the content warning is on like some block of text or someone describing their experience. I've never had any issues with that. I feel like I'm always capable of processing that, dealing with it, engaging with it, and things like that. I think the same goes for pictures as well most of the time. Usually just because I think*

> *the pictures that are attached aren't ever too graphic, at least personally, that I've seen. But definitely, the ones where I do engage, and I kind of maybe hinted at this earlier, is with the videos and the instances where like, I'll definitely skip. (P10)*

Furthermore, some participants (4/15) indicated that when they come across content with a TW/CW on an image or video, they read the accompanying text-based information, such as captions, before deciding whether to view the visual content. This approach allows them to gauge the content's nature and relevance and make an informed decision about whether to engage with it or skip it.

> *And then I, you know, I think it really depends on the format. I am more likely to read something with a trigger than I am to see something with a trigger warning. Especially because reading, I guess you can stop at any time and watching is a little bit harder because like, you can close your eyes and you still have audio, right? (P04)*

> *So to be honest, I mostly skip those parts because I cannot handle those videos. I usually read what's written in the caption and get the gist of what's in the video. (P09)*

> *But if I'm scrolling down, not in the story, but as a post, then reading the caption might help a bit, and then I might open it. (P13)*

In summary, participants' responses indicated that content modality influenced their decision to engage with material featuring TW/CW. Text-based content with warnings was generally less triggering and often engaged with, while videos and images with TW/CW were often skipped.

Poster Of The Content

Additionally, the decision to view content with TW/CW is sometimes influenced by the identity of the author or poster. P15 mentioned that they found it easier to accept warnings when they came from friends in the Facebook groups they've joined, as they had a better understanding of what to expect in such cases.

Moreover, viewers may have an idea of the content behind the blur based on who posted it. This familiarity with the type of content that poster typically shares allowed them to make informed decisions about whether to skip or view the content, depending on their comfort level. For instance, P04 shared that they had friends who regularly posted surgical images from their medical school experiences. When they encountered a blurred image with a warning from one of these friends, they anticipated graphic surgical content and chose to skip it if they found it uncomfortable.

> *There's...I have maybe like three or four friends back home who went to medical school and for some reason like to post pictures of people being opened up, you know, in the surgical room. So I know that if it's one of them and the images blurred out and it's a content warning, I'm like, oh, I'm about to see some intestines out of somebody and I hate that. So I skip immediately. (P04)*

Similarly, P10 indicated that their decision to engage with content featuring TW/CW is often influenced by their confidence in the poster or author. They tend to engage with such content when it is shared by people they follow and trust, as they have an expectation of considerate content. However, when encountering content from random sources, they are more hesitant to engage due to a lack of familiarity and trust. This distinction highlights the role of trust and familiarity with content creators in influencing the decision to view or skip content with warnings.

In summary, the identity of the author or poster can significantly impact users' decisions regarding content with TW/CW. Familiarity with the author's typical content and the level of trust in their postings play a crucial role in determining whether viewers choose to engage with or skip content with warnings.

Specificity Or The Context Given In The Warning

Participants also based their decision to view or skip content with TW/CW on the amount of context or description provided in the warning. The majority of the participants (12/15) said that not giving the necessary information can hurt the viewers' autonomy and decision-making. They then do not know if the content of the post intersects with their personal triggers.

> *But sometimes when I don't know what to expect when I am not given enough information to make an informed decision, those really come as a surprise. Yeah. (P15)*

They preferred warnings that included sufficient information to assess the potential impact of the content on them. For instance, if a TW/CW offered specific details, such as "flashing lights," P04 knew this would not affect them so they could confidently engage with the content. However, if the warning was vague or lacked context, they were more likely to skip or scroll past it.

> *Yeah. So usually first thing would be to read the trigger warning and see if it has enough description that I can gauge if it was going to affect me or not. Like, for example, flashing lights don't affect me, right, Because I really don't*

> *have epilepsy. So if it's something like trigger warning, flashing lights, I'm like, "okay, this is fine". If it's something vague, then I mostly skip it. (P04)*

Similarly, P12 explained how having specific descriptors with the warning can help:

> *It might be better to have it [specific warnings] because, if some people have trigger to the sexual stuff, then it's better to know if it's sexual stuff [behind the warning]. Because if it's for example, racial stuff rather than sexual stuff, it's fine for them. (P12)*

Moreover, P13, who did not experience strong triggers, mentioned that they would view content with TW/CW if there was additional background information provided and the topic interested them. This contextual information influenced their decision to engage with the content or skip it.

> *I don't feel much impact as such, but sometimes it's better to have a warning like "some graphical content is there". So if that is there and if I wish to see, like if it's what we say, not an interesting topic, I would say, but a topic that I should read about or something. Then I will definitely view the content or the graphical image, or else I will just scroll it. (P13)*

General or less context warning is better in some cases: While detailed context is crucial in warnings, there are instances where general or less specific warnings can be effective. In circumstances where it gets *"too complicated"* to decide what warning label to add, P13 suggested that it is better to give a general warning.

From the viewer's perspective, three participants mentioned that on visuals they prefer completely covered warnings with no context of the author, comments, or colors of the photo visible.

> *For the previous question, I just thought about another frustration. So, you know, in some contexts, sorry they blur the picture. Yeah, they blur the picture. I don't know if it's to me personally, it makes me more curious of what's underlying like behind the blurred picture. You know, if it's like skin color, maybe it's a nude; if it's a red color, maybe it's blood. So, like, I feel for some people, like even with that blurred image, they can get triggers. So I think it would be just not to have the blurred image at all. (P12)*

> *And also, I think it will be helpful that we cannot see who post it, because some people there may be curious or just try to find it. (P07)*

In summary, while comprehensive warnings with specific context are generally preferred, there are situations where simplicity and minimal context can enhance the user experience.

More context or specificity is better: Most participants had the opinion that more context or specificity in the warning is useful to make a decision. P14 pointed out that specific warnings are preferred because they help prevent anticipatory anxiety and curiosity that can arise from vague warnings. Knowing what to expect is seen as less distressing than having no information.

> *I think having a non-general warning or like a specific warning is better because just having a general warning is like. Uh, like, we don't know what's going on there, so we don't know if it's something we may or may not have experienced. I think having like an experience of that, like might bring back the trauma. So I think, like, just the curiosity of it is worse than just having it. (P14)*

If the warning is too general or has no context, viewers can give in to their curiosity to see the content. For instance, P12 found themselves more intrigued by what might be concealed

behind the warning, whether it's sensitive images or topics, and this curiosity led them to watch the content.

> *I don't know if it's to me personally, it makes me more curious of what's underlying like behind the blurred picture. You know, if it's like skin color, maybe it's a nude; if it's a red color, maybe it's blood. (P12)*

This increases the risk of encountering triggering content and defeats the purpose of having a warning. Viewers can end up regretting it later, getting affected more than they thought, or feeling betrayed by the author. The absence of severity indicators in warnings, especially when new features related to warnings are introduced, can also lead to a false sense of security (P03). As an example, P10 expressed concerns that overly vague warnings, like the TikTok general warning in the example slides, could lead viewers to click "Watch anyway" out of curiosity, defeating the purpose of the warning.

> *Yeah, I think that might be a little too vague because I think it's kind of what I'm saying with the Discord thing is like, you don't know and you're curious, so you probably are just going to click "Watch anyway" to find out and then skip it, which kind of defeats the point of warning, I guess, because you've already engaged with it. (P10)*

P15 provided a concrete example of encountering a blurred image on Quora without any context, which led to curiosity and a potentially shocking experience. This participant believed that vague or context-less warnings are not effective and could encourage viewers to explore further.

> *To be honest, I didn't know that Quora would just make a blurry picture. Like an inappropriate picture blurry because I didn't know what I was looking at. I*

> *just click on that I was like dang am I looking at, like, pornhub or something. Why didn't Quora like tell me that? It should at least give you like, a piece of information what is it about. You know what I mean. I was really surprised when I, was like, Is it just my internet not good? So the picture is blurry, so I double-tap on it. So I. Because I wanted to see it, right? It's like how human beings are. The more that you don't, you're taught to not do something, the more that you want to do it. And I was like, Dang, I wish I didn't do that. So I do believe, like the blurry picture without any information on it, it's not effective at all. I think it's going to encourage people to want to see more. (P15)*

Moreover, P07 mentioned that they would want specific categories of warning given for abuse because they get triggered by academic abuse but want to watch content on child abuse because they want to educate themselves on it.

Too much context can be triggering: If there are more than necessary specifics given in a warning, then that can trigger the viewer (3/15). P04 mentioned the same when talking about the example from the slides:

> *Like, for example, the middle one, the one of the book is like "it's a true story focusing on sadistic torture and abuse". Like if you go too specific with your trigger warning, you're going to trigger somebody. (P04)*

On viewing the first-slide method-based example of "TW: Gangrape", all of the above participants felt uncomfortable about the word. P13 said that the word gangrape is *"too specific"* and *"reading the word itself might be a trigger for someone"*. P14 suggested that it would be more effective to use a word like 'sexual assault' instead of the exact wording of the crime committed.

In summary, while most participants preferred TW/CW with more context or specificity, this subgroup of participants emphasized that there is a limit to the level of detail that should be included, as excessively specific warnings can themselves be distressing and triggering for viewers.

Giving a reason for the warning helps in decision-making: In terms of specifics, participants agreed that giving a reason for the warning makes the warning effective. For example, explaining why the platform-added photosensitive warnings on TikTok made it effective:

> *[the photosensitive one is effective] because it actually talks about what might be... Like light, like for example, I think that one's good. It actually discusses what the actual trigger for the warning might be versus a general trigger warning. (P14)*

P15 recommended concealing visual content behind a warning, accompanied by explanatory text outlining why the content could be triggering helps in decision-making as a viewer.

> *I think if there's like black in the background so you don't really see anything and then it's just text on it that says that this is, you know, like this might be disturbing because of the following reasons. And then do you still want to proceed or just to skip it? I think those would be like the most effective one. (P15)*

In essence, participants believed that providing a clear reason or explanation for the warning, either through text or other means, contributes to the effectiveness of TW/CW, as it empowers viewers to make informed choices about their content consumption.

Overall, participants' preferences as viewers regarding the context of TW/CW revealed a nuanced tension. While many emphasized the importance of detailed warnings for informed

decision-making, others cautioned against excessive specificity, as it could itself be triggering. Striking a balance between providing sufficient context and avoiding overly explicit details seems crucial, with some participants agreeing that clear explanations for the warning's purpose enhance the effectiveness of TW/CW for the viewer.

Optimal Warning Length and Clarity

The decision to view the content behind the TW/CW also depended on its length and clarity. P14, for example, found concise warnings to be highly effective as a viewer, such as "TW: abuse," as they provided a quick and clear indication of the content's nature. They emphasized that shorter warnings make it easier for readers to decide whether to engage with the content or not.

However, some participants (4/15) highlighted the potential for ambiguity when using abbreviations or acronyms in warnings.

> *P-07: I forgot again, I didn't say it, but what the TW stands for again? What warning?*
>
> *Interviewer: Trigger warning.*

P04 and P15 mentioned that abbreviations like "TW" may not be universally understood, especially by non-native English speakers or those unfamiliar with internet culture. They argued that providing clear, unabbreviated warnings, such as "trigger warning," can help avoid confusion and make the warnings more accessible.

> *P-04: Sometimes I don't love TW like I like trigger warning better because if you don't know what TW is?*
>
> *Interviewer: That is true. Yeah.*

> *P-04: Especially people like... I'm not a native speaker of English. Right? So the first time I saw TW, I'm like, "I don't know what this means". TW abuse. I'm like... (confused face) So I would spell out trigger warning*

Moreover, P07 noted that simply using the term "disclaimer" in the caption is not an effective way to warn about sensitive content. This is because *"disclaimer can be just anything that doesn't necessarily mean content warning or trigger warning."*

Enhancing warning effectiveness with visual separation: Two participants (3/15) emphasized the importance of visually separating the warning from the sensitive content to improve its effectiveness. For text-based posts, having spaces or dashes between the warning and the main content would make it easier for viewers to discern where the warning ends and the actual content begins (P11).

> *I would have put more space between the warning and the text personally, just because, I mean, people have wandering eyes. Like immediately you can shoot down to the next line. (P04)*

Similarly, P10 pointed out that on platforms like Twitter, where creators mark images as sensitive content, the absence of visual separation between the warning and the image renders the warning ineffective for the viewer to discern.

> *Also on Twitter. I think it is up to the creator to mark pictures. As like content warnings or sensitive content. So even though if a creator might put a content warning in the tweet, the image still just shows up regardless. So same thing as the text one, it's not really effective. I think that's the like with how they're used I think is a frustration is just because. It. People are trying to use them, but it ends up not being effective. Which kind of sucks. (P10)*

They further mentioned that using tags, that are visually distinct from the main text, is *"definitely more noticeable"* and consequently more effective.

In short, making warnings clear, visually separate from the content, and keeping them brief can aid viewers in deciding whether to engage with the content or avoid it.

Visibility Of Warnings

Participants found warnings effective in decision-making to view content if the warning was visible and they were aware of the warning. Including the warning in the title of the post than the description (P04, P05), having a font big enough (P12) so that viewers can instantly see it, and having an emoji next to the warning (P13) were some of the ways participants mentioned to grab viewers' attention to the warning on text.

Visual warnings for visual content: For visual content, participants (3/15) did not find the warning in the caption helpful in making a decision since the viewer's attention was more towards the visual content than the warning in the caption text. In the Instagram example shared in the text-based slide, P10 found a problem with how the warnings were just in the caption:

> *But I can already see the immediate issue with on Instagram because it puts the picture first and the text is very small. You're going to read the image before you even have time to look at the description. (P10)*

Likewise, five participants found visual warnings to be a better way to warn viewers than just having a warning in the caption because it *"alerts me more than text" (P12).*

> *If you're on a reel or a video, it's usually the person who uploaded it would have to put like text. So that would be somewhere in the video or somewhere in the*

caption of the video. But at that point, you're already watching the video. So I think it's a little bit pointless when they do that (P04)

Putting it in the caption, I don't think it adds that much value because anyways have seen the picture. So I think it's a little too late. (P02)

P14 also found that when someone reposts visual content on an Instagram story, the caption is hidden by default so the caption-based warning is not visible to the viewer.

Accessibility considerations: Further, two participants mentioned that for videos, adding text-based and spoken warnings could help people with disabilities (who cannot see, cannot hear, or cannot read) to be conscious of the warning.

P-07: And I really appreciate those YouTubers who edit this stuff, and they put both, you know, like they're trying to warn a lot for like a whole minute talking, just to warn. Um, visually. So they post something like this and also they say it out loud. So I guess, because a lot of people on YouTube also, they cannot, they cannot see or they cannot hear. So the only here, you know, so I think that will help for those people with disabilities who watch.

Interviewer: Yeah. And I guess also people who cannot like if there are people who cannot read...

P-07: Yes. Illiterate people, they only can listen. So that that helps.

Interviewer: Yeah, that's a good point. Yeah.

Accidental viewing: Accidental exposure to sensitive content can occur when viewers are not mindful. P12 mentioned that the blurred method works better than the first-slide method because people *"have a habit of automatically swiping"* without reading the text

and can be exposed to triggering content. Similarly, P07 found the pop-up warning on the YouTube mobile application more effective than a browser because it is in the center of the screen and *"you are forced to actually read it"* before clicking to view. Hence, having a user experience where the viewer is forced to pay attention to the warning is effective.

In summary, participants stressed the need for prominent and easily noticeable warnings to enhance their effectiveness as a viewer, especially in a fast-scrolling online environment.

Environment Settings (Public or Private)

Three participants emphasized the importance of warnings in public settings for example on public transit (P10), highlighting the need for informed decisions when dealing with triggering content. As a consumer, they noted that such content can impact not only them but also those in their vicinity. For instance, P07 narrated an incident where graphic content with disturbing audio affected their roommate, prompting the roommate to request warnings for such content in the future. P10 also illustrated a scenario on public transit:

> *One situation that, you know, that I think about sometimes is just like even if you're just like on the bus or next to somebody or somewhere in public. Right. And you're on your phone, and without warning, something pops up. You know, it could affect more than just you. (P10)*

Furthermore, P10 expressed their gratitude for warnings, even when they chose to view the content despite potential triggers. They explained that they prefer not to open such posts publicly. Those posts might affect them, so they would instead open them in the privacy of their home.

In summary, viewers decided to see the content behind the warnings also based on their

physical environment.

4.2.2 Internal Factors

Viewer's Headspace

Some participants revealed that their decision to engage with potentially triggering content depended on their current mental state and overall mood. For instance, P07 mentioned that they consider factors like whether they've taken their antidepressant medication or their general emotional state that day. When feeling stable and open to learning, they might choose to watch such content for educational purposes. However, during periods of vulnerability or low mood, they opt to avoid it.

> *Yeah, it depends on my situation. Let's say today I forgot to take my antidepressant medicine. It's been two days and I'm like, "Oh no, I shouldn't watch those kind of stuff". But sometimes I'm very chill and leveled. Let's just watch that to educate myself. So it really depends. (P07)*

P15 further explained that it also depends on their brain power to process and unpack warnings to determine whether the content will affect them or not:

> *And then it's also depending on the headspace and the mood. Like, if I'm already sad, like, what is the point of reading more? Consuming more negative information. And also, the mood like say it's that if I just really tired and just on social media to like to kill time before I go to bed, I might not have the power, the brain power to actually process and unpack the trigger warnings. I might still just, I might just let it put. I might just keep reading, and keep scrolling. Those*

> *probably wouldn't affect me as much because I'm tired already. But then, like, if I spend, like I'm really focused. Like, for example, I woke up refreshed in the morning, wake up refreshed in the morning, trying to get some work done. But before then, I decided to look at my phone. I saw something like that that I would intentionally choose not to deal with, not to read it and skip the content because I'm refreshed. I know what I'm doing, and I know that I wouldn't let that take over my attention and my capacity to process more information. (P15)*

To conclude, some participants' decision to engage with potentially triggering content behind the warning was influenced by their current headspace.

Viewer's Personal Relation with Topic Behind Warning

In certain instances, participants opted to view content flagged with warnings, driven by their desire to stay informed about relevant topics given in the descriptor of the warning labels. Whether it was a local event, like an incident in their hometown (P02), or a campus-related matter, such as a sexual assault case (P15), they chose to engage with such content to understand the situation better and support relevant causes. One participant (P07) also watched sensitive content to gain knowledge on a specific issue like child abuse demonstrating their commitment to learning about important societal matters. Furthermore, individuals like P12 engaged with content marked with TW/CW as a means to stay aware of global events, even when the subject matter is distressing or unpleasant.

Whereas certain potentially triggering topics with TW/CW, to which the participants had a direct and sensitive connection, were avoided at all costs. For example, a participant who has struggled with suicidal tendencies chooses to steer clear of content related to suicide as part of their ongoing recovery journey (P07). However, they were more open to viewing

content on sensitive subjects that they cannot personally relate to or have not experienced (P07). Similarly for P15, topics like gender-based violence were deemed off-limits due to their emotionally taxing nature.

> *Again, there are topics that I would really rather not touch, like gender-based violence, straight-out domestic violence, and body scene like no go, I just can't deal with it. (P15)*

In contrast, P10 skipped warnings when the topic did not resonate with their own experiences. This perspective suggests that TW/CW is primarily intended for individuals who may have direct personal connections to the content (P14).

> *A lot of the times if. The topic isn't something that's relevant or. Like applicable to me is when I don't really have an issue not engaging, I guess. (P10)*
>
> *The trigger warnings seem like they're more for people who might have had have experienced this. So when it comes to warnings that certain things that I might not have experienced, then I'm more inclined to ignore the warning. (P14)*

Overall, participants' choices to engage with content featuring TW/CW partly depended on their personal connection to the topic. If they wanted to learn more about the topic, they would ignore the warning even if it could affect them. Certain topics were avoided at all costs due to personal history with them. Whereas some participants tended to skip or ignore warnings when the content did not directly apply to their experiences.

Viewer's Tolerance (high or low)

Participants' decision to view the content behind the warning was influenced by their personal ability to tolerate sensitive material. A subset of participants (4/15) indicated that

they tend to view content with TW/CW without getting easily triggered. For instance, one participant (P12) mentioned that they have a high tolerance for such content and rarely get triggered. As a result, they are inclined to watch content with TW/CW anyway, simply because they can handle it without adverse emotional reactions.

> *It was a generic warning, like trigger warning, "Do you really want to watch?" And I press Yes, because I know. I think I have maybe I personally have a high tolerance in those kind of things. I don't get triggered that easily. (P12)*

Moreover, P12 admitted to enjoying violent content in cartoons or prank videos, acknowledging that while these videos might not affect them, they could potentially trigger others. This participant appears to have a robust emotional resilience to triggering content, referring to their tolerance as having a *"brain of steel."*

Similarly, some participants, like P11, mentioned that they do not experience triggers often but still appreciate the inclusion of TW/CW for the comfort and well-being of others.

> *Interviewer: Yeah. Um. And so, I guess, do you want trigger warnings or content warnings as a viewer? And why or why not? I don't...*
>
> *P-11: I really appreciate when they're added. You know, even if they weren't. I wouldn't really be too bothered. But for me, it's more important that like. Other people are comfortable with it.*

Other participants decided to skip viewing the sensitive content, or at least save it for later, because they had a low tolerance for that material and it could have a detrimental effect on their mental well-being. For instance, P02 made a conscious choice to skip content with TW/CW because they believed that consuming such material would have an adverse effect on their mental health.

> *I mostly skip. I don't like to. I mostly skip all that content because I feel like consuming that content has a negative effect on me. I actually avoid it. So maybe that's why I don't follow stuff that show trigger warnings and content warnings a lot. Maybe that's the reason. But I definitely skip it. (P02)*

Similarly, P04 saw little value in exposing themselves to content behind TW/CW because they did not want to test their tolerance.

> *P-04: And just I mean, I'm a curious person, but I'm not curious enough to say like, "oh, let's take a gamble and see what this trigger learning is about". Like, I just won't.*
>
> *Interviewer: Right. You won't take that risk?*
>
> *P-04: No. I mean, because the next thing you know is like, I can't sleep for three days, and I'm like, I don't want that! (laughs).*

Strategies to avoid triggering content: Where warnings did not work or were not present, participants employed various strategies on their end to avoid encountering triggering content on social media. One common approach was muting or blocking users who consistently posted such content. However, P02 acknowledged the downside of this method, as it could lead to missing out on personal updates from the muted users.

> *If there's a person who's posting a lot of political stuff and just like throwing content at my face and that person's a friend, I just mute that person. I think muting is a very good feature. But like the problem with that is, you know, people have phases when they do this and I don't go back and unmute them. So it kind of like breaks the connection. You know, like I don't get updates from them anymore. So like I think it was yesterday that I realized that, "oh, I muted this person and*

> *I never got their update. They got engaged and I never got that update." So yeah like I think it makes... I think there can be something better. (P02)*

Similarly, P10 used a combination of blocking and muting words on Twitter to manage their exposure to triggering content. They also employed a unique workaround by utilizing the "data saver" feature on Twitter. Originally designed to conserve data usage, this feature prevented Twitter videos from automatically playing on the timeline, offering an extra layer of protection against potentially distressing video content.

> *P-10: ...As I learned about this feature and this is on my end personally, and it is a feature that prevents Twitter videos from automatically playing when they're on your timeline. And I think it's called "data saver" because it waits to download the video until you press on it. But if there's any triggering content in the video that, you know, I guess this is a very general, you can't do it for specific things. It'll just any time there's a video, it won't play it till you click on it. Which I think I have enabled currently which...*
>
> *Interviewer: No, I guess the main purpose of it was not to prevent videos, it was just to save data. But it's super useful for this context as well.*
>
> *P-10: Yeah. And I think I personally turn that on.*

Some participants had a strategy of initially viewing content even after encountering a warning, but they would quickly skip it if their tolerance to such content was tested. For example, one participant (P13) mentioned scrolling through such content within the first few seconds if it became overwhelming. Another participant (P04) described a selective approach where they read the end of the caption, and if the content showed personal growth or a positive outcome, they were willing to watch it; otherwise, they would skip it. Similarly, P11 indi-

cated that while they aimed to stay informed, there were moments when the overwhelming volume of sensitive content prompted them to disengage temporarily.

> *Like, again, I kind of just want to be as aware of what's happening as possible. But sometimes it gets to be too much, and that's when I skip. (P11)*

In summary, participants' decisions to engage with content flagged with TW/CW were shaped by their personal tolerance levels for sensitive material. Some had a high tolerance and would view such content without distress, while others with a lower tolerance opted to skip or avoid it to safeguard their mental well-being. To mitigate exposure to triggering content, participants used strategies like muting or blocking users, employing content filters, or quickly bypassing content that became overwhelming.

4.3 RQ3: How can the platform improve users' experience with TW/CW?

4.3.1 Beforemath: Education of Triggers and TW/CW

Most participants (10/15) mentioned that sections of people on social media do not know what a TW/CW is or how to use them correctly. Some (3/15) also acknowledged that not everybody is aware of why there is a trigger warning, and *"not everybody is knowledgeable enough about the significance of it (P15)"*. When P07 used TW/CW on their content, some of their audience perceived it as *"annoying"* and believed that there was *"nothing wrong"* with sharing sensitive content without any warning. P15 mentioned that their folks from mainland China have less awareness of warnings. P07 further gave us the example of friends from their home country posting sensitive content openly because they do not know what a warning is:

> *So for those people who don't really put trigger warnings, especially the ones from back home, they will just share on this case "Someone committed suicide" and then they don't really say it. I mean, they don't know what content warning is. (P07)*

One of our participants, P12, even signed up for the study because they wanted to learn more about TW/CW and they were *"not even sure where to get an education for this kind of stuff."*

> *Interviewer: So, like, why? Why were you curious about signing up for this study?*

P-12: That's a good question. I think I was. (laughs) I don't know. I mean, I see them a lot. I mean, I've used like, social media for a while, but recently I've seen them more. And while more and more while. So I was curious about, like, what this really is. I mean, even still right now, I mean, I don't know that much about trigger warnings.

Ways to Educate Around TW/CW Usage on Social Media:

Participants mentioned that the lack of awareness and education around warnings can be improved by having discourses on how to use them effectively.

I think education on it and like, why and how to use it is important. ... So that yeah, I guess education would definitely be a pretty important component to being able to use this. You need to know when and how to use it. (P08)

P-07: ...I think this thing should be taught to everyone, to kids, especially. Everyone. Teenagers and old people. Because there's people back home who send crazy stuff are usually old people who just have their phones and just have their internet. Um. Yeah, I think it would be nice if they have some sort of. Not curriculum, what's that? Like you said.

Interviewer: Guidelines.

P-07: Guidelines. And actually be taught to everyone. So. Because who is not in social media these days? No one. So. So etiquette for social media I think. Trigger warning.

These could include teaching about warnings in the school curriculum (P03), guidelines (P07), platform-led suggestions (P02, P09), and training (P15).

Platform Guidelines: Participants were asked about having platform guidelines around warnings. According to P12, there should be a general basic guideline of what to avoid, but *"people should be able to feel free to add more."* P07 wished for guidelines on approaching people in their circle to nudge them to add a warning.

> *If they are your parents, this is what you should do. If they're your kids, what you should do. If they're your colleagues at work, this is what you should do. It's it's really nice to learn, actually. And I don't know to be honest. So do you do you? I don't know, call them out in public, in the group, or you actually talk with them privately? Stuff like that. So. Because it's it's it's it's not a. It's not a secret any more. People actually commit suicide just because of these triggers, random triggers on social media. And I was one of them when I was having my those years. (P07)*

However, there were a couple of concerns and challenges raised around having guidelines:

- **Hard to have universal or platform-wide guidelines:** Developing universal or platform-wide guidelines for adding warnings proves to be a complex endeavor, as highlighted by six participants. They emphasized that creating such guidelines can be problematic due to the intricate nuances of triggers and the diverse ways individuals interpret and respond to warnings. For example, P15 mentioned when Facebook was trying to implement a universal policy about nudity, there was outrage because they wrongfully labeled some really radical feminist artists' performances as violating guidelines. P15 further goes on to say:

  > *You know, like, actually, it's kind of hard to implement like a universal platform-wide policy given how dynamic and diverse human beings' experiences are. I do believe that the choice should be in the user's hands. (P15)*

Secondly, two participants expressed that *"there can be universal guidelines for certain triggers but it will not be for every one (P09)."* P11 also mentioned that having general guidelines is enough as specific topics will be harder to regulate:

> *It would be good for like a general statement or general like. Guidelines kind of thing. But again, there's other like smaller topics that are like. It might be hard to regulate for them. (P11)*

P08 also noted that establishing universal guidelines is complicated due to variations in how different platforms are constructed.

> *I mean, they're all so different. Like on Reddit, you can go from, like, places where it's okay to post, like, literally anything to places that are like, super, super heavily moderated. (P08)*

- **Hard to capture people's attention:** Educational efforts face a significant challenge in capturing users' attention. Giving guidelines might not work because it is hard to focus the user's attention on reading the guidelines. Two participants mentioned that educating on a new platform feature roll-out will not be effective because it will be like the terms and conditions of an application, and *"no one reads it" (P09).* P13 further emphasized that users tend to rely on their instincts and the familiar experiences they have encountered in other applications, rather than delving into provided guidelines. This tendency leads users to follow established usage patterns rather than actively seeking out educational materials.

Training: To help users better understand the concept of TW/CW, P15 drew a parallel with Robinhood, a stock-trading app that provides mini-training to educate users about wise investment practices. Similarly, P15 proposed that platforms could offer training to inform

users about the purpose and usage of TW/CW. P08 suggested that the platform should educate users on how to add warnings when a new user onboards and makes their first post:

> *Yeah, I mean. I think so. I think it would be like, let's say I create an account on Facebook or whatever and I go to make my first post. It might have like a short little thing like, Hey, this is a tool. You can use it, you should use it. Here's why you might want to use it and then let them be. (P08)*

Platform-led suggestions: P02 thought that post-engagement metrics should be used by the platform to see if the post is triggering to people and then send notifications to the author on *how* to add a warning *using their tools*.

> *I think maybe the platform itself can step in in the education part of it and if it notices that... So the platform should be able to notice that someone's posting triggering stuff because that will spike engagement or there are metrics for it. If they're not putting like trigger warnings and stuff and people are getting triggered, if they can measure that, then I think this should be able if the person's not doing it, then they can maybe send a notification to that person, "Hey, these are the tools available to you to add those trigger warnings and stuff." That would be a good idea. (P02)*

P09 highlighted that platforms, like Instagram or Facebook have a history of promoting new features through stories and other in-platform advertising methods. P09 suggested employing a similar strategy to raise awareness about features to add TW/CW. This approach would facilitate user education without incurring significant costs for the platforms.

In conclusion, participants emphasized the need for comprehensive education and awareness campaigns regarding the effective use of TW/CW on social media platforms.

4.3.2 Design Recommendations for Warning Systems

Participants provided design recommendations for warning systems that they believed platforms should consider incorporating, drawing from their personal experiences and discussions with us.

> *I guess the platform at least should provide the feature, but how to use it? Again, it depends on the community, how they are using it, but at least having the feature is like I would consider it as a duty for the platform to do it, to add such features. (P13)*

We divided these recommendations into the viewer's perspective and the poster's perspective.

Viewer's Side - Personalized Filtering Of Content Having Warnings

Because of the nuances in triggers and the complexity of giving warnings, the participants (7/15) wished for mechanisms where the viewers could filter content based on their personal triggers.

> *And then for triggering, because triggering is like it's different for different people, I would like it to curate the stuff for me so that the things I find triggering, you know, can be avoided. So it would be a great feature if, you know, like if it could ask me, "Hey, do you want to see about political? Do you want to see political stories?" I would definitely mark that as No. (P02)*

P08 shared that with having specific categories for warnings, there should be an option for users in *Settings* to toggle specific content on and off based on the warning categories given to posts.

Let's just say it's super basic, like a filter that covers an image that has different types of descriptions that you could use. So it's a trigger warning for sexual assault, trigger warning for graphic content, whatever it is. You have different categories you could apply to that filter so that when somebody scrolls to that post, they can immediately see the category of things that you were trying to prepare people for. With that, then in your settings, you could have a toggle that says "Always", I always make sure that I get that prompt for these different types of warnings. Or maybe I don't care about graphic content like that doesn't upset me. I can turn off all of the filters that would be applied to things that have been declared graphic content so that I don't have to go through that interaction step or it'll just always come through. I mean, I don't know why you'd want to do that, but I think that would be. The best way to give the user control of what they see and what they choose to be filtered based on the input by other users who post. (P08)

Such a feature would also enhance the user experience for users like P13 who are not easily triggered. They can turn the setting off and will not have to interact with the warning repeatedly.

P-13: Yes, I would like have more such features like adding a granularity settings option for the viewers as well as the poster.

Interviewer: Right.

P-13: That would be really helpful. Like if I don't feel a trigger from a flashing light, then I can just change it from my setting. It's not like every person had to like click again and again multiple times. Like it's a bad UI experience or user UX experience. Same thing again for YouTube videos as well. If I'm watching a

> *certain video, a certain kind of video, if I'm watching with it, then if the algorithm might understand it like this person is okay and not show him the understand button again and again.*

P13 further justified that the *"granularity setting feature*" would help reduce the load on the platform content moderation team. That is due to users seeing less potentially triggering content and reducing the number of reports for review.

P15 wished for user input of personal triggers similar to asking what content they want to see in the onboarding process on a platform:

> *I really wish that, you know, like, you know, when you make, like, a new account, when you sign up for a service, like they ask you, like, what kind of news do you typically would like to read? I wish there was, like, filters like that. I was like, I don't need to see that. I don't need to hear that. Like, don't show those things to me. I wish there are things like that. (P15)*

However, P12 raised concern that personalized filtering based on TW/CW will make people more polarized in their beliefs as they might use such features to hide opposing views.

> *I believe it will increase the bias. Liberal people will more and more see liberal people, even though it's already like that. (P12)*

- **Skip-all Feature:** The TikTok Photosensitive warning example had buttons to "Skip All" photosensitive content or "Watch Anyway". On seeing this, P04 said they want to avoid any content that needs a warning, so they want such an option for most triggers.

P-04: I like the photosensitive warning one, especially because it lets you skip all.

Interviewer: So then you don't have to do it again and again.

P-04: Yeah. Well, especially because it's specific enough, right? Like a photosensitive warning. It's like a very specific type of trigger. So I like that one. Then again, I feel like most content that needs a trigger warning I would skip anyways. So I would like an option like that for most trigger warnings.

This would also help to avoid repetitive interactions. But P10 worries that with the *"Skip all"* option they will miss out on important resources and call-to-action being shared for specific topics, like transphobia or sexual abuse:

Just because even though the content warnings are there, I think it's oftentimes like in posts of like sexual abuse or when people are talking about transphobia or things like that, they're usually talking of it in terms of like... Or at least not usually, but sometimes it's like a call to action or like links to resources and things like that. And I worry that, like if you were to skip all of those, at least not on the photosensitive side, but on the specific categories... That you would kind of that you would miss out on maybe some of the important stuff. (P10)

- **Disable Warnings:** Two participants who hardly got triggered mentioned that having a way to disable warnings would be a better user experience for them. P13 questioned the photosensitive warning on TikTok UI having only *"Skip all"* or *"Watch Video"* options. They said that if they are fine with watching a certain category of content, there should be another option to watch all such videos instead of clicking

"watch video" constantly. Similarly, P12 proposed that platforms can add a filter where they enable or disable overall warnings because folks who do not get triggered would prefer to disable warnings.

> *Or maybe like they can have a filter where they enable warning or disable warning for, like, for people like me I can disable it. For people that need it, they can enable the warning for them. (P12)*

- **Collecting Information To Predict Triggers:** P12 proposed that in the on-boarding process, platforms can ask for basic information like politics, and religion to identify probable triggers. Based on the information collected, the platform algorithm can then update the potential triggers as the user engages with more posts.

  > *They* [the platform] *can ask for basic information like politics, religion to identify possible triggers specifically. The algorithm can learn through machine learning as they engage in more posts. (P12)*

  However, P12 acknowledged that such a feature can raise privacy issues.

  > *I wish they did not [collect information]. But like they will, I think they'll still do it. So, yeah. I mean, yes, there will be privacy concerns, but they'll do it anyway. So I don't mind. (P12)*

In conclusion, participants recommended a number of personalized content filtering options to enhance user control and customization while navigating triggering content.

Recommendations On The Poster's Side

- **Specific Categories of TW/CW:** From the author's perspective, (4/15) participants suggested that the platform should give posters the tools to add specific cate-

gories of warnings. P11 and P14 envisioned having a feature to add descriptive warning tags while making a post. P11 additionally proposed that platforms, like Instagram, should make the images editable in case the poster forgot to add warning tags on top of the image.

P13 recommended the addition of granularity settings on the poster's side for common triggers, categorizing them by natural causes and human fault, with sub-categories for intentional and unintentional triggers.

> *There. I mean, besides making it accessible and very easy, like when you go to post, maybe there is like a right on the side of the screen, just the, like the pull-down menu for like choosing them [the topic of warning]. (P08)*

P13 recommended adding granularity settings for some common triggers for posters as well. According to P13, one way the platform can categorize triggering content is: by natural causes and due to human fault, with sub-categories of intentional and unintentional.

> *So again, like sensitive content, you can characterize it in two ways. One is like happening due to natural causes and one which is happening due to human fault. Or maybe. Again, in that two-case scenario, one intentional and one (un)intentional, something like that, like an airplane crash, it could be an unintentional human-made.*

Challenges with Specific Categories: Two participants highlighted challenges related to the recommendation for specific categories in TW/CW. They pointed out that while personalized warnings could be valuable, they might also pose difficulties for platforms in terms of regulation. For example, P12 mentioned that while sexual content and violence tend to trigger masses, *"there are also topics that are like they*

only trigger specific kind of people, for example, that religious posts, like political posts" that would be tougher to manage. Additionally, P13 raised a concern about user experience. If someone frequently posts triggering content, repeatedly flagging it with the appropriate warnings might lead to a poor user experience on the poster's side. They further stated that if the authors have to give specific warnings, then it will worsen the user experience from their end. Giving specific options to mark content as sensitive may result in too many clicks for the author to add a warning.

More granularity would be like what kind of content and that option should be provided to the author. I think that is a good way to hide this content specifically. But again, it will have lesser, what we say, not bad, but comparatively worse user experience because you have to do too many clicks. (P13)

In conclusion, some participants suggested providing specific categories for TW/CW to enhance user control and clarity during content creation. However, challenges such as regulating diverse triggers were also raised, indicating the need for a balanced approach to implementation.

- **Nudging Authors To Add Warnings:** Some participants (7/15) suggested that the platform should not let sensitive content without warning be widespread and should nudge the users to add warnings if it's detected they are not doing so. P10 mentioned that authors can be nudged when they are about to post by integrating it into the user interface:

  If it was like a checkbox or something that asked you a question, maybe as you were making a post, leaving a comment. I think it'd be easier to not forget or like, have it like, built into the UI of whatever app you're using. (P10)

Another route mentioned for the platform, by two participants, was to observe the engagement and then send in a nudge to edit the post and add a warning. If the author does not listen or if the post gets too popular, then the platform should step in and add the warning anyway. For instance, P14 mentioned that the comments section of a post can be evaluated by the platform to prompt the poster to add a warning.

> *Interviewer: Because they are the ones who are perpetrating the hateful content so. I guess in those cases, do you think like the platform should step in or try to nudge?*
>
> *P-14: Yes, I think like in those cases, again, use the comment section. If this is obviously not good, then I would like give them like a time to say, "hey, like either put it as this category or the website has to step down and say, Hey, we're going to add a sensitive content warning anyway based on the comments that we've gotten, and then we'll add a category. So either the user does it or if the user does not respond, they'll do it anyway. If it gets popular enough.*

If the creators are not adding a warning to sensitive content, then two participants mentioned that there should be some form of consequences for the creators unless a warning is included.

> *P-07: Yeah, like, maybe like YouTube would give them some warnings. "Hey, you have done this once". Yeah.*
>
> *Interviewer: Nudge them, I guess.*
>
> *P-07: Yeah. Yeah, something like that. Some sort of consequences if these content creators don't put these.*

Further P13 mentioned that unless there is some form of consequence, the creators will not start using warnings:

> *Maybe like they* [platform] *can put up a penalty of monetization penalty. Like your video has been reported so many times for such content, you will not earn that much or a penalty has been added to your account. So then they will start using such features to put a trigger warning in the post itself.*

– **Making It Easier For Posters To Add Warnings:** Two participants suggested that the platform should make it easier to add warnings. For instance, P12 thought *"if there's a button that will automatically add trigger warning for us that will be good"*. According to P10, it is less work and very direct on the poster's side if they just have to select from a check box to add a warning for the post. To increase the accessibility of warnings, P15 suggested that warnings should be added in the alt text.

– **Nuanced Community Feedback:** Interviews with participants suggested that there should be nuanced community feedback for nudging posters with accountability in adding warnings. P09 suggested an option for the audience to report content as triggering if the author had not properly marked it. With regards to such a feature, P13 expressed contentment with the user experience (UX) for reporting content on Instagram, especially for reporting triggering content. They found the three-click process, involving menu access, reporting, and category selection, to be straightforward and user-friendly. However, P09 cautioned with an example that when a majority of users flag content as controversial, the platform should step in and review that content.

> *P-09: But again, for example. We saw this Indian movie "Pathaan", which came out, people* [Hindu Extremists] *protested against the heroine*

wearing that saffron-colored dress.

Interviewer: Yeah.

P-09: People might find it triggering. They might find that... They might report it. But it's just that it's illogical and all. Even though a large set of people agreed with it, it's just lame that they did that. In that case, the company should not. The social media platform should not... [listen to them]. I'm like, Yeah, that's what I feel. This is an example I could give you.

Community feedback was also perceived as useful when the intention of the author was not to add warnings. P08 suggested that community feedback and platform intervention can be combined to tell the author to add a warning:

P-08: let's say they don't put a trigger warning on it and it gets reported for then the need or necessity to have one, that their posting capabilities may be sanctioned or something of the sort where they say, hey, clearly the community says this requires a content filter. You do not put a content filter on, Here's why you should or you need to do that. We're not going to let you post in this community for whatever period of time. Please don't do that again or whatever. That's the only thing I can think of. I don't know.

Interviewer: Yeah, that makes sense. Because, like, it's sort of like telling you that, hey, we don't condone this behavior. Like, you have to learn.

P-08: Right. It's community feedback

In summary, participants highlighted the importance of nudging posters and suggested various methods, including integrating warning prompts into the user interface, com-

munity feedback, and potential consequences for non-compliance, to add warnings for sensitive content. They also emphasized the need to make the process of adding warnings easier for posters and incorporating nuanced community feedback.

New Ways to Tackle Warning Systems

Few participants (2/15) suggested some new ways to tackle giving warnings. P01 suggested something similar to the extension *Shinigami Eyes*, where the content is colored green if it is trans-friendly and red if it is transphobic. According to P01, color-coding social media content based on how triggering that topic is should be done more often so that you do not accidentally stumble upon it. The color coding can be community-built.

> *I think stuff like that should be done more often. Like just, hey, you know, looking this person posts something about, you know, so you might accidentally come across it. I like color-coding things. ... What I like about the extension, like when say, I'm on Wikipedia or Twitter, they'll be, they'll have the different colors there. So like sometimes they'll be certain Twitter accounts that I used to follow. But if they're red well, I probably don't want to see them, so I'll just unfollow them and then I don't have to accidentally find these things. If someone else has already found them and made sure people knew about it. And that's the other thing that's good about it is community-built as to the warning about it. (P01)*

In addition, P15 suggested something like a PG-rated system from the art and entertainment industry to be implemented on social media because *"TikTok, like literally, it's like a mini-movie empire these days"* (P15).

4.3.3 Post-TW/CW Measures

An Aftermath Response: Doom-Scrolling

Participants highlighted that one of the responses of consuming sensitive and triggering content led to a cycle of seeking out more distressing information, in other words doom-scrolling. Doom-scrolling is a state of media use typically characterized by individuals persistently scrolling through their social media newsfeeds with an obsessive focus on distressing, depressing, or otherwise negative information [34]. P14 described it as *"the idea of you want to know more, and it's giving feeding into that trigger a little bit."* For instance, P07 mentioned that in the past, they would actively seek out triggering content, such as videos depicting accidents, without fully understanding its impact on their mental health.

> *A video of someone having a road accident, so maybe their head just burst. It's something. I don't know why. It's something that back home people are really drawn to. I really don't know. I stopped that. But I remember back home when I didn't understand about mental health and stuff. I also will seek to that. I had a lot of those, but not anymore. (P07)*

Additionally, P07 acknowledged that it's difficult to resist the urge to continue engaging with such material, despite recognizing its adverse effects. Furthermore, P12 expressed skepticism regarding the potential impact of TW/CW on social media platforms' profit motives. They suggested that social media platforms may benefit from users engaging with triggering content for extended periods, as increased engagement often translates into higher profits for these platforms and *"social media is a for-profit company"*.

> *It's good for the algorithm because the more they engage with the social media platform, they get paid more. So I think that's why they keep recommending,*

even though if you don't want to watch it. (P12)

Impact of Doom-scrolling: Doom-scrolling made some participants feel drained, or even numb (P03) when subjected to such content. P02, for example, mentioned that the constant presence of certain issues on social media platforms can lead to a sense of emotional exhaustion.

And when I see it [news, social movements] repeatedly on the internet, like on social media, it kind of drains you a little. (P02)

Similarly, P11 indicated that while they aimed to stay informed, there were moments when the overwhelming volume of sensitive content prompted them to disengage temporarily.

Like, again, I kind of just want to be as aware of what's happening as possible. But sometimes it gets to be too much, and that's when I skip. (P11)

P07 emphasized the need for TW/CW, as many individuals might consume such content without considering the potential consequences.

Aftermath Recommendations

Beyond implementing TW/CW as preventative measures, interviews with participants revealed the need for additional resources to assist users when they are exposed to triggering content. These post-exposure measures aim to complement TW/CW practices and address the complexities involved.

SOS System with Resources: Some participants (5/15) felt an SOS system for mental health resources and emergency contact information for crisis hotlines would be beneficial on

social media. The pop-up could be similar to the COVID-19 pop-ups with CDC resources that were common on Facebook and Instagram during the pandemic. The platform should also keep *"updating them as needed" (P14)*. One participant's comment described the need effectively.

> *Everyone has the resources to be on Instagram right now, but not everyone has the resource or the language to access the support that they might need or even understand that they're triggered at that moment.* (P03)

Since *"everybody's unique and they have their own ways of dealing with things"* (P08), personalizing the resources would make them more effective. According to P07, having different resources for specific sensitive topics should be recommended.

> *So it should be more maybe oh for academia abuse, for child abuse or sexual abuse should be more different, you know, directions where do you need to go if you get triggered. (P07)*

Designate Support Contacts: Four participants mentioned that depending on the topic they are triggered by, they reach out to different folks in their close circle. The platform can personalize the self-care resources further by helping the user designate certain friends/family members to be recommended contact points when triggered by a social media post.

> *When you talk to people and you understand that I have gone through the same process and all, it's not just me. Then you feel a little comforted. (P09)*

Panic or Triggered Button: P08 envisioned something like a panic or triggered button *"that would then maybe remove you from what you're reading or what you're seeing"* and help provide tools that *"would work best for you"*.

Curated Playlist of Comforting Content: Five participants mentioned that watching comforting content or music helps to calm down after getting triggered, or to break the doom-scrolling cycle. So, the platform could recommend users create their own playlist of heart-warming content which the platform can recommend in moments of crisis. They could also detect after a series of heavy posts/videos that you may want a break with personalized lighthearted content. For instance, P07 already has a playlist on YouTube which they watch when they realize they are doom-scrolling. They further shared that listening to music also helps.

> *Interviewer: And I guess, do you personally follow something? I think one thing you said was you got some lighthearted content after you get triggered.*
>
> *P-07: Yeah, I do that. I have actually list (laughs) a playlist, yeah I have actual playlist on YouTube. I'll go there when I'm already like, I feel like, "Oh, okay, I'm done". This disturbing stuff I watch. So.*
>
> *Interviewer: Is there anything else you do like to calm down?*
>
> *P-07: Uh. Listening to music that's that helps so much. And go out simply go out, for fresh air. Just do something else then, because social media, YouTube, and stuff like that, that's endless. So I can keep clicking new channels of crime stuff like new videos, just nonstop.*

Intervention by the Platform: Some participants (4/15) thought that the platform should intervene when a user is viewing a lot of potentially triggering content (doom-scrolling) since the user *"might need to take like mental breaks"* (P15). On YouTube, P07 thought there should be a feature where the platform can have a pop-up notification when you watch too many flagged or sensitive content videos:

> *"Oh, you watch too many cat videos". It's not a problem. But when you watch too many content-flagged videos. Yeah, I think it would be nice if YouTube has some automatic pop-up for me. "Hey, you've been spending 3 hours on this". (P07)*

To avoid doom-scrolling, adding breaks here and there, like muting, avoiding a topic (P15), to *"just being forced to like not be able to do anything about it is"* (P14) could help. P14 suggested there could be a pop-up if the user has been searching through similar topics, or tags for a while and the platform can provide them with something else to view as a break.

> *Yeah, I think Instagram, for example, has like a general, like, uh, like. It allows you like, "Hey, like you've been on this poster. ... For so long". Or for example, or like, you've been like, how certain posts are connected to one another, um, through like, tags and stuff like that. "Hey, like you've been searching through these topics", maybe like, um, "let's provide something else for you to", like, look through, uh, for like a few minutes. (P14)*

Most of these platform interventions, however, rely on social media users telling when they are triggered by a post. Since many users may not report or block the triggering content, it becomes a challenge for the platform to know if a user is triggered. P02 suggested that the platform could calculate a *"well-being score"* based on metrics like the content a user viewed, how long they spent on certain pages or posts, what links/buttons they clicked on, and maybe even eye-tracking. After the score is detected to drop below a certain threshold, the platform can give a pop-up to recommend you utilize some of the mental health resources or take a break from the platform. However, this comes at the price of giving up users' privacy, and the platform accessing and quantifying personal sensitive data.

> *I think I'm fine with them tracking my eye and like having my mental health score if they are transparent about it. (P02)*

Mindful Practices: In addition, mindful practices like gratitude journalling and taking decompression time to process after getting triggered also can be incorporated into the platform's self-care resources.

> *I certainly do, like, get emotional or angry and like, but I'm aware like, what caused it, I think. So at least for me, it's just like some decompression time just to. Like I know that it'll go away. (P10)*

> *Interviewer: I guess you're more aware of like what you should, you are able to handle, and what you are not able to handle?*
>
> *P-15: Yes, I've been doing gratitude journaling. I think it's more like that helping me to understand that, you know, people have really different experiences. Yeah. The fact that I should be more accepting. And loving. Even to all the people that I don't like. Not really. But it's more accepting than loving. Yeah.*

In conclusion, the interviews highlighted the importance of not only implementing warnings but also providing robust post-exposure measures to support users when they encounter triggering content on social media platforms.

Chapter 5

Discussion

In general, our participants expressed a preference for the use of warnings on social media platforms. During the interviews, they provided valuable feedback aimed at improving the efficacy of these warnings. This adds another layer to the ongoing discourse surrounding the use of TW/CW in educational and clinical settings, which has primarily focused on the effectiveness of such warnings [9]. The concerns voiced in this ongoing discussion, concerning both the necessity and effectiveness of warnings, may be rooted in factors influencing their overall impact or lack thereof.

Our interviews unveiled a complex landscape of communication practices and diverse experiences with warnings among participants, indicating a lack of shared understanding. Consequently, content containing potentially triggering elements poses socio-technical challenges, stemming from the diverse nature of triggers and the various ways individuals engage with and interpret such content. Building on our findings and the challenges identified, we examine the implications for designing platform features related to TW/CW.

The challenges discussed by participants, as outlined above, shed light on potential avenues for improving the design of computer-mediated communication systems concerning TW/CW. However, our findings suggest that addressing these challenges may require a more intricate solution than simply introducing an additional feature. We delve into the delicate balance between warnings, user engagement, and specificity, as well as discuss features related to personalized content filtering, emphasizing considerations, incentives, nudges, and interventions

as strategies for effective design around warnings. Furthermore, we explore how platforms can enhance post-trigger support to better serve their users.

5.1 Impact of Warnings on User Engagement

We observed a tension between warnings and user engagement with posts. The prevailing viewpoint among our participants was that including a warning often leads to a reduction in post views. This perspective aligns with the idea that warnings empower viewers to make informed choices about engaging with potentially distressing content, a preference commonly appreciated by the audience [9, 19, 25]. Our research builds upon this perspective by exploring the effects of warnings on content creators and their motivations within the social media landscape. Our findings indicate that content creators are primarily focused on enhancing viewership and engagement, potentially discouraging them from utilizing warnings.

In specific instances, the inclusion of a warning has actually led to increased engagement. Our research revealed a phenomenon where being triggered by content can initiate a *doom-scrolling* behavior (discussed in 4.3.3), where viewers continue to consume similar content, especially during viral events or discussions related to controversial topics. In such cases, content creators might intentionally incorporate warnings to attract a larger audience. However, this situation presents a paradox. While content with warnings, which is contentious, can gain greater visibility due to heightened user interaction [12], it also raises the likelihood of viewers being exposed to warning-labeled content and potentially experiencing triggers themselves [18].

This complex interplay between warnings and their impact on user engagement warrants further exploration, taking into account the perspectives of both content creators and viewers. To gain a further understanding of this issue, a quantitative study is warranted. Such a study

could provide insights into whether content featuring TW/CW experiences an increase or decline in user engagement taking both viewer and poster perspectives into account.

5.2 Specificity in Warnings: A Social Translucence Phenomenon

In many cases, general warnings prove to be ineffective as they often lack the depth required to empower viewers to make confident decisions. This deficiency can result in anticipatory anxiety, a phenomenon that aligns with the findings of Bridgland et al.[8]. Instead, the preference leans towards more specific warnings tailored to the content, mirroring the insights related to photosensitive warnings in media [35]. However, we expanded this result from South et al. with our finding that overly specific warnings can sometimes have counter-productive effects and trigger individuals. This highlights the delicate balance must be struck in terms of specificity in warning systems. The presence of context plays a pivotal role in enabling viewers to make informed choices about whether to engage with or skip content.

The tension of offering a warning with adequate specificity, without delving into excessive detail, resonates with the concept of *social translucence* [14]. This tension underscores the importance of the three properties of socially translucent systems: visibility, awareness, and accountability, all of which play a crucial role in the warning process. From our findings, for content creators, being *aware* and *accountable* is essential in delivering effective warnings. On the viewers' side, we found that *awareness* and the clear *visibility* of warnings are vital for navigating content that may require such advisories. Maintaining this balance in warning specificity and effectiveness hinges on *shared awareness* of the inherent constraints within

the warning system.

5.3 Balancing Personalizing Warnings and Filtering with Filter Bubbles and Privacy Concerns

Participants expressed a desire for personalized content filtering based on their individual triggers (see Section 4.3.2). This aligns with the recommendations outlined in the trauma-informed computing framework [11]. Personalized warnings can be especially valuable in cases where content that society generally considers positive or harmless might trigger trauma. For instance, they can assist individuals in recovery from substance-use disorder in managing digital triggers [30]. However, it's essential to note that one participant raised concerns about the potential consequences of this approach, particularly the risk of increased polarization. This concern mirrors the possibility of creating "filter bubbles", which Chen et al. caution against [11]. While personalized filtering can help users avoid triggering content, it may also be misused as a means to entirely evade uncomfortable but non-triggering topics.

5.3.1 Privacy Considerations

Social media platforms commonly employ algorithms to personalize content, encompassing advertisements and recommendations, a practice that has previously stirred privacy concerns [38]. However, if these platforms have access to information regarding users' triggers, there is a potential risk of using this knowledge to target individuals with distressing content, potentially resulting in adverse psychological effects [16, 30]. Current algorithms lack an understanding of the evolving nature of triggers, particularly in response to life transitions such as recovery from substance use [30], experiences with disordered eating [16], or going

through a gender transition [20]. This aligns with recommendations aimed at ensuring that online platforms support marginalized individuals during significant life transitions [15].

Moreover, in cases where users respond to distressing content with doom-scrolling, platforms armed with knowledge of triggers may perpetuate this cycle by presenting more such content, thereby increasing user engagement and time spent on the app. The possession of knowledge about users' triggers can grant excessive power to the platform, contradicting participants' desires for the inclusion of warnings.

In addition, the risk of data breaches or leaks potentially places this sensitive trigger information in the wrong hands, paving the way for its misuse. Such misuse could result in harm, emotional distress, or exploitation, even leading to instances of discrimination. It is essential for users to have control over their data, including details about their triggers. They should have the ability to access, delete, or manage this information based on their preferences. Consequently, the storage and security of trigger-related data must be carefully addressed while designing for personalized filtering.

5.4 Author-Based Focus on Features

In the existing literature, there has been a notable lack of attention directed towards authors' perspectives when it comes to employing content warnings. We have illuminated the perceived challenges authors face in utilizing TW/CW for social media content. Future research could delve deeper into testing strategies to address these challenges.

5.4.1 Cultivating Author Accountability

In certain scenarios, posters have the freedom to decide whether to include a warning in their content. When the process of adding warnings is not user-friendly or when authors are unaware of it, they may be less likely to include warnings. Consequently, it becomes crucial for the platform to establish incentives that encourage authors to incorporate warnings for sensitive content or impose penalties when they neglect to do so. One identified approach involves implementing demonetization or notifying content authors when a warning is missing in situations where it is deemed necessary.

5.4.2 Nudging and Intervention Strategies

One potential intervention strategy focuses on nudging authors to include warnings during the content posting process. This approach serves a dual purpose by increasing awareness and providing content creators with a hands-on, learning-by-doing experience.

Additionally, an intervention that involves post-editing to include a warning can also be synchronized with community feedback mechanisms. This can include features such as the report function or heightened engagement in the comments section. This integration becomes essential when authors disregard the previous nudge during the posting process or when the platform's sensitivity to detecting sensitive content falls short. A point to note here is that there exist some platforms like Twitter where the post-editing option is not available, so this strategy might not work unless posters are allowed to edit after such feedback notification.

5.5 Enhancing Post-Trigger Support

While the existing literature has predominantly focused on the effectiveness of warnings, our research has highlighted that achieving flawless warnings may be an unattainable goal due to the complexity of addressing triggers. Consequently, there is a growing recognition of the need to prioritize strategies aimed at assisting users *after* they have encountered triggering content (see Section 4.3.3). This perspective aligns with Phelan et al.'s findings, which suggest that an essential aspect of the substance-use recovery process involves reducing sensitivity to triggers and negativity to safeguard one's progress, as exposure to triggers is an inevitable part of the journey [30].

We propose that platforms introduce mindful features to help users build resilience and cope with the overwhelming feelings that can follow a triggering experience. These features can include something as simple as a personally curated playlist of comforting content for users to turn to when triggered, as suggested by participants in our study (See 4.3.3). An illustration of such an intervention is TypeOut, which employs a just-in-time intervention approach designed to combat excessive smartphone usage [40]. This method comprises an in-situ typing-based unlocking process to boost user engagement and the integration of self-affirmation-based typing content to amplify its efficacy. Such an intervention could be adapted to assist users who experience triggers, providing timely support when needed.

As participants have indicated, interventions offer a potential solution to address excessive exposure to triggering content (see Section 4.3.3), especially when personalized warnings might prove ineffective. For example, implementing a pop-up notification during mindless scrolling, indicating the duration of exposure to triggering content and offering relevant resources, could effectively help users refrain from further consumption of distressing material.

Given that triggers may evolve over time, interventions could also periodically assess whether

users' triggers have changed. For instance, a pop-up notification every few months could prompt users to update their trigger preferences or adjust the content they wish to avoid, based on their current life circumstances and transitions [15, 30].

However, it's essential to seek user feedback when they encounter triggering content to ensure that recommendations align better with their needs. Nevertheless, caution should be exercised when trying to predict triggers through metrics, as inaccuracies could frustrate users and raise privacy concerns.

In essence, the emphasis should shift toward designing measures to support users after exposure to triggers, potentially in conjunction with the use of warnings. This approach holds the potential to alleviate some of the difficulties in decision-making that arise when navigating sensitive content through warnings on social media platforms.

5.6 Emotional Impact on Researchers - Personal Reflections

While conducting semi-structured interviews with a smaller gap between sessions, I unexpectedly started having vivid and graphic dreams related to discussions about trigger warnings and content warnings (TW/CW). This experience raised my awareness of the potential mental health effects on individuals who work in manual content moderation roles as their jobs. It also posed challenges for me in conducting numerous interviews throughout the study.

During the analysis phase, I unexpectedly experienced grief that had a profound impact on me. At that time, juggling my research on trigger warnings and content warnings (TW/CW) while being in the United States felt overwhelming. I found it challenging to distance myself from my personal life and find solace in my research. This distressing event significantly

delayed the progress of my project, leading us to reduce the number of participants from 20 to 15. As a researcher, I've come to realize the complexities of separating personal triggers from the process of analyzing user interviews on this topic *(See Figure 5.1)*. This underscores the need for more research in this area. Looking ahead, I strongly recommend that researchers have access to mental health support or resources to help them navigate challenging situations and safeguard their well-being.

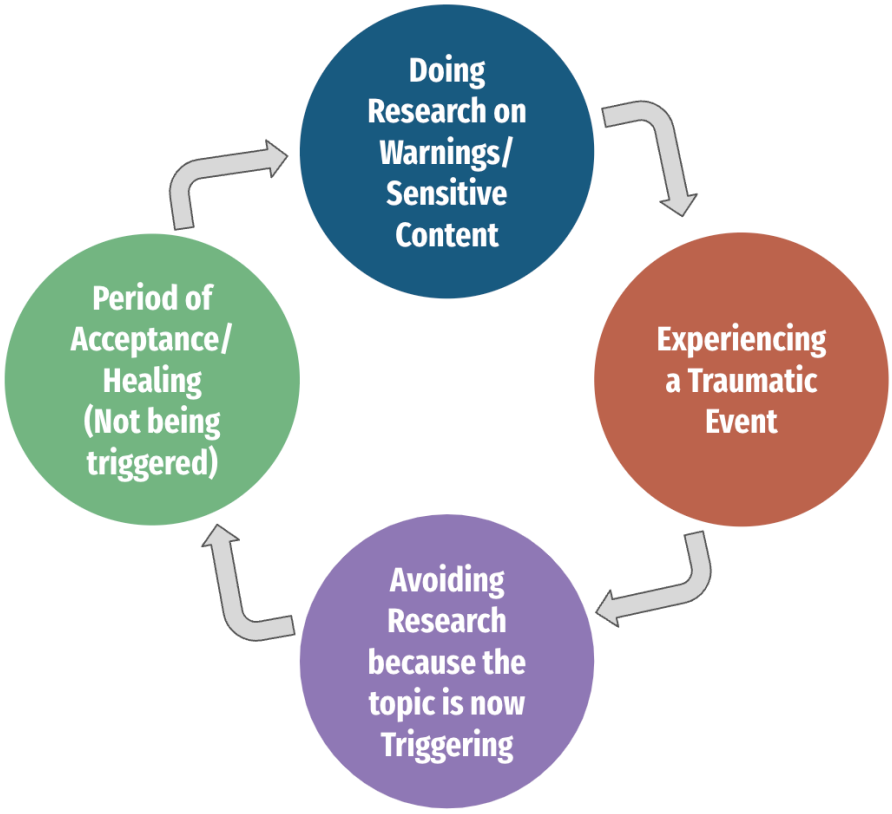


Figure 5.1: The Cycle of Working as a Researcher on Warnings While Going Through a Traumatic Personal Event

5.7 Limitations and Future Work

One limitation of our study is related to participant selection. The majority of participants we recruited demonstrated a pre-existing interest in and proactive approach towards TW/CW. However, we encountered challenges in recruiting individuals with opposing viewpoints—those who strongly disliked TW/CW. As a result, our study lacks representation from this particular group of participants.

Another limitation to consider is the broad scope of this study. For future research, narrowing the focus to a single platform could potentially yield more detailed and granular insights. Nonetheless, it is important to note that the absence of platform restrictions served in providing cross-platform insights.

As this research was conducted as part of my thesis, the data analysis was primarily undertaken by a single researcher, which may introduce a potential limitation. In qualitative research, involving multiple researchers can help mitigate bias and enhance the rigor of results. Future studies employing qualitative methods may benefit from collaborative analysis to achieve more robust outcomes.

As mentioned earlier, a quantitative study is needed to find out if and how the presence of TW/CW correlates with changes in user engagement. Additionally, further research should explore the usage of TW/CW in the context of targeted advertisements and the utilization of trigger information by the platform.

Finally, future research should prioritize the prototyping and user testing of nudging and intervention techniques aimed at encouraging the addition or viewing of TW/CW. These studies will help to determine the efficacy and impact of such methods.

Chapter 6

Conclusions

In conclusion, this thesis has undertaken a comprehensive exploration of social media users' attitudes and preferences regarding the utilization of TW/CW in the social media landscape. Through qualitative interviews with social media users, our aim was to delve deep into the challenges users perceive when employing TW/CW for social media content, the factors that sway users' decision-making processes when encountering content with TW/CW, and the design recommendations to enhance users' experiences with TW/CW.

Our findings have revealed the intricate nature of both adding and viewing content with warnings. We've identified factors that influence users' choices when it comes to engaging with content featuring TW/CW, explored ways to improve trigger education and TW/CW implementation, and distilled design recommendations rooted in participants' experiences. Furthermore, our results revealed potential post-trigger measures that platforms could consider implementing.

Subsequently, we discussed the design implications surrounding warning systems. These encompass navigating the delicate balance between warnings and user engagement, striking the right equilibrium between specific and generalized warnings, and addressing concerns regarding personalized filtering. We also acknowledged the emotional toll such research can have on researchers. Our study underscores the importance of enhancing author accountability, implementing nudges and interventions to encourage content creators to include warnings, and placing a strong emphasis on refining post-trigger support within the design of TW/CW

systems.

Looking ahead, there are avenues for future research to explore. Future research could focus on a single platform, involve multiple researchers for robustness, conduct quantitative studies on TW/CW's impact on user engagement, explore TW/CW around targeted ads, and prioritize user testing of nudging and intervention techniques.

In essence, our study contributes to the existing literature by providing user perspectives on warnings, shedding light on authorship challenges, suggesting ways in which authorship accountability can be improved based on the challenges, addressing the complexities of nuanced content, and moving the conversation in literature forward by putting focus on post-trigger support methods. These findings provide a valuable foundation for future research and the ongoing design of warning systems, ultimately fostering a safer and more inclusive social media environment for all users.

Chapter 7

Summary

Exposure to distressing content on social media can negatively impact users' mental well-being. To safeguard vulnerable individuals online, many platforms use trigger warnings (TW) and content warnings (CW). However, our understanding of social media users' preferences and challenges related to these warnings is limited. To address this gap, we conducted in-depth interviews with 15 participants, combining pre-survey data for selection and a two-phase thematic analysis.

Our findings reveal that users generally appreciate these warnings. However, comprehending and deciding when to use them is intricate, and influenced by various factors. Challenges arise when applying TW/CW on social media, involving decisions about which topics require warnings, navigating usage norms, and considering their impact on engagement. External aspects like presentation and internal factors like the viewer's mindset significantly shape user choices when engaging with content bearing TW/CW. Participants stressed the need for more education about warnings and triggers on social media. Participants also suggested post-trigger support measures to help users cope after encountering distressing content.

From these insights, we explored design recommendations that encompass personalized filtering while addressing privacy concerns, navigating the tension between warnings and engagement, enhancing specificity, nudging content creators, implementing interventions to mitigate excessive scrolling of triggering content, and integrating community feedback systems. Furthermore, our contribution extends to shedding light on the ways in which warning

systems can aid users in coping after encountering triggers.

While our research spans various platforms, we acknowledge limitations related to recruitment bias and the study's broad scope. Future research should explore individual platforms in greater detail and consider promoting author accountability, implementing nudges and interventions to add warnings, and enhancing post-trigger support to refine TW/CW design for social media users' evolving needs.

Bibliography

[1] Does the dog die: Crowdsourced emotional spoilers for movies, tv, books and more. URL https://www.doesthedogdie.com/.

[2] Shinigami eyes: A browser addon that highlights transphobic and trans-friendly social network pages and users with different colors. URL https://shinigami-eyes.github.io/.

[3] Samantha Artiga, Latoya Hill, Usha Ranji, and Ivette Gomez. What are the implications of the overturning of roe v. wade for racial disparities. *Kaiser Family Foundation*, 2022.

[4] Keely Ball. Let's talk about trigger warnings | keely ball | tedxwarwick. URL https://www.youtube.com/watch?v=Zcn1RMZCVI4&ab_channel=TEDxTalks.

[5] Corina Benjet, Evelyn Bromet, Elie G Karam, Ronald C Kessler, Katie A McLaughlin, Ayelet M Ruscio, Vicki Shahly, Dan J Stein, Maria Petukhova, E Hill, et al. The epidemiology of traumatic event exposure worldwide: results from the world mental health survey consortium. *Psychological medicine*, 46(2):327–343, 2016.

[6] Jack Bowker and Jacques Ophoff. Reducing exposure to hateful speech online. In *Science and Information Conference*, pages 630–645. Springer, 2022.

[7] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[8] Victoria ME Bridgland, Deanne M Green, Jacinta M Oulton, and Melanie KT Takarangi. Expecting the worst: Investigating the effects of trigger warnings on re-

actions to ambiguously themed photos. *Journal of Experimental Psychology: Applied*, 25(4):602, 2019.

[9] Victoria ME Bridgland, Payton J Jones, and Benjamin W Bellet. A meta-analysis of the efficacy of trigger warnings, content warnings, and content notes. *Clinical Psychological Science*, page 21677026231186625, 2022.

[10] Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Onni Gust, Fiona Ng, Elizabeth Evans, Emily Knox, et al. Typology of content warnings and trigger warnings: Systematic review. *PloS one*, 17(5):e0266722, 2022.

[11] Janet X Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. Trauma-informed computing: Towards safer technology experiences for all. In *CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2022.

[12] Zoey Chen and Jonah Berger. When, why, and how controversy causes conversation. *Journal of Consumer Research*, 40(3):580–593, 2013.

[13] Pedro VA de Freitas, Gabriel NP dos Santos, Antonio JG Busson, Álan LV Guedes, and Sérgio Colcher. A baseline for nsfw video detection in e-learning environments. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pages 357–360, 2019.

[14] Thomas Erickson and Wendy A Kellogg. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)*, 7(1):59–83, 2000.

[15] Jessica L Feuston, Michael Ann DeVito, Morgan Klaus Scheuerman, Katy Weathington,

Marianna Benitez, Bianca Z Perez, Lucy Sondheim, and Jed R Brubaker. " do you ladies relate?": Experiences of gender diverse people in online eating disorder communities. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–32, 2022.

[16] Liza Gak, Seyi Olojo, and Niloufar Salehi. The distressing ads that persist: Uncovering the harms of targeted weight-loss ads among users with histories of disordered eating. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–23, 2022.

[17] Evan George and Angela Hovey. Deciphering the trigger warning debate: a qualitative analysis of online comments. *Teaching in Higher Education*, 25(7):825–841, 2020.

[18] Amit Goldenberg and James J Gross. Digital emotion contagion. *Trends in cognitive sciences*, 24(4):316–328, 2020.

[19] Onni Gust. I use trigger warnings—but i'm not mollycoddling my students. *The Guardian*, 14, 2016.

[20] Oliver L Haimson, Justin Buss, Zu Weinger, Denny L Starks, Dykee Gorrell, and Briar Sweetbriar Baron. Trans time: Safety, privacy, and content warnings on a transgender-specific social media site. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27, 2020.

[21] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35, 2021.

[22] Larke N Huang, Rebecca Flatow, Tenly Biggs, Sara Afayee, Kelley Smith, Thomas Clark, and Mary Blake. Samhsa's concept of truama and guidance for a trauma-informed approach. 2014.

[23] Orla Hyland. *A qualitative investigation of people's attitudes towards the use of trigger warnings on social media*. PhD thesis, Dublin, National College of Ireland, 2023.

[24] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33, 2018.

[25] S Karasek. Trust me, trigger warnings are helpful. *The New York Times*, 2016.

[26] Caitlin Lustig, Artie Konrad, and Jed R Brubaker. Designing for the bittersweet: Improving sensitive experiences with recommender systems. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[27] Kaitlin Mahar, Amy X Zhang, and David Karger. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[28] Irene V Pasquetto, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*, 2020.

[29] Morgan E PettyJohn, Grace Anderson, and Heather L McCauley. Exploring survivor experiences on social media in the# metoo era: Clinical recommendations for addressing impacts on mental health and relationships. *Journal of Interpersonal Violence*, 37(21-22):NP20677–NP20700, 2022.

[30] Chanda Phelan, Jeremy Heyer, Rachel Pfafman, Connie Kerrigan, Golfo K Tzilos Wernette, Lynn Dombrowski, Andrew D Miller, and Jessica Pater. The work of digital

social re-entry in substance use disorder recovery. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–33, 2022.

[31] Casey Randazzo and Tawifq Ammari. "if someone downvoted my posts—that'd be the end of the world'': Designing safer online spaces for trauma survivors. 2023.

[32] Herbert J Rubin and Irene S Rubin. *Qualitative interviewing: The art of hearing data*. sage, 2011.

[33] Henrik Skaug Sætra and Jo Ese. Shinigami eyes and social media labelling as a technology for self-care.

[34] Bhakti Sharma, Susanna S Lee, and Benjamin K Johnson. The dark at the end of the tunnel: Doomscrolling on social media newsfeeds. 2022.

[35] Laura South, Caglar Yildirim, Amy Pavel, and Michelle A Borkin. Exploratory thematic analysis of crowdsourced photosensitivity warnings. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2023.

[36] Manuka Stratta, Julia Park, and Cooper deNicola. Automated content warnings for sensitive posts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[37] H Tankovska. Social media use during covid-19 worldwide–statistics & facts. *Statistica. https://www. statista. com/topics/7863/social-media-use-during-coronavirus-covid-19-worldwide*, 2021.

[38] Imdad Ullah, Roksana Boreli, and Salil S Kanhere. Privacy in targeted advertising: A survey. *arXiv preprint arXiv:2009.06861*, 2020.

[39] Nina Vlodder et al. Social representations of harm: Trigger warning practices on facebook watch. 2023.

[40] Xuhai Xu, Tianyuan Zou, Han Xiao, Yanzhang Li, Ruolin Wang, Tianyi Yuan, Yuntao Wang, Yuanchun Shi, Jennifer Mankoff, and Anind K Dey. Typeout: leveraging just-in-time self-affirmation for smartphone overuse reduction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[41] Nan Zhao and Guangyu Zhou. Social media use and mental health during the covid-19 pandemic: Moderator role of disaster stressor and mediator role of negative affect. *Applied Psychology: Health and Well-Being*, 12(4):1019–1038, 2020.

Appendices

Appendix A

Eligibility Survey

Investigating Social Media Users' Perceptions on Trigger and Content Warnings

We are going to be discussing about trigger and content warnings on social media. This includes user-added warnings to a post, whether it's a text-based post, image post, video post, or some combination of these mediums. While we know content moderation, where certain posts are deleted/hidden based on their
content, is also common on social media, we will not be discussing that today.

Participants must be:
1. Age of 18 years or older
2. Currently a US citizen or legal resident
3. Regular social media users who have seen trigger or content warnings on others' posts

Preferred:
1. Regular social media users who post frequently
2. Regular social media users who frequently add trigger or content warnings to their posts

See more details on the study:
https://bit.ly/3R3I6dl

This survey will take about 5 minutes to complete. At the end of this survey, you will be able to state if you're interested in a follow up interview by the researchers (and getting 15$ amazon gift card!)

Thank you for your time spent taking this survey!

* Required

1. Are you over the age of 18? *

Mark only one oval.

- ◯ Yes
- ◯ No

2. Are you currently a US citizen or resident (meaning legally authorised to live in the US, like being on F1 student visa or H1B)? *

Mark only one oval.

- ◯ Yes
- ◯ No

3. I consent to my involvement in this survey. *

Mark only one oval.

- ◯ Yes, I consent to my involvement in this survey

Part 1: Demographic Information

We are collecting this information to ensure we have a diverse set of participants in our study.
Don't worry, this is completely anonymous.

4. Age *

_____

5. Race/Ethnicity *

Check all that apply.

- ☐ Asian or Pacific Islander
- ☐ Black or African American
- ☐ Hispanic or Latino
- ☐ Native American or Alaskan Native
- ☐ White or Caucasian
- ☐ Other: _____

6. How do you describe your gender? *

Mark only one oval.

- ◯ Woman/Female
- ◯ Man/Male
- ◯ Non-binary
- ◯ Prefer not to answer
- ◯ Other: _____

7. Are you transgender? *

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Prefer not to answer

Your Experience with Trigger/Content Warnings

8. What social media platforms do you **regularly** use? *

(For purposes of this study, we are not considering strictly messaging sites/apps, like iMessage or Facebook Messenger.)

Check all that apply.

- [ ] Facebook
- [ ] Instagram
- [ ] Snapchat
- [ ] Reddit
- [ ] TikTok
- [ ] Discord
- [ ] Pinterest
- [ ] Twitter
- [ ] YouTube
- [ ] Tumblr
- [ ] Medium
- [ ] LinkedIn
- [ ] Quora
- [ ] Twitch
- [ ] VSCO
- [ ] BeReal
- [ ] Imgur

9. How often do you **view content** on social media?

Mark only one oval.

- ◯ Never
- ◯ Rarely
- ◯ Sometimes
- ◯ Often

10. How often do you **see** trigger and/or content warnings on others' social media posts? *

Mark only one oval.

- Never
- Rarely
- Sometimes
- Often
- Always

11. How often do you **post** on social media? *

Mark only one oval.

- Never
- Rarely
- Sometimes
- Often

12. How often do you **add** trigger and/or content warnings to your social media posts? *

Mark only one oval.

- Never
- Rarely
- Sometimes
- Often
- Always

13. Are you interested in being interviewed to share more about your answers? * The interview would be held on Zoom for 1-1.5 hours and you would be compensated with a $15 Amazon gift card.

Mark only one oval.

- Yes
- No

Interest in being Interviewed

This section is for those who are interested in being interviewed. If you are not, please go back to the previous section and select "no". Those who are interested will be reviewed by the researchers and then a subset will be selected.

14. Name *

_____

15. Email *

_____

16. Since this research deals with trauma and triggers, sensitive subjects may arise. Are there any topics you do not wish to discuss at all during the interview?

_____

_____

_____

_____

_____

Appendix B

Interview Questions

B.1 Introduction to the Interview

We are going to be discussing trigger and content warnings on social media. This includes user-added warnings to a post, whether a text-based post, image post, video post, or combination of these mediums. While we know content moderation, where certain posts are deleted/hidden based on their content, is common on social media, we will not discuss that today. Do you understand what I just explained?

Disclaimer: At any point, if you feel uncomfortable, please don't hesitate to let us know, and we will stop immediately.

We do not know as part of our study who is and who isn't a Virginia Tech student, and we won't ask, but as part of our Title IX reporting responsibilities, researchers have to report if sexual violence is reported to us that took place either on Virginia Tech's campus or involved Virginia Tech students.

Do you have any questions about the consent document? Can you sign and email it to us now?

B.2 Participant's use of social media

1. Just for background, what is your use of social media like? What sort of apps, what content do you view, and who is your audience? (No messaging apps)

B.3 Social media user as a viewer/consumer with current UI

2. First, we will discuss what you see as you spend time on social media platforms. What is your experience with seeing TW/CW on social media?

3. If/when you do see them, what do these warnings typically look like? What do they say? Where are they located? Can you give an example? *(This could be different depending on the type of post - image with caption, video, text-post - and possibly even the platform.)*

4. What topics usually have a TW/CW present? Describe the scenarios for expecting a TW/CW present on a post.

5. How do TW/CWs change your viewing/reading behaviors on social media? Do you primarily ignore them and read them anyway? Do you not view/read posts that could be triggering to you? Do you save for later when your mental health may be in a better place or where you are in a safe or private space? Why?

6. How do you decide to view or skip a post with TW/CW? Do you want them as a viewer? Why or why not?

7. What has been your experience seeing a post with TW/CW and deciding to view the content anyway?

(a) Has the content ever been unexpected behind the warning? Why was it unexpected? How did you react?

(b) The content did not match the warning

(c) Underestimating the effect it had on you

8. What has your experience been on reading/viewing something on social media that you wish had a TW/CW present? Any specific topics?

9. How do you currently avoid triggering content for you on social media?

B.4 Social media user as a poster/creator with current UI

10. Now, we will discuss how you, as a user (can) add warnings to posts. What is your experience with using TW/CW on social media? How do you decide when to use those? *(Most likely, it's relevant to their answers from 1 so refer to those answers.)*

11. If/when you do add TW/CW in your posts, what do these warnings typically look like? What does it say? Where is it located? Can you give an example? *(This could be different depending on the type of post - image with caption, video, text-post - and possibly even the platform.)*

12. For what topics do you use a TW/CW? Describe the scenarios when you expect to add them to your post/comment. What are the topics that you are fine with but still use TW/CW?

13. How does using TW/CW affect your posting style on social media? Does it not affect it at all? Does it make you feel more comfortable when posting about sensitive topics

that some folks might be triggered by?

14. Is posting TW/CW helpful or needed when something is important to you?

15. What are your views on putting TW/CW when posting content (including comments) on social media? Why do (or don't) you use them?

16. What are your thoughts on experiencing hesitation or difficulty when deciding whether or how to add TW/CW to your posts? Why do you think it is there?

17. What's your experience on realizing later on that you should've added a TW/CW to your post? How did you realize this? If not, why do you think that is?

18. What kinds of topics should have TW/CWs?

19. What are your thoughts on the difference between viewing or using a "content warning" and a "trigger warning"? Do they cover the same or different topics? Are the terms used interchangeably? Do you see or use one term more frequently? Which one? (If yes, give examples - both viewer and poster perspective)

B.5 Social media user and their ideal UI

B.5.1 Short version

20. Now, we are going to be discussing the ideal way you would like to see TW/CW shown on social media. What do you think the platform should do? Should it do anything? or should it leave it completely to users? If so, why or why not?

OR

B.5.2 Longer version if time permits

21. Now, we are going to be discussing the ideal way you would like to see TW/CW shown on social media. What are your frustrations so far with TW/CW on social media?

22. How do you wish TW/CWs were shown on social media? What would they say? What would they look like? Where would they be located in a post?

23. What better ways can other social media users add warnings to be more considerate of you and your triggers?

24. What are your views on having universal guidelines, or at least platform-wide guidelines on how and when to add or view TW/CW? Will they be helpful or not useful? Why?

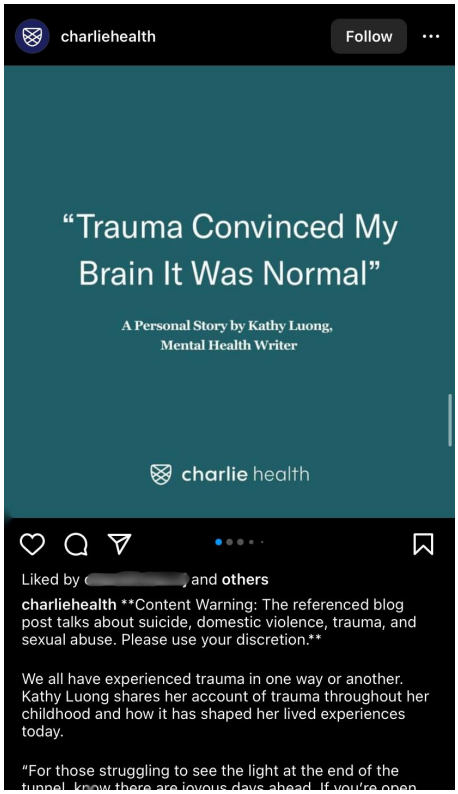
25. TW/CW are preventative measures; how about aftermath measures? (practices to calm down)

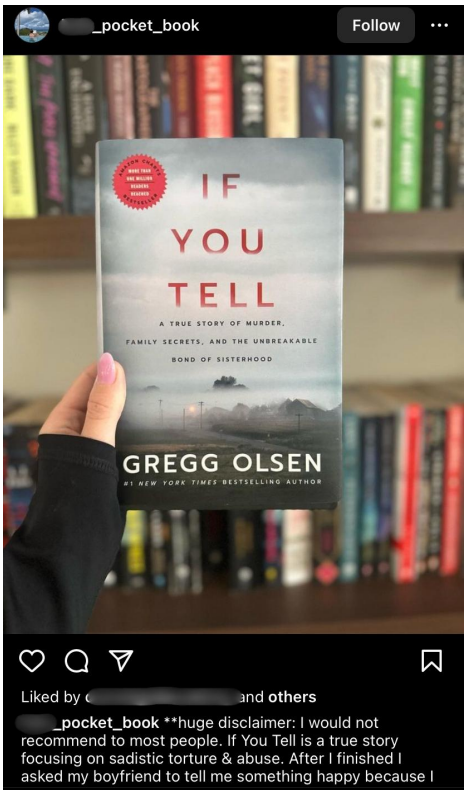
B.6 Wrap-Up

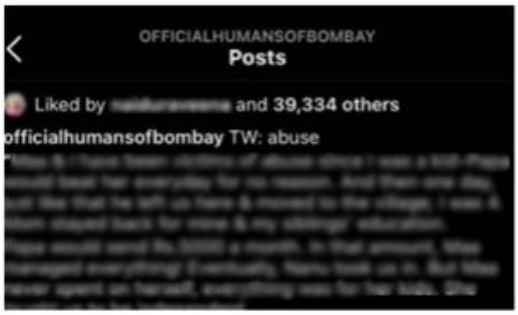
26. Is there anything we didn't talk about on the subject of TW/CW on social media that you would like to share?

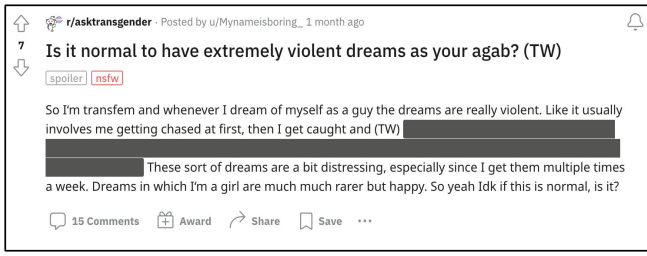
Appendix C

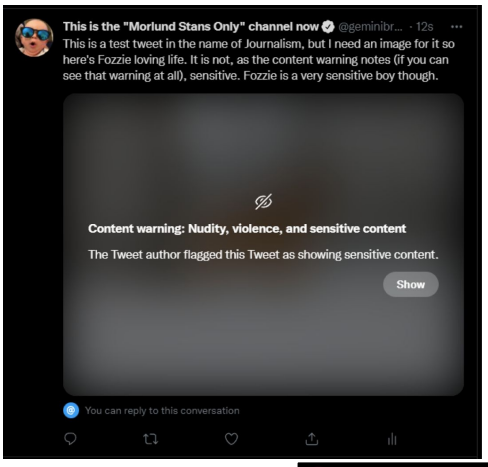
TW/CW Example Slides Shown in Interviews


charliehealth
Follow
"Trauma Convinced My Brain It Was Normal"
A Personal Story by Kathy Luong, Mental Health Writer
charlie health
Liked by [redacted] and others
charliehealth **Content Warning: The referenced blog post talks about suicide, domestic violence, trauma, and sexual abuse. Please use your discretion.**
We all have experienced trauma in one way or another. Kathy Luong shares her account of trauma throughout her childhood and how it has shaped her lived experiences today.
"For those struggling to see the light at the end of the

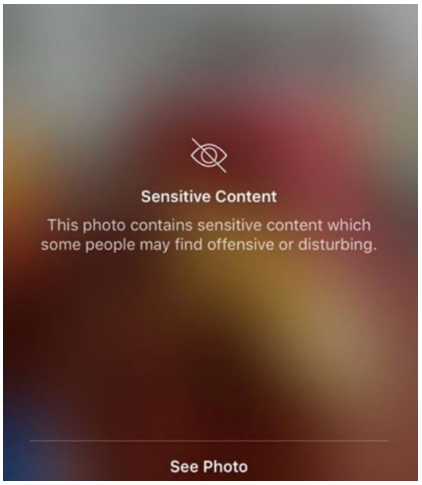

[redacted]_pocket_book
Follow
IF YOU TELL
A TRUE STORY OF MURDER, FAMILY SECRETS, AND THE UNBREAKABLE BOND OF SISTERHOOD
GREGG OLSEN
#1 NEW YORK TIMES BESTSELLING AUTHOR
Liked by [redacted] and others
[redacted]_pocket_book **huge disclaimer: I would not recommend to most people. If You Tell is a true story focusing on sadistic torture & abuse. After I finished I asked my boyfriend to tell me something happy because I

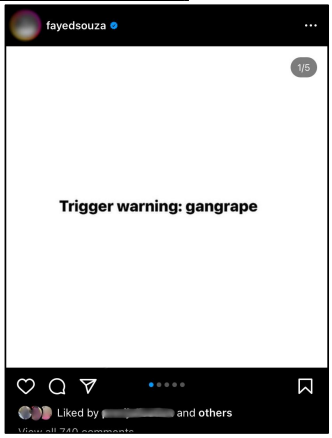

OFFICIALHUMANSOFBOMBAY
Posts
Liked by [redacted] and 39,334 others
officialhumansofbombay TW: abuse

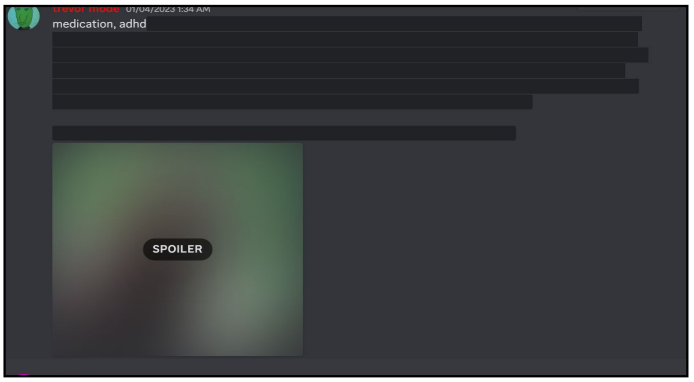

r/asktransgender · Posted by u/Mynameisboring_ 1 month ago
7
Is it normal to have extremely violent dreams as your agab? (TW)
spoiler
nsfw
So I'm transfem and whenever I dream of myself as a guy the dreams are really violent. Like it usually involves me getting chased at first, then I get caught and (TW) [redacted] These sort of dreams are a bit distressing, especially since I get them multiple times a week. Dreams in which I'm a girl are much much rarer but happy. So yeah Idk if this is normal, is it?
15 Comments
Award
Share
Save


This is the "Morlund Stans Only" channel now @geminibr... · 12s
This is a test tweet in the name of Journalism, but I need an image for it so here's Fozzie loving life. It is not, as the content warning notes (if you can see that warning at all), sensitive. Fozzie is a very sensitive boy though.
Content warning: Nudity, violence, and sensitive content
The Tweet author flagged this Tweet as showing sensitive content.
Show
You can reply to this conversation

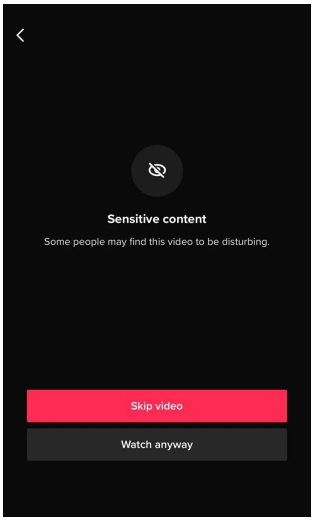

Zerquix18
This photo may show violent or graphic content.
We covered this photo so you can decide if you want to see it.
Uncover Photo

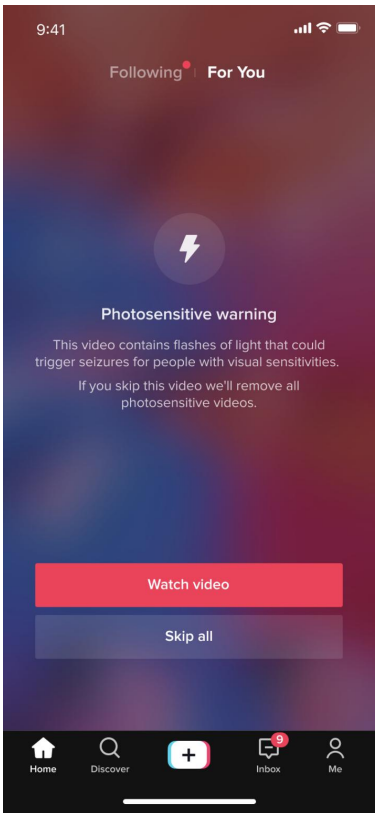

Sensitive Content
This photo contains sensitive content which some people may find offensive or disturbing.
See Photo

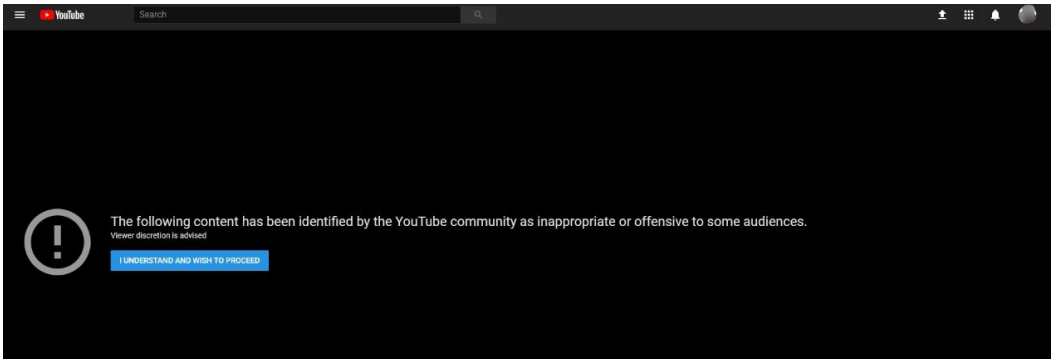

fayedsouza
1/5
Trigger warning: gangrape
Liked by [redacted] and others


medication, adhd
SPOILER


Content Warning:
The following video contains material that may be harmful or traumatizing to some audiences.


Sensitive content
Some people may find this video to be disturbing.
Skip video
Watch anyway


9:41
Following
For You
Photosensitive warning
This video contains flashes of light that could trigger seizures for people with visual sensitivities.
If you skip this video we'll remove all photosensitive videos.
Watch video
Skip all
Home
Discover
Inbox
Me


YouTube
Search
The following content has been identified by the YouTube community as inappropriate or offensive to some audiences.
Viewer discretion is advised
I UNDERSTAND AND WISH TO PROCEED