Beyond the blur: The empirical basis of Instagram's sensitive-content screens

By

Erin Tayla Simister

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy (Clinical Psychology)

College of Education, Psychology and Social Work

9th May 2024

TABLE OF CONTENTS

SUMMARY ..... **IV**
DECLARATION ..... **VI**
ACKNOWLEDGEMENT OF COUNTRY ..... **VII**
ACKNOWLEDGEMENTS ..... **VIII**
LIST OF FIGURES ..... **IX**
LIST OF TABLES ..... **I**
1 LITERATURE REVIEW **1**
1.1 Sensitive-Content Screens: A History 2
1.2 Advocates' Claims from an Emotion Regulation Perspective 8
1.3 Do Sensitive-Content Screens Deter People from Viewing Sensitive Content? 12
1.4 Do Sensitive-Content Screens Emotionally Prepare People to View Sensitive Content? 20
1.5 Existing Research Gaps and Implications 30
1.6 Summary 33
2 OVERVIEW OF THESIS STUDIES **34**
Chapter 3: Studies 1a and 1b 34
Chapter 4: Study 2 36
Chapter 5: Studies 3a and 3b 37
Chapter 6: Studies 4a and 4b 38
3 INVESTIGATING WHETHER INSTAGRAM'S SENSITIVE-CONTENT SCREENS DETER PEOPLE FROM VIEWING NEGATIVE CONTENT **40**
Abstract 40
Introduction 41
Study 1a 45
Study 1b 57
General Discussion 67
Supplementary Materials 72
4 INVESTIGATING THE ROLE OF INFORMATION SEEKING BEHAVIOUR IN THE DECISION TO UNCOVER SENSITIVE-CONTENT SCREENS **94**
Abstract 94
Introduction 95
Study 2 99
Supplementary Materials 117
5 INVESTIGATING WHETHER ADDING CONTENT-RELATED INFORMATION TO SENSITIVE-CONTENT SCREENS CREATES AN EMOTIONAL COST **122**

Abstract 122
Introduction 123
Study 3a 127
Study 3b 135
General Discussion 143
Supplementary Materials 148
6 INVESTIGATING WHETHER ADDING COGNITIVE EMOTION REGULATION INSTRUCTIONS TO SENSITIVE-CONTENT SCREENS REDUCES DISTRESS 159
Abstract 159
Introduction 160
Study 4a 163
Study 4b 175
General Discussion 182
Supplementary Materials 189
7 GENERAL DISCUSSION 203
7.1 Summary of Findings 203
7.2 Theoretical Implications 219
7.3 Methodological Implications 222
7.4 Practical Implications 223
7.5 Clinical Implications 228
7.6 Limitations and Future Directions 229
7.7 Conclusion 238
REFERENCES 240
APPENDICES 268
Appendix A: Image Stimuli (Study 1a and 1b) 268
Appendix B: Image Stimuli and Corresponding Brief and Detailed Content Descriptions (Studies 2, 3a and 3b) 275
Appendix C: Pre-Task Questions (for all studies) 277
Appendix D: Reasons for Uncovering Questionnaire (prior to PCA) 278
Appendix E: The Short-form Spielberger State-Trait Anxiety Inventory (STAI-6; Spielberger, 1983) 281
Appendix F: The Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) 282
Appendix G: The Depression, Anxiety and Stress Scale (DASS-21; Lovibond & Lovibond, 1995) 283
Appendix H: The Scale of General Well-Being short form (SGWB-14; Longo et al., 2018) 284
Appendix I: Trauma History Screen (THS; Carlson et al., 2011) 285

Appendix J: The Posttraumatic Stress Disorder Checklist (PCL-5; Weathers et al., 2013) 286
Appendix K: The Centrality of Event Scale Short Form (CES; Berntsen & Rubin, 2006) 287
Appendix L: The Five-Dimensional Curiosity Scale Revised (5DCR; Kashdan et al., 2020) 288
Appendix M: Acceptance and Action Questionnaire-II (AAQ-II; Bond et al., 2011).....290
Appendix N: Perth Emotion Regulation Competency Inventory (PERCI; Preece et al., 2018) 291
Appendix O: Intolerance of Uncertainty Scale-Short Form (IUS-12; Carleton et al., 2007) 293
Appendix P: Task Instructions and Post-Task Questions for Studies 1a and 1b 294
Appendix Q: Task Instructions and Post-Task Questions for Study 2.....297
Appendix R: Task Instructions and Post-Task Questions for Studies 3a and 3b.....297
Appendix S: Task Instructions and Post-Task Questions for Studies 4a and 4b 300

Summary

Instagram, along with other social media platforms, blur sensitive images and provide a warning—with the intention to minimise harm—but there is no empirical basis for these *sensitive-content screens*. My thesis aimed to address existing research gaps by examining behavioural and emotional responses (e.g., anxiety) to sensitive-content screens, before investigating adaptions to improve their utility as a harm minimisation tool.

First, I developed a simulated Instagram image-viewing task and gave participants the opportunity to uncover sensitive-content screens (Chapters 3, 4 and 5). I found most participants, including vulnerable people (e.g., with higher rates of depression), uncovered the first sensitive-content screen they saw, and many participants *repeatedly* uncovered screens. Participants also reported their emotional reactions to sensitive-content screens and the forewarned content (Chapters 5 and 6). Consistent with existing research, I found sensitive-content screens created a noxious anticipatory period that did not translate to an emotional benefit when participants viewed the forewarned content. Together, these findings suggest sensitive-content screens do not deter people from viewing sensitive content *or* help them emotionally prepare for it.

Second, I explored the reasons underpinning people's uncovering behaviour. I first developed a questionnaire based on existing theory and related literature (e.g., on uncertainty; Chapter 3). Participants rated their endorsement with items (e.g., "I uncovered the screened image(s) because I was eager to learn what the image was"), and I ran principal component analyses to identify the *key* uncovering reasons—information seeking behaviour, positive and negative affect driven behaviour, and avoidance behaviour. I then focused on information seeking behaviour because it was the most strongly endorsed; I manipulated the amount of content-related information *on* sensitive-content screens to examine whether screens prompt uncovering behaviour (Chapter 4). Consistent with this idea, I found participants uncovered

screens *most* often when screens appeared in their current format (i.e., without content-related information).

Finally, I investigated adaptions to improve sensitive-content screens. I first examined whether adding content-related information to sensitive-content screens reduced uncovering behaviour (Chapter 4). It did: participants uncovered sensitive-content screens *least* often when screens had content-related information. Importantly, I found no evidence of an emotional cost to adding *brief* content-related information to screens: participants reported similar anticipatory anxiety and image-related distress whether they saw sensitive-content screens with or without brief content descriptions (Chapter 5). I next examined whether providing emotion regulation instructions *on* sensitive-content screens reduced image-related distress (Chapter 6). They did: participants had lower image-related distress after negative images where I instructed them to use distraction and reappraisal (vs. no instructions).

Overall, my thesis provides a new and original contribution to the literature in three key ways. It demonstrates: 1) sensitive-content screens in their current format do not function as intended, 2) people uncover sensitive-content screens for different reasons, and 3) adapting sensitive-content screens can improve their utility as a harm minimisation tool. My thesis has implications. Theoretically, my findings help develop a framework for understanding *how* and *why* people respond to sensitive-content screens. Methodologically, my thesis influences how we investigate behavioural and emotional responses to warning systems. Practically, my findings suggest Instagram and social media platforms alike (e.g., TikTok) should move beyond merely warning about upcoming content. Clinically, my findings raise considerations for clinicians working with people (e.g., with depression) who seek out sensitive and potentially distressing content.

Declaration

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted or degree or diploma in any university; and
2. the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

Signed: Erin Tayla Simister

Date: 27/02/2024

Acknowledgement of Country

I would like to acknowledge that this work was produced on the lands of the Kaurna nation. I recognise the Traditional Custodians of the land where my research was conducted and pay my respects to their Elders past, present, and emerging.

Acknowledgements

First and foremost, thank you to my supervisor, Professor Melanie Takarangi. It has been a pleasure to learn from you over the past 4 years, and to have been part of the ever-strengthening Forensic and Clinical Cognition Lab. Thank you for being so generous with your time, for sharing your wealth of knowledge, and for pushing me (in your ever so kind way) to pull this body of work together. I admire your dedication and commitment to all that you do.

I would also like to thank my secondary supervisor, Associate Professor Ryan Balzan. I have greatly appreciated your support and advice along the way. A big thank you also to my mentors (and collaborators), Dr Victoria Bridgland and Dr Ella Moeck. Your guidance and encouragement have been crucial in shaping this body of work into what it is today.

To the rest of the Forensic and Clinical Cognition Lab, thank you for being alongside me every step of the way. Nothing about this process has been easy, but you have all made it that little bit easier (and more enjoyable too). I owe a special thank you to both Catherine and Meghan, who have lifted me up on many occasions. To my clinical peers, thank you for helping me celebrating the small wins along the way. Our friendships have been a constant light throughout this process.

To the clinical staff, especially Professor Reg Nixon and Associate Professor Lisa Beatty, thank you for helping me develop my skills as a clinician.

To my biggest supports of all, Mum, and Dad, thank you. You have been with me through every high and low, and I appreciate all that you have done to allow me to pursue this career. I could not have done it without you both. To my life-long best friend and running partner, Kayla, thank you for listening to my complaints and for celebrating my successes. And finally, thank you to my partner, Lewis (and his family), you have been my place of respite. Thank you for your unwavering patience, love, and support.

List of Figures

Figure 1.1 *Example NAPS Image Modified to Look Like an Instagram Image with a Sensitive-Content Overlay* **.....48**

Figure 2.1 *Example NAPS Image Modified to Look Like Instagram Images with a Sensitive-Content Overlay and (a) No Content Description, (b) Brief Content Description, and (c) Detailed Content Description* **.....103**

Figure 4.1 *Example NAPS Image Modified to Look Like Instagram Images with a Sensitive-Content Overlay and (a) No Instruction to Regulate, (b) Instructions to Use Reappraisal, and (c) Instructions to Use Distraction* **.....166**

Figure 4.2 *Trial Structure for the Main Image Task (an Example of a Reappraisal Trial)* **.169**

Figure 4.3 *Distress Rating Estimates by Cognitive Emotion Regulation Condition* **.....172**

Figure 4.4 *Example NAPS Image Modified to Look Like Instagram Images with a Sensitive-Content Overlay and (a) No Instruction to Regulate, and (b) Instructions to Use Distraction* **.....178**

Figure 4.5 *Distress Rating Estimates by Cognitive Emotion Regulation Condition* **.....182**

List of Tables

Table 1.1 *Means (and Standard Deviations) for Vulnerability Measures* **.....53**

Table 1.2 *Correlations Between Uncovering Behaviour and Vulnerability Measures* **.....56**

Table 1.3 *Means (and Standard Deviations) for Reason Factors* **.....57**

Table 1.4 *Means (and Standard Deviations) for Vulnerability Measures* **.....63**

Table 1.5 *Correlations Between Uncovering Behaviour and Vulnerability Measures* **.....66**

Table 1.6 *Means (and Standard Deviations) for Reason Factors* **.....66**

Table 3.1 *Means (and Standard Deviations) for State Anxiety, by Condition and Time* **.....133**

Table 4.1 *Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Condition on Distress Ratings* **.....171**

Table 4.2 *Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Conditions on Distress Ratings* **.....181**

1 Literature Review

Molly Russell was 14-years old when she took her own life in 2017, after entering—what her father termed—Instagram's "dark rabbit hole of depressive suicidal content" (Crawford, 2019). In response to scrutiny following Molly's suicide, Instagram prohibited all graphic self-injury content and added sensitive-content screens—a specific type of *trigger warning*—to non-graphic self-injury related content (Mosseri, 2019a). However, Instagram first began using sensitive-content screens prior to Molly's suicide—suggesting the warning system, primarily designed for harm minimisation, may not function as intended.

Beyond Instagram, sensitive-content screens are now widely used across most social media platforms, including Facebook, TikTok and Reddit—to name a few. Advocates claim that such warning systems deter people from viewing sensitive content, or *if* people decide to view the content, allow them time to emotionally prepare for the content (e.g., Manne, 2015). Yet, critics argue that trigger warnings "coddle" people (e.g., by sheltering them from the "real world"; Lukianoff & Haidt, 2015) and may cause harm by encouraging avoidance. Most research to date has focused on more traditional (e.g., text based) trigger warnings, and the impacts they have on people's behaviour and emotional reactions. So far, this research suggests that trigger warnings in their current form are not beneficial and may instead lead to a risk of emotional harm (see Bridgland et al., 2023). However, there is limited research investigating the effects that sensitive-content screens, *specifically*, have on people's behaviour and emotional reactions.

The emerging literature on sensitive-content screens has found limited evidence in support of advocates' claims. Specifically, in one study, when researchers warned participants about an upcoming sensitive image via a sensitive-content screen, most decided to view the forewarned content (Bridgland, Bellet et al., 2022)—contrary to the claim that such warning systems act as a deterrent. In another study, participants experienced a noxious

anticipatory period (characterised by state anxiety) at the time of viewing a sensitive-content screen, and this anxiety did not translate to an emotional benefit when people eventually viewed the forewarned content (Takarangi et al., 2023)—contrary to the claim that such warning systems assist with emotional preparation. Taken together, existing research provides further support for the idea that sensitive-content screens may not function as intended.

However, there are many existing research gaps from the first (albeit small) wave of research. We do not know *how* people respond to sensitive-content screens when they see more than one screen, or *why* they respond the way they do. My thesis aims to first address these important research gaps, and then, using this empirical work as a foundation, investigate potential ways in which social media platforms can adapt sensitive-content screens to improve the screens utility as a harm minimisation tool.

1.1 Sensitive-Content Screens: A History

Instagram and Sensitive Content

At launch in 2010, Instagram was primarily a photo-sharing platform where people could post photos, follow other users, and like and comment on posts. However, since then, the overall user experience has evolved with the introduction of numerous features (Hackett, 2023). In 2011, Instagram introduced *hashtags*—a combination of letters, numbers, and/or emojis preceded by the # symbol (e.g., #depression). Users can add hashtags to their posts to categorise the content and make it more discoverable to interested users (Newberry, 2023). In 2016, Instagram introduced their *algorithm*; the algorithm works out what type of content users like and recommends more of it to them. However, when Instagram first introduced these features, they did not differentiate the *type* of content made more readily accessible to users. Therefore, not only could users more easily view sensitive or graphic content by searching for related hashtags (e.g., #depression, #selfharm), but after engaging with the content, the algorithm would then show users even more of this content. Thus, the once

relatively harmless social media platform began to become a potentially dangerous environment.

Sensitive-Content Screens

The Beginnings

Instagram have always maintained that they have tried to make Instagram a *safe* place for everyone. In 2017, they strengthened this commitment, announcing new policies for "sensitive" content (Systrom, 2017). Specifically, Instagram began *screening* images (and videos) that users reported as sensitive (and content moderators subsequently confirmed as such) but that did not violate their community guidelines. The exact nature of what content Instagram screened or not was relatively unclear at this time. Notably though, at the time of introduction, Instagram's *sensitive-content screens* had two hallmark features. First, the screens obfuscated sensitive images via an image processing technique called Gaussian blur; the resulting image had reduced noise (e.g., variations in brightness or color) and detail, which made it difficult for people to determine exactly what the image depicted. Second, the screens included a warning message (e.g., "Sensitive Content: This photo contains sensitive content which some people may find offensive or disturbing"), a form of *trigger warning*—a statement intended to help people prepare for or avoid content likely to trigger memories or emotions relevant to past experiences (Bridgland et al., 2023). In their announcement, Instagram explained that such screening would balance out *their need* to create a safe space for people to talk about their experiences—including self-injury and post related non-graphic content online—with *their responsibility* to reduce the potential harm that such content might have on other people, especially "vulnerable" people who may see it (Mosseri, 2019a). In this case, it appears that Instagram operationalises vulnerable people as people with more severe psychopathological symptoms (e.g., of depression)—but they do not explicitly define the group. The idea was that screening sensitive content removes the surprise people may experience coming across such content "unprepared", whilst also giving people an

opportunity to completely avoid the content, or if they want to view it, an opportunity to emotionally prepare for it. Therefore, sensitive-content screens were originally intended as a harm minimisation tool.

In 2019, Instagram made further changes to their policies after investigations revealed that Molly Russell—the 14-year-old who died by suicide—had saved, liked or shared 2,100 pieces of content related to suicide, self-injury and depression in the six months before she died—some of which the algorithm recommended to her (Naughton, 2022). At the conclusion of the inquest into Molly's death (that came later in 2022), a coroner ruled that Molly died from "an act of self-harm while suffering from depression and *the negative effects of online content*". This ruling was significant—because it was the first time that social media platforms were held formally responsible for the death of a child (Molly Rose Foundation, 2023)—but the circumstances of Molly's death are far from unique. In the United Kingdom, exposure to suicide-related content was reported in 24% of deaths by suicide among young people ages 10 to 19 between 2014 and 2016 (Rodway et al., 2023). Nonetheless, in response to the ongoing scrutiny that followed Molly's death specifically, Instagram prohibited all graphic self-injury related images (including fictional depictions), removed non-graphic self-injury related content from hashtag searches, and stopped the algorithm from recommending sensitive content to users (although at this time the content was still available when users searched for it; Mosseri, 2019a, 2019b). Instagram also added sensitive-content screens more *broadly* to sensitive content across Instagram—but again, the exact nature of what content they screened or not remained relatively unclear at this time.

Current Guidelines and Processes

Sensitive-content screens, and warning systems alike, are now used across most social media platforms, including Facebook, TikTok and Reddit—to name a few. Instagram's sensitive-content screens still have their original hallmark features: the forewarned image is blurred and accompanied by a warning message, though Instagram have changed the wording

of the message several times. In 2024, the message usually warns about the type of content (i.e., "Sensitive Content: This photo may contain graphic or violent content"), but in some instances, it still warns about the reactions people may have to the forewarned content—per the original warning. Culturally, there are debates regarding the type of sensitive content that should have a warning. Although some people have called for warnings to be added to anything that may be potentially distressing (e.g., material that depicts exclusion and oppression; Johnson et al., 2015), others have argued that shielding people from such content is unnecessary and possibly even harmful (e.g., because people may not learn how to tolerate emotional discomfort, and awareness of important issues, such as suicide, may be hindered; Lukianoff & Haidt, 2015; Filipovic, 2014). More broadly speaking, we also know that "triggers" are complex and closely connected to personal vulnerabilities (Riachi et al., 2022), so what one person may find sensitive (or triggering) may be very different to another. Understandably then, there remains some ambiguity and confusion about what sensitive content really is, and thus, what type of content we should expect to see beneath a sensitive-content screen.

Meta, the company that has owned Instagram since 2012, has policies and community standards (available at the Transparency Center: https://transparency.fb.com/en-gb/) that clearly and comprehensively define what content is and is not allowed on their platforms (change logs reflecting frequent policy updates are also available). According to their current policies, Meta removes content from Instagram that is "particularly violent or graphic" such as *videos* depicting dismemberment, visible internal organs, or charred bodies, as well as content that encourages suicide or non-suicidal self-injury (including content related to eating disorders, and fictional content such as memes or illustrations). But they recognise that people are differently affected by violent or graphic imagery; they also note the importance of allowing people the space to share their experiences, raise awareness, and seek support from one another. Therefore, they allow certain types of sensitive content on

Instagram. For example, they allow *images* depicting dismemberment, visible internal organs, or charred bodies, and content depicting older instances of non-suicidal self-injury (e.g., healed scars or other non-graphic self-injury imagery) in a *recovery* context (Meta, 2023)—though what might appear recovery focused, according to Meta, may be perceived otherwise by users. Notably, allowed sensitive content is covered by a sensitive-content screen. Thus, Instagram follows a narrow set of community standards that stipulate whether content is to be removed or screened based on the perceived risk of viewing a *single* piece of content.

The processes by which Meta identifies sensitive content on Instagram have also evolved with the increasing popularity of the platform. With more active users than ever before (~2.00 billion users monthly*;* Kemp, 2023), billions of pieces of content are posted online every day (Meta, 2023). Previously, content moderators individually reviewed all the potentially violating content reported by users, but with increases in the volume of content, artificial intelligence now detects and acts on violating content, often before anyone reports it. This process includes removing content, adding sensitive-content screens to photos or videos that may be distressing, and/or disabling accounts. Artificial intelligence also detects *potentially* violating content (e.g., when the sentiment of a post is unclear, or the content is context-dependent) and sends it to content moderators for further review. Data available on the Transparency Center suggests Meta's current processes are effective in quickly detecting and acting on violating (and potentially violating) content. In fact, Meta acted on 6.2 million pieces of "violent and graphic content" from Instagram between April and June 2023—97.5% of which was acted on *before* users reported it. However, given the threshold for *removing* content is arguably high, it is likely much of this action resulted in the screening of sensitive images, rather than their complete removal. Indeed, when researchers scraped the data on Instagram (in October 2023) they found substantial amounts of potentially harmful content (45% of which

had more than 1000 likes), including posts that promote and glorify suicide and self-injury (including eating disorders), actively reference suicide ideation, and contain intense themes of misery, hopelessness, and depression (Molly Rose Foundation, 2023)—which are all seemingly in violation of the community standards. Worryingly, a recent *eSafety Commissioner* report (Australian Government, 2022) revealed that almost two-thirds of young people aged 14–17—who are some of the most vulnerable users on Instagram—had been exposed to this type of sensitive content over the past year. Therefore, despite Meta's efforts to reduce the accessibility of such content, users still have access to a considerable amount of sensitive content on Instagram, suggesting there are significant ongoing issues with Meta's current guidelines and policies, and/or the way they are implemented.

Aside from removing or adding sensitive-content screens to sensitive content, Instagram has recently also changed the way they allow people to use hashtags, and restricted the type of content they recommend to users (Instagram, 2024). Instagram now hides content when people, especially young people (i.e., <18), search for terms related to suicide, self-harm and eating disorders (e.g., #suicide), and instead directs them to resources (e.g., helplines) for support. However, we do not know that providing resources for supports means that people will opt to use them. Indeed, people can easily click past the "help is available" pop-up to view the sensitive content—even though the "continue to search results" button is arguably inconspicuous (i.e., the text is small and located at the bottom of the page). Another issue with this process is that new hashtags (e.g., #suicidalll) begin to emerge as others are restricted, which gives people access to the same sensitive content—albeit in a more secret community (Fulcher, 2020; Molly Rose Foundation, 2023). Instagram now also aims not to recommend sensitive content to users, but to date, there remains some concern regarding their success in achieving this aim (e.g., Molly Rose Foundation, 2023). Instagram has also introduced other features to increase the control people have over the content they see on their feeds. In 2021, they introduced

the “Sensitive Content Control” setting, which allows users to see “more” or “less” sensitive content than the standard setting (as of early 2024 people under 18 were automatically set to the most restrictive setting; Instagram, 2021, 2024). Although this feature was seemingly developed with users’ wellbeing in mind, it assumes that *if* people are given an opportunity to avoid unpleasant experiences—or in this case, to opt to see less sensitive content on their feeds—they will take it. However, we know that sometimes people do not behave in a manner likely to increase their experiences of pleasure (e.g., Hsee & Ruan, 2016; Oosterwijk, 2017). Therefore, it is unlikely that every user will opt to see less sensitive content, despite having the opportunity to do so. Thus, Instagram seemingly relies primarily on sensitive-content screens to protect users from the sensitive content that remains on the platform.

1.2 Advocates’ Claims from an Emotion Regulation Perspective

The idea that sensitive-content screens *protect* users from sensitive content mirrors the claims advocates often make about traditional trigger warnings (e.g., Manne, 2015; Lockhart, 2016). Broadly speaking, advocates claim that trigger warnings can influence people’s emotions (e.g., by reducing their experiences of sadness or fear). To explore whether the existing evidence supports advocates’ claims, I first examine what an emotion is, the processes involved in emotion regulation, and how people’s engagement with sensitive content may involve a range of emotion regulation strategies. I then examine advocates’ claims from an emotion regulation perspective.

Emotions and Emotion Regulation

At a basic level, emotions can be positively (e.g., happiness) or negatively (e.g., sadness) valenced (Bradley & Lang, 2007), and encompass experiential (e.g., feelings of anxiety), behavioural (e.g., urge to escape), and physiological (e.g., increased heart rate; Evers et al., 2014) reactions. *Emotion regulation*, therefore, refers to the automatic or controlled processes by which people influence which emotions they have, when they have

them, and how they experience and express them (i.e., the emotions' intensity, duration, and quality; Gross, 1998). Traditional *hedonic* accounts of emotion regulation claim that people are motivated to decrease (or down-regulate) negative emotions and increase (or up-regulate) positive emotions (e.g., Larsen, 2000). A common misunderstanding is that emotion regulation encompasses *only* these hedonically driven attempts to influence emotions; but the term, broadly speaking, refers to efforts to achieve any emotion *goal* (or desired end-state; Tamir, 2016), including negative emotions. Indeed, there are some situations in which instrumental goals—that is, goals that lead to delayed rather than immediate reinforcement (e.g., to make meaning of a traumatic experience; Tamir, 2016)—motivate *counterhedonic* regulation, such that people seek to up-regulate negative emotions and/or down-regulate positive emotions.

The Process Model of Emotion Regulation

To achieve an emotion *goal* (or desired end-state; Tamir, 2016), people employ a variety of emotion regulation strategies. According to the *process model of emotion regulation* (Gross, 2015)—the most dominant model in the field—there are five families of emotion regulation strategies that can occur during difference stages of the emotion experience: (1) situation selection, (2) situation modification, (3) attentional deployment, (4) cognitive change, and (5) response modulation. Notably, people's engagement with sensitive content on social media platforms may involve a range of these emotion regulation strategies. Indeed, when people decide to view sensitive content (or not) they demonstrate a form of situation selection—which refers to choosing situations based on their likely emotional impact. For example, people whose emotion goal is sadness may deliberately seek out sensitive content as a means of experiencing sadness. While viewing sensitive content, people can change its emotional impact by directly altering the situation (situation modification), intentionally directing their attention in the situation (attentional deployment), or by reappraising the situation altogether (cognitive change). For example, people may

choose to view a sensitive image for an extended period or move quickly to the next image, focus their attention away from unpleasant aspects of the image (e.g., deceased person) and towards more pleasant aspects (e.g., the green grass on which the deceased person lies), or think "it's only a photo" rather than seeing the situation depicted in the image as a real-life event. Therefore, depending on which strategies people employ, and to what degree, they may experience more or less of the emotions they otherwise would have experienced (i.e., without such regulation strategies). Furthermore, even once people experience emotions related to the sensitive content (e.g., sadness), they can influence how they experience and express them; for example, by holding back tears, or maintaining a neutral facial expression to hide their sadness (i.e., expressive suppression; Gross, 2015)—which is a form of response modulation.

Advocates' Claims

Turning now to the claims advocates make about trigger warnings, with emotion regulation in mind; I explore the two key claims in turn. First, advocates claim trigger warnings deter people from viewing sensitive content by giving them an opportunity to avoid it—when the choice to employ situation selection may otherwise not be afforded (e.g., if they came across such content unexpectedly). Advocates claim that people with trauma histories especially, should be able to decide if they want to avoid content that may trigger re-experiencing symptoms (e.g., intrusive memories of the content), arguing that such avoidance can aid recovery (Cripps, 2020). Paradoxically though, avoidance is also one of the purported harms of trigger warnings. Specifically, critics (e.g., Lukianoff & Haidt, 2015) argue that encouraging avoidance is ultimately harmful—despite any temporary relief it may provide. Indeed, avoidance is a primary maintaining factor in post-traumatic stress disorder (PTSD; Badour et al., 2012) as well as a central characteristic of a broad range of mental disorders (e.g., anxiety disorders; Krypotos et al., 2015).

Second, advocates claim that trigger warnings allow people time to emotionally prepare for sensitive content (e.g., by employing anxiety management techniques, such as meditation; Manne, 2015). The idea here is that such preparation could increase the likelihood that people engage in emotion regulation strategies (e.g., attentional deployment or cognitive change; Gross, 2015) that help people better manage (or reduce) their negative emotions while viewing the sensitive content. Advocates claim that it is particularly important for vulnerable people (e.g., people with trauma histories) to emotionally prepare to view sensitive content (e.g., Manne, 2015); presumably, the idea here is that vulnerable people are at greater emotional risk when viewing such content unprepared. However, critics argue that trigger warnings "coddle" people (e.g., by sheltering them from the "real world"; Lukianoff & Haidt, 2015), and may create a false impression that experiencing trauma *always* has long-lasting negative emotional impacts (e.g., Bellet et al., 2018). In reality though, most people who experience trauma are resilient and show few symptoms of PTSD after an initial period of adjustment (Breslau & Kessler, 2001).

Taken together, advocates' claims as well as critics' responses raise a number of important issues. For example, should people have the right to decide what content they would like to engage with, even if it inevitably causes them harm (e.g., by maintaining their PTSD or sheltering them from the "real world")? And does the impact that such warning systems may have on societal perceptions of recovery from traumatic events, and mental disorders (e.g., PTSD) more generally, counteract their purported benefits? These issues parallel with a known conundrum in ethics debates related to warning use: balancing concerns over non-maleficence (i.e., not causing harm) and the right to autonomy (i.e., to make an informed decision, e.g., Stirling et al., 2022). However, these issues are beyond the scope of my thesis. My thesis, broadly speaking, aims to examine the empirical basis of the two key claims made by advocates, specifically in relation to sensitive-content screens—which we know are a form of trigger warning. Thus, I now explore existing theory and

related literature, as well as the evidence (thus far) for advocates' two key claims, beginning first with the claim related to deterrence.

1.3 Do Sensitive-Content Screens Deter People from Viewing Sensitive Content? Predictions Based on Existing Theory and Related Literature

Uncertainty and Curiosity

There is existing theory and related literature that suggests sensitive-content screens may *increase* engagement with—rather than deter people from—sensitive content. By design, sensitive-content screens are ambiguous; they warn of content that "may contain graphic or violent content" but do not provide any information about the exact type or nature of the content. Coming across such screens then, people likely wonder about what the content might be; for example, whether it is something they want to see. According to the information-gap hypothesis (Loewenstein, 1994), when people perceive a gap in knowledge—that is, when what *they want to know* exceeds their current level of knowledge—they experience feelings of deprivation. These feelings are aversive and motivate people to obtain information to eliminate, or at least reduce, their perceived gap in knowledge (Loewenstein, 1994). Thus, in such a situation, people may uncover sensitive-content screens to get more information about the content beneath the screen—even though the warning indicates it may be distressing.

In fact, the "Pandora effect" suggests that in some cases people seek to resolve curiosity *even though*, and in some cases *because*, doing so will have a negative (or aversive) effect (Hsee & Ruan, 2016; Yagi et al., 2023). In one series of experiments, people were *more* likely to engage with stimuli (e.g., open a box) if the consequences of doing so were uncertain (vs. certain) and negative (vs. neutral) in nature (e.g., electric shocks, unpleasant sounds, and disgusting images; Hsee & Ruan, 2016). Indeed, although people may come to regret such decisions—perhaps when they experience the negative (or aversive) effects—the desire to resolve curiosity (even under uncertain conditions) is seemingly more important

than regret aversion (Van Dijk & Zeelenberg, 2007). For some people, knowingly engaging in such risky behaviour may be driven by sensation seeking—a personality trait that is defined broadly as the willingness to take risks for the sake of novel and intense experiences (Zuckerman, 2007). Thus, warning of negative (and potentially distressing) content is unlikely to deter people from viewing sensitive images, especially for people who are high sensation seekers.

We also know that some people are morbidly curious, such that they intentionally seek out highly negative information (Oosterwijk, 2017). In one series of experiments, people willingly viewed images that portrayed death, violence, or harm when they had the option to instead view a neutral alternative (e.g., images of household items, plants, buildings; Oosterwijk, 2017). Although it seems counterintuitive to intentionally seek out negative (and potentially distressing) content, we know that some people experience pleasure from these experiences; the popularity of horror movies and true crime shows (e.g., Netflix's *Making a Murderer*; Bonn et al., 2016), as well as the interest people show in news coverage of violence and terrorism, supports this notion. It has been theorised that viewing such content gives people an opportunity to experience difficult emotions (e.g., fear and sadness) in a safe, contained and chosen environment (e.g., Princing, 2021)—which may have a therapeutic benefit, akin to exposure-based interventions (e.g., exposure therapy; Abramowitz et al., 2019). We also know that negative content may offer stronger informational gain than positive or neutral information because of its unique (and sometimes, socially deviant) nature (Oosterwijk, 2017). For example, such information may help people build a realistic understanding of the world (Baumeister et al., 2001), and/or prepare people for negative events through the experience of others (Bartsch & Mares, 2014; Hoffner et al., 2009). Therefore, labelling negative content like sensitive-content screens do may assist people, and especially people who are morbidly curious, with readily identifying (or locating) highly negative content, and thus increase its accessibility.

The Forbidden Fruit and Boomerang Effects

Restricting access to content, like sensitive-content screens do, may also increase its attractiveness, and thus exacerbate the likelihood that people *intentionally* seek it out. The "forbidden fruit effect" occurs when an experience or behaviour becomes more attractive to people *because* their freedom to engage in that experience or behaviour is restricted (or is perceived as being off-limits; Weaver, 2011). This effect has been found across contexts; for example, warning labels (vs. no label) on cigarette packages increase existing smokers desire to smoke cigarettes (Hyland & Birrell, 1979), dieters (vs. non-dieters) experience stronger cravings for restricted foods (Massey & Hill, 2012), and warnings (e.g., "Viewers discretion is advised") and age-restrictive labels (e.g., "MA 15+") increase people's desire to watch violent television programs (Bushman, 2006) and play violent video games (Bijvank et al., 2009). In some cases, restricting people's freedom to engage in an experience or behaviour can also create a "boomerang effect" (Brehm, 1966)—whereby people intentionally engage in the restricted experience or behaviour. For example, when the United States increased their drinking age from 18 to 21, newly underage college students (i.e., whose freedom to drink was now restricted) consumed more alcohol than adult students—a pattern that had not previously been seen (Engs & Hanson, 1989). Such behaviour can be understood in reference to psychological reactance theory (Brehm, 1966; for review see Rosenberg & Siegel, 2018)—which posits that people's freedom of behaviour is important, and if threatened or eliminated, people are motivated to restore it. Therefore, sensitive-content screens may not only be ineffective at deterring people from viewing sensitive content, but they may also increase the likelihood that people intentionally seek out such content.

Notably, marketing teams appear to realise the impact that such warning systems may have on behaviour. In recent years there have been many advertisements for food and alcohol products (e.g., by Pizza Hut, McDonald's, and White Claw) that bear a striking resemblance to Instagram's sensitive-content screens. For example, in 2019, Pizza Hut released a new

menu item: the advertisement was blurred and came with a warning message, “Sensitive Content: This post contains sensitive content which some people may find delicious and irresistible” (Kobach, 2019). These advertisements are used to *maximise* consumer attraction and engagement (the “teasing effect; Ruan et al. 2018), which aligns with what we know from existing theory and related literature about behaviour following such warnings. Taken together, all this information begs the question, why do social media platforms expect sensitive-content screens to deter people from viewing content beneath screens?

Deterrence and Vulnerable Populations

Of particular concern, vulnerable people may be *least* deterred by sensitive-content screens. Although they are the very people sensitive-content screens were originally designed to protect, we know that vulnerable people often engage with sensitive content. For example, people experiencing a major depressive episode, or persistent depressive disorder (American Psychiatric Association, 2022) have been found to expose themselves to more negative (and sadness-inducing) images than participants without such depressive symptoms; they are also more likely to listen to sad music, compared to happy or neutral music (Millgram et al., 2015). Indeed, in line with these behaviours, people with major depressive disorder demonstrate a greater desire for sadness, compared to participants without depression (Arens & Stangier, 2020). Furthermore, some people with a history of trauma engage in self-triggering—which, put simply, refers to the act of intentionally seeking out reminders of the traumatic experience(s) (e.g., by exposing themselves to related graphic images or media coverage; Bellet et al., 2020). One study, for example, found that people with prior lifetime exposure to violence are more likely to seek out and watch disturbing content online, such as the graphic ISIS beheading video (Redmond et al., 2019). In fact, such behaviour is associated with PTSD symptom severity (Bellet et al., 2020)—which suggests self-triggering may exacerbate PTSD symptom severity and/or that people with worse PTSD symptom severity may be more likely to self-trigger.

There are also online recovery communities—where users post content and engage, socially, with one another—which seemingly facilitate access to sensitive content. For example, people with a tendency to engage with non-suicidal self-injury and people experiencing eating disorders have been found to seek out sensitive content through such online communities (e.g., Fulcher et al., 2020; Wang et al., 2018). Although this behaviour may provide benefits related to social connectedness (e.g., Park et al., 2022; Juarascio et al., 2010), it is in some ways akin to self-triggering, and may lead to harmful consequences (e.g., increases in non-suicidal self-injury and eating disorder symptomology; Arendt et al., 2019; Feldhege et al., 2021). In fact, in one study, one-third of participants engaged in the same or similar types of non-suicidal self-injury after viewing it on Instagram (Arendt et al., 2019). There is also a risk that such behaviour may become a normalised coping mechanism, and that such online communities may preclude people from seeking out professional support by establishing a sense that such professionals "would not understand" their difficulties (e.g., Lavis & Winter, 2020). Although it is necessary to discuss how people with a tendency to engage with non-suicidal self-injury and people experiencing eating disorders may engage with sensitive content—given that content related to non-suicidal self-injury and eating disorders is commonly found (and screened) on Instagram—my thesis is primarily focused on the former vulnerable populations (i.e., people with symptoms of depression, PTSD, and/or who self-trigger), and on sensitive content more generally. Therefore, my literature review will not specifically discuss these populations further.

Thus, returning to vulnerable users in general, why might they be susceptible to seeking out sensitive content—even in the presence of sensitive-content screens warning them of possible distress? Perhaps it is to obtain information or gain insight into their own feelings and situations—for example, to make meaning, and/or to improve their current feelings or situation (e.g., through social comparison or problem-focused coping; Reinecke et al., 2016). Indeed, the desire to make meaning of a traumatic experience was the best

predictor of how often participants self-triggered (Bellet et al., 2020). Additionally, vulnerable people may deliberately seek out negative content as a means of gaining (a sense of) control and predictability of their psychopathological symptoms. For example, some people with depression and PTSD deliberately try to maintain their negative emotional states, possibly because those states are familiar and they want to avoid contrasting emotional states and/or unexpected shifts in symptomology (e.g., Bellet et al., 2020; McGhie et al., 2022; Millgram et al., 2015). Indeed, a preference for avoiding contrasting emotional states is established in other clinical disorders (e.g., generalised anxiety disorder). More specifically, the Contrast Avoidance Model postulates that some people deliberately engage in negative thinking (e.g., worry or rumination) to perpetuate a negative mood and thereby avoid the shift from positive or neutral moods into negative moods (Crouch et al., 2017).

Taken together, labelling sensitive content, in the way that sensitive-content screens do, may not deter vulnerable people from viewing it. In fact, somewhat counterproductively, sensitive-content screens may assist vulnerable people with accessing such content.

The Evidence

Trigger Warnings

Now, I turn to the existing evidence on deterrence; first, I draw on the trigger warning literature to understand how sensitive-content screens might operate, specifically with respect to deterrence. Notably, only a handful of studies have explicitly examined whether people are deterred by trigger warnings, and broadly speaking, these studies show that participants who receive a warning are no more likely to avoid forewarned content (e.g., films and images) than participants in a control condition (i.e., who do not receive a warning). For example, Gainsburg and Earl (2018) found no difference in how often participants selected to watch films based on titles accompanied by a trigger warning or not. Similarly, Bridgland and Takarangi (2021) found that participants did not passively (by remaining on an instruction screen) or actively (by covering images) avoid negative content prefaced with a trigger

warning any more than content without a warning—apart from a minor increase in avoidance when a warning appeared in the first few trials. Indeed, a recent meta-analysis found that trigger warnings have a negligible effect on *avoidance*—defined broadly as the act of bypassing or otherwise blocking exposure to content (e.g., choosing to skip content entirely, engaging with alternative, non-distressing content, or dropping out from the experiment following the warning; based on 11 unique effect sizes; Bridgland et al., 2023). In fact, in one study—which was an outlier in the meta-analysis—articles were selected *more* often when they carried a warning (Bruce & Roberts, 2020), suggesting warnings possibly *increase* engagement with sensitive content. Notably though, avoidance was operationalised in a slightly different way compared to other studies; rather than randomising to a single warning or no warning condition, in this study, participants were asked to choose between four article titles, two with trigger warnings (i.e., "Trigger Warning: Sexual Abuse") and two without. Having four article titles available, within-subjects, may have changed participants' behaviour: for example, by increasing their relative curiosity for the article titles with a trigger warning (vs. without). Indeed, we know that people typically experience an initial increase in curiosity as information increases (e.g., via such trigger warnings), and that such curiosity is characterised by exploration, and approach-driven behaviour (Day, 1982)—similar to that observed in the aforementioned study.

Furthermore, trigger warnings do not appear to deter vulnerable people from viewing negative content either. Specifically, research has found no evidence of increased avoidance of forewarned content among vulnerable participants (e.g., people with a history of trauma; Bruce & Roberts, 2020; Kimble et al., 2021)—including participants with prior experiences *related* to the forewarned content (Bridgland & Takarangi, 2022). In one study though, certain individual differences (e.g., the belief that trigger warnings are protective) were associated with increased avoidance of forewarned content (Gainsburg & Earl, 2018); but there was no indication that these individual differences were related to vulnerabilities. Taken

together, extant literature suggests that there *may* be some important individual differences, but largely, trigger warnings do not appear to deter people, including vulnerable people, from viewing sensitive content.

Sensitive-Content Screens

The evidence on whether sensitive-content screens, specifically, deter people from viewing sensitive content is even more limited (than that on traditional trigger warnings). In a preliminary investigation of sensitive-content screens, participants saw a single sensitive-content screen and reported how likely they would be, hypothetically, to uncover it (Bridgland, Bellet et al., 2022; Study 1). In line with the idea that sensitive-content screens may increase the attractiveness of, and inadvertently increase engagement with, sensitive content, 80% of participants said they *would* uncover the screen (to view the sensitive content underneath). That is, few participants said they would be deterred from the sensitive content. Furthermore, the intention to uncover the screen was associated with poorer wellbeing, higher depression symptoms and experiential avoidance (i.e., tendency to avoid thoughts), and lower perceived life meaningfulness—all factors that are in turn linked to a range of psychopathologies (Beck, 2009; Kashdan et al., 2006; Keyes et al., 2010). Thus, there is some evidence to suggest that vulnerable people *may* be more likely, relative to people with less severe psychopathological symptoms, to seek out sensitive content. In a follow-up study examining participants' actual uncovering behaviour, rather than their intentions to uncover, participants viewed a single sensitive-content screen—presented after neutral and positive images—and had the option to "uncover" it (though the negative image was not shown; Bridgland, Bellet et al., 2022; Study 2). In line with intentions in the first study, most people (~85%) chose to uncover the screen. However, unlike the first study, the decision to uncover the screen was not associated with vulnerability characteristics (e.g., higher depression symptoms). Put differently, vulnerable people were no more likely to seek out sensitive content, but they were also no more likely to use the screens to avoid sensitive content.

Across both studies (Bridgland, Bellet et al., 2022), participants also reported what factors *would*, hypothetically (Study 1), or *did* (Study 2) influence their decision to uncover the sensitive-content screen. In Study 1, where participants reported their intentions to uncover, 35.8% of participants reported they simply wanted to see the image, and of these, 75.3% specifically mentioned reasons related to curiosity—consistent with the information-gap hypothesis (Loewenstein, 1994), the "Pandora effect" (Hsee & Ruan, 2016; Yagi et al., 2023), and morbid curiosity (Oosterwijk, 2017). Other participants (36.2%) indicated they would decide whether to uncover sensitive-content screens based on the context of the image (e.g., posting account, and content descriptions). In Study 2, where participants decided to uncover a sensitive-content screen or not, curiosity remained the primary reason for uncovering (46.2%). Other participants (10.7%) mentioned they would decide to uncover or not based on their ability to cope with distressing content—a form of situation selection informed by a person's anticipated future emotion states, and their belief in their capacity to manage the distress (Gross, 2015).

Taken together, this preliminary work supports the idea that sensitive-content screens may not deter people, including vulnerable people, from viewing sensitive content—at least for *one* image. There also appears to be several possible reasons *why*—some of which align with existing theory and related literature. I now explore existing theory and related literature, as well as the evidence (thus far) for advocates' second key claim related to emotional preparation.

1.4 Do Sensitive-Content Screens Emotionally Prepare People to View Sensitive Content?

Predictions Based on Existing Theory and Related Literature

Recall, emotional preparation—in terms of advocates' claims—refers to the idea that people can employ some kind of strategy *before* viewing the content, to better manage (or reduce) their negative emotions *while* viewing the forewarned content. The existing theory

and related literature provide a mixed account of whether sensitive-content screens emotionally prepare people to view sensitive content.

Bracing for the Worst

Bracing for the worst, or simply bracing, is one example of emotional preparation that sensitive-content screens may elicit. Bracing involves people intentionally and strategically managing their expectations; the idea of bracing is that by expecting an unfavourable outcome people can avoid experiencing negative emotions while they view the forewarned content (see Moeck, 2023). Indeed, people report bracing for this reason (Sweeny & Falkenstein, 2015), and do so in everyday life while waiting for a range of potentially negative outcomes (e.g., exam grades, and medical tests; Shepperd et al., 1996; Sweeny & Cavanaugh, 2012). The bracing literature relates more specifically to anticipating something negative over a longer period of time (i.e., being notified of a potential stressors, waiting for an outcome, and then receiving the outcome), whereas sensitive-content screens provide an almost immediate outcome by comparison. Nonetheless, the bracing literature provides insight into the processes at play when we experience an outcome subsequent to our expectations.

Existing theories provide a conflicting account about the likely outcomes of bracing. Decision affect theory (Mellers et al., 1997) posits that people's emotional responses to situations are influenced by comparing actual outcomes with what could have been. For example, if a student expects to do poorly on their exam and they do well, they exceed their expectation and experience happiness; whereas, if a student expects to do well on their exam and they do poorly, they fail to meet their expectation and experience disappointment. Thus, lowering expectations (or bracing) may help people pre-emptively avoid disappointment (van Dijk et al., 2003). However, in contrast, the affective expectations model posits that expecting a negative outcome *heightens* negativity when the expected outcome occurs (Wilson et al., 1989). For example, if people expect to feel disappointed, and a situation fails

to meet their expectations, then they are likely to experience *more* disappointment than they otherwise would have (had they not expected such an outcome). Therefore, according to existing theories, bracing may either up-regulate positive emotions and down-regulate negative emotions *or* simply, make people feel worse. But does bracing actually help?

Existing evidence provides partial support for decision affect theory but suggests that the emotional benefit of bracing is *immediate*, yet short-lived, after the outcome of an anticipated situation is known. For example, students who expect to do poorly on exams and receive bad grades, feel better immediately after receiving their grades, compared with students who expect to do well but receive bad grades (Sweeny & Shepperd, 2010). *But* regardless of whether they expected to do poorly or not, all students feel the same 24-hours later (Lench et al., 2021). However, somewhat counterproductively, bracing appears to have negative impacts during the *anticipatory* period (i.e., before the outcome is known). Specifically, bracing elicits an immediate negative psychological impact (e.g., negative affect; Golub et al., 2009; Sweeny et al., 2016)—which in some cases, lasts two to three hours before the outcome is known (Neubauer et al., 2018)—as well as negative *physiological* impacts (e.g., increased blood pressure; Spacapan & Cohen, 1983). Notably, in one study, higher levels of anticipatory negative affect (in advance of receiving exam results) was associated with negative affect at 5-month follow-up—suggesting there may also be long term consequences of such a noxious anticipatory period (Kalokerinos et al., 2022; Study 1). Therefore, bracing makes people feel worse as they wait—and in some ways, may be akin to experiencing the actual situation—with little to no benefit after the outcome is known. Anticipatory negative affect may also have long term consequences in and of itself. Thus, preparing to view sensitive content by bracing may have negative to null effects—and although null effects are not harmful per se, relying on bracing may come at the expense of using other evidence-based strategies (e.g., cognitive emotion regulation strategies; Gross, 2015) that could provide people with an emotional benefit.

Nocebo Effects

While bracing for forewarned content, people may also begin to anticipate their responses; for example, how they might feel while they view the forewarned "graphic or violent" content (e.g., "I will feel distressed"). Such emotional preparation—while a purported benefit of sensitive-content screens, according to advocates' claims—may counterproductively increase the likelihood that people experience the very negative outcomes sensitive-content screens warn of (e.g., distress). This phenomenon is known as the "nocebo effect", which put simply refers to the tendency for negative outcomes to occur when people expect them (Hanh, 1997; Rooney et al., 2022). Indeed, a growing body of literature provides evidence for nocebo effects across a diverse range of health outcomes, from experimentally induced pain/itches to Parkinson's disease (e.g., Bartels et al., 2016; Keitel et al., 2013). In one example, participants given information about pain before an upcoming injection (e.g., "This is the worst part of the procedure") reported markedly worse pain immediately following the injection, relative to participants who received the same injection without information about pain (Varelmann et al., 2011).

Existing theories tend to agree that *expectancy* is the primary mechanism driving the nocebo effect, but there are several different explanations for how expectancies result in nocebo effects. Response expectancy theory posits that people receive information from their environment (e.g., via other people) regarding the likely outcomes of an event (e.g., pain following an injection) and begin to anticipate, and then internally generate, their anticipated outcomes (e.g., pain; by [subconsciously] changing their behaviour; Kirsch, 1997). Consequently, their experiences of the event, and in some cases, physiological functioning, are altered in line with their expectations. Barsky et al. (2002) offer a different account; specifically, they posit that negative expectancies increase attention to the event and create state anxiety—which causes people to over-attend to negative information (e.g., increasing pain following an injection) and unfavourably interpret ambiguous situations (e.g., pre-

existing body tension)—thereby worsening outcomes. Thus, although both theories agree that expectancy is the primary mechanism, the former posits that expecting a negative outcome can worsen the *actual* outcome, whereas the latter includes attention and state anxiety as mediating variables. Notably, both theories have empirical support; a recent meta-analysis on nocebo effects (of 59 studies with varying study designs and health outcomes) found strong evidence for the role of expectancy, and some evidence for the role of state anxiety (Rooney et al., 2022).

Taken together, sensitive-content screens may elicit nocebo effects, which may induce (rather than reduce) negative emotions (e.g., distress) while people view the forewarned content. Notably, while bracing is an active process by which people intentionally and strategically manage their expectations, nocebo effects can occur without such awareness. Therefore, it may be difficult for people to overcome nocebo effects once their expectations have been shaped by sensitive-content screens.

Priming Effects

Sensitive-content screens may also influence the emotional reactions people have towards forewarned content by eliciting priming effects. In one example study testing priming effects outside of the warning context, participants were primed with either negative (e.g., mean, selfish, rude) or positive (e.g., sincere, creative, wise) trait adjectives, and were shown an ambiguous image of a person: participants primed with negative traits rated the person higher on these negative traits (Ferguson et al., 2005). Therefore, it is possible that viewing sensitive-content screens may prime a negative mindset and cause people to interpret—and therefore, respond to—subsequent content in a more negative way than they otherwise would have (i.e., without the screens). As with nocebo effects, such priming effects can occur without awareness, and the resulting emotions—which are likely negative in this situation—may be difficult to down-regulate once they are fully formed (Gross, 2015).

Emotion Regulation Efforts

Sensitive-content screens may not only elicit problematic expectancies (which may worsen how people feel in the anticipatory period and cause nocebo and priming effects), but they may also directly influence people's efforts to down-regulate negative emotions *while* viewing forewarned content. Specifically, anticipatory information about the nature of upcoming content may interfere with people's ability to use attentional deployment and cognitive change effectively (Shafir & Sheppes, 2018, 2020). Indeed, knowing that an image is graphic or violent may cause people to focus on the unpleasant aspects of the image (e.g., deceased person), rather than the more pleasant aspects (e.g., the green grass in which the deceased person lay), or see the image as a real-life event, rather than thinking "it's only a photo". Thus, although advocates claim that sensitive-content screens assist with emotional preparation, it is possible that the screens make it more difficult for people to effectively manage their emotions. Consequently, people may experience more, rather than less, negative emotion while viewing the forewarned content.

Emotional Preparation and Vulnerable Populations

Vulnerable people may also be the *least* likely to emotionally prepare for sensitive content. As I have already discussed, vulnerable people often deliberately seek out negative content, and do so for several reasons (e.g., to maintain negative emotional states)—many of which conflict with advocates' ideas about emotional preparation (i.e., using strategies to reduce subsequent negative emotions). Thus, although vulnerable people (e.g., people with depression; Koval et al., 2012) may become "stuck" in negative affect by repeatedly engaging with sensitive content—a concept termed emotional inertia (e.g., Kuppens et al., 2010)—they may not want to repair their affect.

Additionally, even *if* vulnerable people *want* to emotionally prepare for the forewarned content, they may find it especially difficult to do so. Sensitive-content screens in their current format do not include instructions on *how* to emotionally prepare; therefore,

social media platforms employing such screens assume that people can spontaneously implement emotion regulation strategies. But vulnerable people may have few strategies to choose from—potentially because they tend to rely on certain strategies (e.g., avoidance of reminders of their trauma experiences). Vulnerable people may also maladaptively weigh the cost and benefits associated with using strategies, and thus select inappropriate strategies. For example, people who engage in non-suicidal self-injury value self-injury as an effective means of regulating negative emotions (e.g., to avoid or suppress negative feelings; McKenzie & Gross, 2014), even when the short-term relief associated with avoiding such feelings can come at a longer-term cost (e.g., by increasing negative feelings, such as shame; Gunnarsson, 2021). Aside from potentially selecting inappropriate emotion regulation strategies, vulnerable people may not believe in their own capacity to effectively employ a particular strategy (i.e., they may have low emotion regulation self-efficacy; Gross, 2015). Indeed, as we know from cognitive behaviour therapy, believing that they are incapable of employing a particular strategy may then inhibit a person from attempting to initiate that strategy (Beck, 2021). Finally, even if vulnerable people decide to initiate a strategy, they may experience difficulties during its implementation; for example, people with depression often find it difficult to repair sad moods, in part, due to an impaired ability to recall happy memories—an impairment that persists even after recovery from depression (Joormann et al., 2007).

Taken together, although advocates claim that it is particularly important for vulnerable people to emotionally prepare to view sensitive content, it is possible that they may not want to emotionally prepare to view sensitive content or find it particularly difficult to do so. Sensitive-content screens in their current format also seemingly fail to address the difficulties vulnerable people may experience in implementing such preparation strategies by not including explicit instructions on *how* to emotionally prepare.

The Evidence

Trigger Warnings

Again, I draw on trigger warning literature to understand how sensitive-content screens might operate, here with respect to emotional preparation. Recall, emotional preparation involves processes in both the anticipatory (i.e., before the outcome is known) and content-viewing (i.e., after the outcome is known) periods. Thus, I first examine how trigger warnings impact people's emotional experiences in the anticipatory period, before turning to their impact on people's emotional reactions while viewing the forewarned content.

Anticipatory Period.

Consistent with the bracing literature, trigger warnings increase negative expectancies and create a noxious anticipatory period. For example, across five experiments, Bridgland et al. (2019) found participants who received a trigger warning (vs. no warning) expected the forewarned content to be significantly more negative and had higher state anxiety and negative affect during the anticipatory period. Indeed, the recent meta-analysis showed that trigger warnings have a small to medium-to-large effect on anticipatory affect—across both subjective (e.g., rating scales) and objective (e.g., psychophysiological measures) markers of distress (based on 32 unique effect sizes; Bridgland et al., 2023). Thus, trigger warnings appear to make people feel worse during the anticipatory period—and arguably elicit distress akin to that elicited by viewing the forewarned content itself.

Emotional Reactions to Forewarned Content.

In theory, a noxious anticipatory period could mitigate the emotional impact of viewing forewarned content—and thus, be conceptualised as a form of emotional preparation. However, existing research suggests otherwise. For example, trigger warnings (vs. no warning) have trivial effects on negative affect following exposure to negative text passages, film clips (Sanson et al., 2019), and lecture content—including among people with personal

experiences that match the lecture topics (e.g., sexual assault; Boysen et al., 2021). Emerging research has also found trigger warnings (i.e., anticipating neutral, positive, negative emotional reactions) have no impact on distress in the longer-term (at Day 1, 2 and 14)—even for people with higher PTSD scores at baseline (Kimble et al., 2022). Indeed, the recent meta-analysis found that a meaningful effect in either direction (i.e., towards a benefit or cost of trigger warnings) is unlikely (based on 86 unique effect sizes; Bridgland et al., 2023). Therefore, despite eliciting negative expectancies and creating a noxious anticipatory period, trigger warnings do not appear to mitigate the emotional impact of viewing forewarned content—and thus, are unlikely to elicit the kind of emotional preparation advocates claim.

So, *why* do trigger warnings fail to emotionally prepare people to view sensitive content? Emerging research on this issue has revealed two key possibilities. First, trigger warnings may not help people bring coping strategies to mind. In one study, participants reported what they would do when encountering a trigger warning related to their most stressful/traumatic experience (Bridgland, Barnard et al., 2022). The strategies these participants reported (e.g., leave [situation selection] or reappraise [cognitive change] the situation; Gross, 2015) were comparable to those reported by participants who imagined encountering trauma-related content without a trigger warning. Therefore, trigger warnings do not appear to make it more likely that people will bring coping strategies to mind. Second, trigger warnings may not help people pause so they can emotionally prepare for the forewarned content. In another study, participants viewed a traumatic film and then viewed images from the film, preceded by either a trigger warning or a neutral task instruction (Bridgland & Takarangi, 2022). There was no difference in the average time participants spent on the trigger warnings compared to the control screens (Bridgland & Takarangi, 2022). In fact, within the first two trials (which were always a warning and control screen) participants spent more time waiting on the control screen rather than the warning screen. Therefore, people do not appear to pause following trigger warnings. Taken together, trigger

warnings neither equip people with strategies for emotional preparation, nor assist people in taking a moment to pause before proceeding to the forewarned content. Therefore, unless people already have coping strategies—which we know may be unlikely for some people, especially vulnerable people—it seems that traditional trigger warnings may be ineffective at emotional preparation.

Sensitive-Content Screens

The evidence on whether sensitive-content screens, specifically, emotionally prepare people to view sensitive content is limited, but is more comprehensive than that for deterrence. Here, I draw on a recent multi-experiment study (of which I am a co-author; Takarangi et al., 2023). Again, I first examine how sensitive-content screens impact people's emotional experiences in the anticipatory period, before turning to their impact on people's emotional reactions while viewing the forewarned content.

Anticipatory Period.

To examine how sensitive-content screens impact people's emotional experiences in the anticipatory period, we presented participants with either a series of sensitive-content screens or control screens (i.e., grey masks), amongst neutral and positive images (Takarangi et al., 2023; Experiment 1). We then examined whether exposure to sensitive-content screens (without an option to view the forewarned content) increased participants' state anxiety and negative affect from pre- to post-task. Consistent with the bracing literature and research on traditional trigger warnings, participants in the sensitive-content screen condition reported a larger increase in state anxiety and negative affect compared to participants in the control condition. Thus, as we predicted, sensitive-content screens, like traditional trigger warnings, appear to create a noxious anticipatory period.

Emotional Reactions to Forewarned Content.

Again, it is possible that the noxious anticipatory period suggests people are emotionally preparing for the forewarned content; but much like the trigger warning research,

our results indicate otherwise. In a follow up experiment, we presented participants with negative images that either had preceding sensitive-content screens or not, amongst neutral and positive images (Takarangi et al., 2023; Experiment 3). We examined whether seeing sensitive-content screens (vs. not) prior to viewing negative images changed participants' state anxiety and negative affect. Indeed, if sensitive-content screens help people emotionally prepare—as advocates claim—then we would expect to see a reduction in negative emotional reactions for participants who see negative images preceded with sensitive-content screens (vs. not). However, we found no evidence of such emotional preparation; exposure to negative images increased people's state anxiety and negative affect *regardless* of whether those images were preceded by a sensitive-content screen.

Taken together, this preliminary work supports the idea that sensitive-content screens may not emotionally prepare people to view sensitive content. In fact, like trigger warnings, sensitive-content screens create a noxious anticipatory period that does not translate to an emotional benefit when people view the forewarned content.

1.5 Existing Research Gaps and Implications

Overall, research on traditional trigger warnings is limited, but the body of research examining sensitive-content screens *specifically* is especially lacking. Therefore, many questions remain from the first (albeit small) wave of research. Notably, there are more questions relating to deterrence than emotional preparation because there is more comprehensive existing evidence for the latter. Thus, in my thesis I focused on the first claim related to deterrence (Chapters 3 and 4), as well as adaptions related to deterrence (Chapters 4 and 5), because I was already well positioned to also focus on adaptions related to emotional preparation (Chapter 6).

Do Sensitive-Content Screens Deter People from Viewing Sensitive Content?

In relation to deterrence, we know that people tend to uncover one sensitive-content screen, but what happens when people see more than one screen? Do they repeatedly uncover

subsequent sensitive-content screens, or stop after uncovering the first screen? Additionally, we know that vulnerable people (e.g., people with higher depression symptoms) often engage with sensitive content, but are they *more* susceptible to uncovering sensitive-content screens, relative to people with less severe psychopathological symptoms?

Why do People Respond to Sensitive-Content Screens the way they do?

We know that most people uncover (or say they will uncover) a single sensitive-content screen because they are curious; other people say they will decide based on the context of the image (e.g., posting account, and content descriptions) and/or their perceived ability to cope with distressing content (Bridgland, Bellet et al., 2023). But do people's reasons for uncovering stay the same or change when they see multiple screens? It is possible that people initially uncover sensitive-content screens because they are curious but after seeing the sensitive (and potentially distressing) content underneath one screen, they decide not to uncover subsequent screens. Indeed, we know that people learn from past experiences and tend to avoid future aversive experiences (Krypotos et al., 2015). However, it is also possible that people continue uncovering because they learn they *are* able to cope with their image-related distress. Additionally, *each* sensitive-content screen may create a unique information gap (and/or elicit the "Pandora effect", morbid curiosity, or the "forbidden fruit effect") such that people are drawn to the content, and therefore repeatedly uncover subsequent screens—irrespective of their perceived ability to cope with distressing content.

How can Social Media Platforms Adapt Sensitive-Content Screens to Improve the Screens Utility as a Harm Minimisation Tool?

Reducing Uninformed Engagement

If each screen creates a unique information gap, such that people *are* drawn to repeatedly uncover sensitive-content screens, what would happen if people received more information about the content beneath the screens? Would people still uncover screens or would providing such content-related information reduce their information gap and

subsequently change their behaviour? And how would viewing such content-related information change people's emotional experiences while they view sensitive-content screens, or, if they decide to uncover screens, while they view the forewarned content? Perhaps providing content-related information would have differential costs and benefits on people's behaviour and emotional experiences.

Mitigating the Impact of Exposure to Sensitive Content

If sensitive-content screens provide no emotional benefit when people view the forewarned content—because screens do not equip people with strategies for emotional preparation—would providing explicit instructions detailing *how* to emotionally prepare (i.e., by encouraging people to engage in hedonically driven emotion regulation to down-regulate negative emotions and up-regulate positive emotions; Larsen, 2000), provide an emotional benefit? It is possible that this approach would address some of the challenges people—especially vulnerable people—experience when they need to implement strategies for emotional preparation. Indeed, we know that participants given instructions to use cognitive emotion regulation strategies (e.g., distraction, which involves directing attention away from negative situations, and reappraisal, which involves reinterpreting the meaning of negative situations) while viewing negative images (e.g., Ray et al., 2010; Thiruchselvam et al., 2011) and films (e.g., Wolgast et al., 2011) report lower negative emotions compared to when they are not given instructions/are asked to respond naturally. But does this emotional benefit also apply to negative images in a social media context?

Answering these questions is an important next step in this line of research and has theoretical, methodological, practical, and clinical implications. Indeed, although I can draw on existing research and related literature to make predictions about how people might respond when they are faced with more than one sensitive-content screen, and why they might respond the way they do, examining these questions *specifically* will provide the foundations for a theoretical framework on warning systems. This research will also help us

to develop (and refine) appropriate methodology for investigating the effects that sensitive-content screens have on people's behaviour and emotional experiences. With an improved theoretical understanding and appropriate methodology, we will then be better positioned to investigate potential evidence-based ways in which social media platforms can adapt sensitive-content screens (as well as traditional trigger warning)—with the ultimate aim to improve the screens utility as a harm minimisation tool. Together, this research could have practical implications for the way Instagram use and design sensitive-content screens, and more broadly for other social media platforms that use similar warning systems. Finally, improved harm minimisation tools would contribute to a safer online environment for users on these social media platforms, which may then have accumulating clinical implications in terms of improving users' overall mental health and wellbeing (e.g., Funder & Ozer, 2019). Although such improvements would come too late for Molly Russell's family, there remains hope that future deaths caused by the negative effects of online content may be prevented.

1.6 Summary

We know that exposure to sensitive content online may occur relatively frequently for some people and can have negative consequences (as in the case of Molly Russell). Instagram's current solution, to blur images and provide a warning, may come at an emotional cost, despite advocates' claims that such warning systems are beneficial for users. Indeed, existing theory and related literature (e.g., on trigger warnings) suggests that sensitive-content screens may not only *increase* engagement with—rather than deter people from—sensitive content, but they may also fail to emotionally prepare people to view such forewarned content. Preliminary research examining sensitive-content screens *specifically* appears to align with these claims, but—despite their widespread use across social media platforms—the body of research is lacking, especially with respect to how people respond behaviourally. Therefore, my thesis aims to address some of the many remaining research gaps from the first wave of research before investigating potential ways in which social

media platforms can adapt sensitive-content screens to improve the screens utility as a harm minimisation tool.

2 Overview of Thesis Studies

My thesis, broadly speaking, aims to investigate the empirical basis of sensitive-content screens. First, my thesis aims to fill existing research gaps by examining *how* people respond to sensitive-content screens when they are faced with more than one screen, both in terms of people's behaviour as well as their emotional experiences, and *why* they respond the way they do. I will focus more on the first claim related to deterrence because there is more comprehensive existing evidence for the second claim related to emotional preparation. Then, using this (and prior) empirical work as a foundation, my thesis aims to investigate two potential ways in which social media platforms can adapt sensitive-content screens to improve the screens utility as a harm minimisation tool. Specifically, my thesis examines whether adding 1) brief content-related information can reduce uninformed engagement with sensitive content, and 2) emotion regulation instructions can mitigate the impact of exposure to such content.

Chapter 3: Studies 1a and 1b

In Chapter 3 (Studies 1a and 1b), I sought to examine the claim that trigger warnings, specifically sensitive-content screens, increase deterrence. Across two studies, I built upon the preliminary work on sensitive-content screens (Bridgland, Bellet et al., 2022) to examine *how* people—including people vulnerable people—behave when they see more than one sensitive-content screen, and *why* they behave the way they do. I specifically examined behaviour over a series of sensitive-content screens because we know that people—especially people who seek out sensitive content—are likely to see more than *one* sensitive-content screen in real life, thereby improving the ecological validity of our design.

In Study 1a, I presented participants with a *series* of sensitive-content screens that appeared among neutral and positive images; participants could choose to uncover screens and view the negative image (or not) at their own discretion. Participant also completed a battery of vulnerability measures (i.e., for depression, anxiety, stress, PTSD symptom severity, self-triggering frequency, and wellbeing) and responded to 32 statements providing possible reasons why they uncovered sensitive-content screens or not *during* the task (e.g., "I uncovered the screened image(s) because my freedom to view the image was restricted"). I found most participants opted to uncover the first sensitive-content screen they came across, and over half continued to uncover every screen they saw during the task. I also found no evidence suggesting vulnerable people (e.g., people with higher rates of depression or PTSD) were more likely to avoid sensitive content: people's uncovering behaviour was similar irrespective of their vulnerabilities. Additionally, I found five key reasons for uncovering behaviour—information seeking behaviour, thrill-seeking behaviour, positive and negative affect driven behaviour, and avoidance behaviour.

In Study 1b, I sought to replicate the findings from Study 1a using a slightly different methodology; I fixed the number of images, so that participants made the same number of behavioural choices, and I introduced a 3s image-response delay to limit participants' ability to rush through images. Participants completed the same battery of vulnerability measures and an updated version of the reasons for uncovering questionnaire. Overall, results from Study 1b were consistent with Study 1a. However, uncovering behaviour over the entire task was slightly lower in Study 1b; many participants still uncovered multiple screens, but a smaller proportion of participants uncovered every screen they saw during the task. Additionally, the thrill-seeking behaviour factor collapsed into the information seeking behaviour factor in Study 1b; thus, I found four (not five) key reasons for uncovering behaviour. Taken together, these findings suggest that sensitive-content screens may be

ineffective at deterring people, including vulnerable people, from engaging with sensitive content, and there appears to be several reasons underpinning people's uncovering behaviour.

Chapter 4: Study 2

In Chapter 4 (Study 2), I experimentally examined the most strongly endorsed reason for uncovering sensitive-content screens that I identified in Chapter 3—information seeking behaviour. Specifically, I wondered if sensitive-content screens in their current format prompt information seeking behaviour *because* of their highly ambiguous/uncertain nature (e.g., Loewenstein, 1994). I also wondered if intolerance to uncertainty—which refers to having negative beliefs about uncertainty and its implications (e.g., "uncertainty keeps me from living a full life"; Carleton, Mulvogue et al., 2012)—plays a role in such behaviour.

To investigate these ideas, I built upon Studies 1a and 1b by manipulating the amount of content-related information, in the form of content descriptions, presented *on* sensitive-content screens during a simulated Instagram image-viewing task. There were three conditions varied within-subjects: no content descriptions, brief content descriptions, and detailed content descriptions. Participants viewed images/sensitive-content screens one at a time and could choose to uncover screens and view the negative image (or not) at their own discretion. Participants also completed a measure of intolerance to uncertainty and explained (using open text) whether (and how) content descriptions influenced their decision to uncover sensitive-content screens or not. I found participants uncovered sensitive-content screens irrespective of content description type but did so *most* often when the screens had no content description, and *least* often when the screens had a brief or detailed content description. I found no evidence to suggest that intolerance to uncertainty moderated the relationship between the level of information provided and uncovering behaviour. Moreover, most participants indicated that content descriptions influenced their decision to uncover sensitive-content screens, and specifically, knowing *what* the sensitive content depicted bolstered their ability to make an informed uncovering decision.

Taken together, these findings suggest that sensitive-content screens in their current format may promote engagement with, rather than deterrence from, sensitive content. My findings also raise the possibility of adapting sensitive-content screens to include content-related information, with the intention of reducing uninformed engagement with sensitive content.

Chapter 5: Studies 3a and 3b

In Chapter 5 (Studies 3a and 3b), I experimentally examined if the reduction in exposure to sensitive content that I found in Chapter 4 (i.e., in the brief and detailed content description conditions) comes at an emotional cost for people. Specifically, I wondered if including content-related information on sensitive-content screens increases people's anxiety during the anticipatory period (i.e., before the outcome is known e.g., Blackwell, 2019; 2021), and/or their distress if they decide to uncover the screens and view the forewarned content (i.e., after the outcome is known). Across two experiments, I investigated these possibilities.

In Study 3a, I compared participants' change in state anxiety pre and post a passive image-viewing task when exposed to sensitive-content screens *with* brief or detailed content descriptions or *without* content descriptions. I used the same content descriptions from Study 2, but this time I varied the conditions between-subjects. State anxiety was similar for participants who saw sensitive-content screens with and without *brief* content descriptions, but participants who saw sensitive-content screens with *detailed* content descriptions showed larger increases in state anxiety (relative to brief content descriptions).

In Study 3b, I presented participants with a single sensitive-content screen, either with or without a brief content description, and gave them the opportunity to uncover it. Participants who uncovered the screen rated their distress after viewing the negative image. I found participants uncovered the screen, and experienced similar levels of distress, irrespective of whether they saw screens with *or* without brief content descriptions. In other

words, I found no evidence to suggest that brief content descriptions create an emotional cost when people view sensitive-content screens, or, if they decide to uncover them, when they view the forewarned content. Therefore, including brief content-related information on sensitive-content screens, explicitly indicating what the content depicts, may be one way Instagram can adapt screens to improve their utility as a harm minimisation tool.

Chapter 6: Studies 4a and 4b

In Chapter 6 (Studies 4a and 4b), I sought to examine whether explicitly instructing people to use emotion regulation instructions could assist with emotional preparation—given that we know they are ineffective at emotional preparation in their current format, and that *some* people will uncover sensitive-content screens, regardless of whether they also include content-related information. Specifically, I experimentally examined if encouraging people to use distraction (by directing attention away from negative images) or reappraisal (by reinterpreting the meaning of negative images) could reduce their distress. Across two experiments, I investigated this possibility.

In Study 4a, I first trained participants to use distraction and reappraisal, then showed them a series of sensitive-content screens (accompanied by reappraisal, distraction, or no instructions) followed by a negative image. After viewing each image, participants rated how distressed they felt in that moment. I found participants reported *lower* distress after images where I instructed them to use reappraisal or distraction, compared to images without instructions to regulate. Although participants reported the lowest distress following reappraisal instructions, they preferred using distraction. Notably, the effects of reappraisal and distraction were small, but getting participants to switch between two regulation strategies and/or including no regulation trials (i.e., varying regulation instructions within-subjects) may have dampened the effects.

Therefore, in Study 4b, I addressed the limitations of using a within-subjects design and sought to replicate the effect of distraction vs. no instruction using a between-subjects

design. I focused on distraction because participants preferred distraction over reappraisal, and distraction is easier to teach. I used the same method as in Study 4a, but randomly allocated participants to a distraction or no instruction condition. Participants who received distraction instructions reported substantially lower distress than participants who received no instructions.

Taken together, these findings suggest sensitive-content screens in their current format (without instructions) fail to help people emotionally prepare and suggest that providing explicit instructions detailing *how* to emotionally prepare—using cognitive emotion regulation strategies—can reduce the negative impact of exposure to sensitive content. Therefore, adding cognitive emotion regulation instructions to sensitive-content screens may be another way Instagram can adapt screens to improve their utility as a harm minimisation tool.

3 Investigating Whether Instagram's Sensitive-Content Screens Deter People from Viewing Negative Content

Chapter 3 is published as:

Simister, E. T., Bridgland, V. M. E., & Takarangi, M. K. T. (2023). To look or not to look: Instagram's sensitive screens do not deter vulnerable people from viewing negative content. *Behaviour Therapy.* https://doi.org/10.1016/j.beth.2023.06.001

Authors Contributions: I developed the study design with the guidance of MKTT and VMEB. I collected the data, performed the data analysis and interpretation, and drafted the manuscript. Data from Study 1a was collected as part of my honours study, however the data were re-analysed (using additional analyses, such as factor analysis) for the current thesis. MKTT and VMEB contributed equally by making critical revisions to the manuscript. All authors approved the final version of the manuscript for submission.

Abstract

By blurring sensitive images and providing a warning, Instagram's sensitive-content screens seek to assist users—particularly vulnerable users—in making informed decisions about what content to approach or avoid. Yet, prior research found most people (~85%) chose to uncover a single screened negative image (Bridgland, Bellet et al., 2022). Here, we extended on and addressed shortcomings of this previous research. Across two studies, we presented participants with a series of sensitive-content screens covering negative content that appeared among neutral and positive images; participants could choose to uncover screens (or not). We found most participants opted to uncover the first screen they came across, and many continued to uncover screens over a series of images. We also found no evidence suggesting vulnerable people (e.g., people with higher rates of depression) are more likely to avoid sensitive content: people similarly uncovered sensitive-content screens irrespective of their vulnerabilities. Thus, sensitive-content screens may be ineffective in

deterring people from exposing themselves to sensitive content. Additionally, avoidance behaviour, information seeking behaviour, negative affect driven behaviour, and positive affect driven behaviour appeared to underpin participants' decisions to uncover screened content.

Introduction

"Social media helped kill my daughter" claimed the father of 14-year-old Molly Russell, who took her own life in 2017 after entering Instagram's "dark rabbit hole of depressive suicidal content" (Crawford, 2019). In response to scrutiny following Molly's suicide, Instagram prohibited all graphic self-harm images and added sensitive-content screens to non-graphic self-harm related content (Mosseri, 2019a). However, Instagram *first* began blurring sensitive images, and applying a warning, in 2017. These sensitive-content screens intend to reduce the "surprising or unwanted experience" of discovering sensitive content, and deter people—particularly people with mental health vulnerabilities (e.g., operationalised here as depression, post-traumatic stress, history of self-triggering; referred hereafter as "vulnerable people" to align with Instagram's terminology; Mosseri, 2019a)—from viewing the content altogether. But recent research indicates sensitive-content screens may *not* act as a deterrent. In fact, most people—particularly vulnerable people—said they would uncover sensitive-content screens (~80%), and did so during a simulated Instagram task (~85%; Bridgland, Bellet et al., 2022). However, we do not know whether people *repeatedly* uncover screens, if they stop after initially uncovering, or *why* they choose to uncover. We addressed these important research gaps here.

Sensitive-content screens are a specific type of trigger warning—trigger warnings are *any* alert/message that informs people upcoming content may be distressing. Thus, we can draw on trigger warning literature to understand how screens might operate. Advocates claim trigger warnings help people *prepare* themselves to process negative content (e.g., Lockhart, 2016). However, there is no evidence to support this claim; at best, people's emotional

responses (e.g., distress) are similar irrespective of whether content is accompanied by a warning or not (Boysen et al., 2021; Bridgland et al., 2019; Bridgland, Barnard et al., 2022; Sanson et al., 2019). At worst, trigger warnings may *increase* people's negative reactions (e.g., anxiety) to sensitive content (Bellet, Jones, Meyersburg et al., 2020; Jones et al., 2020). Also, advocates claim trigger warnings help people avoid potentially negative content (e.g., Manne, 2015). Yet studies have found no difference in participants' preference to view titles (e.g., for videos) with or without a warning (Gainsburg & Earl, 2018). Similarly, only a small number of participants (< 6%)—including people who met criteria for probable Posttraumatic Stress Disorder (PTSD)—avoided reading potentially triggering text by selecting a non-triggering alternative (Kimble et al., 2021). Overall, trigger warnings seem ineffective at achieving their purported goals, suggesting that sensitive-content screens might also fail to prepare people emotionally and/or help them avoid sensitive content.

Related research suggests sensitive-content screens may *encourage* people to look at sensitive content. When freedom to experience activities is restricted (e.g., via warning labels), people often find them *more* attractive (i.e., "forbidden fruit effect"; Weaver, 2011). Moreover, screens may foster uncertainty—an aversive state people are motivated to reduce (Berlyne, 1954)—which may drive people to uncover screened content. Worryingly, people may be more likely to engage with stimuli if consequences are uncertain *and* negative (the "Pandora effect"; Hsee & Ruan, 2016). Uncertainty may also drive curiosity, which—according to the information-gap hypothesis (Loewenstein, 1994)—arises when what people *want to know* exceeds their current knowledge. In fact, some people are "morbidly" curious, and deliberately expose themselves to negative content (e.g., images that portray death; Oosterwijk, 2017). Therefore, sensitive-content screens may make avoidance unlikely.

Extant literature also suggests *vulnerable* people may be particularly susceptible to uncovering sensitive-content screens. People with major depressive disorder demonstrate higher desire for sadness (vs. non-depressed controls; Arens & Stangier, 2020) and often

down-regulate positive emotions (e.g., by listening to sad music, Millgram et al., 2015). The familiarity of negative emotions—albeit unpleasant—may serve *instrumental* motives (e.g., sustaining a sense of self; Arens & Stangier, 2020). Similar counter-hedonic motives are evident among a subset of people with PTSD who "self-trigger" by intentionally exposing themselves to reminders of their trauma (Bellet, Jones & McNally, 2020). Self-triggering may help people make meaning of traumatic experiences or maintain consistent levels of PTSD symptoms to avoid being surprised by unexpected increases in symptomology (Bellet, Jones & McNally, 2020). Indeed, a preference for avoiding contrasting emotional states is established in other clinical disorders (e.g., generalised anxiety disorder); more specifically, the Contrast Avoidance Model postulates that some people deliberately engage in negative thinking (e.g., worry or rumination) to perpetuate a negative mood and thereby avoid the shift from positive or neutral moods into negative moods (Crouch et al., 2017). Taken together then, vulnerable people may be *more* susceptible to uncovering sensitive-content screens.

In the first investigation of uncovering behaviour, participants—predominantly young (~36 years) European American/White female Instagram users, who were recruited from the United States using Mechanical Turk (MTurk)—saw a single sensitive-content screen and reported how likely they would be to uncover it (Bridgland, Bellet et al., 2022; Study 1a). Eighty percent of participants said they would uncover the screen and this intention was associated with factors including poorer wellbeing, higher depression symptoms and experiential avoidance (i.e., tendency to avoid thoughts), and lower perceived life meaningfulness; these factors, in turn, are linked to a range of psychopathologies (Beck, 2009; Kashdan et al., 2006; Keyes et al., 2010). In Study 1b, participants viewed a single sensitive-content screen presented after neutral and positive images and could "uncover" the screen (or not, though the negative image was not shown). Replicating Study 1a, most people (~85%) chose to uncover the screen. However, unlike Study 1a, the "uncover" decision was not associated with vulnerability characteristics. Participants also reported (using an open-

text box) what factors would affect whether they would uncover the screen (Study 1a) and why they did or did not uncover the screened image (Study 1b). In Study 1a, 35.8% of participants reported they simply wanted to see the image, and of these, 75.3% (26.9% of the total sample) specifically mentioned they would uncover the image because of reasons related to curiosity. Other participants (36.2%) indicated they would decide whether to uncover based on the context of the image (e.g., posting account, and caption). In Study 1b, curiosity remained the primary reason for uncovering screened images (46.2%). Other participants (10.7%) also mentioned they would uncover/keep covered based on their general tendency to cope with or not cope with distressing content.

This preliminary work supports the idea that sensitive-content screens may not deter people from viewing sensitive content—at least for *one* image. This work also raises several reasons that might help us understand why. However, real-life exposure to sensitive content on Instagram would likely exceed one image, especially for users who follow accounts frequently sharing sensitive content. Currently, it is unclear whether seeing sensitive content deters further uncovering or has no effect on subsequent behaviour. Due to discrepant findings between Bridgland, Bellet et al.'s (2022) two studies, it is also unclear whether vulnerable people are more susceptible to uncovering sensitive-content screens (than people lower in vulnerability characteristics). It is also unclear whether the reasons people uncover screened images changes when exposure to screens increases.

Here, we addressed these limitations: our primary aim was to investigate how people—including vulnerable people—interact with sensitive-content screens over a series of images. In Study 1a, participants viewed *multiple* sensitive-content screens for a fixed time (5-min)—amongst neutral and positive images. We measured frequency of uncovering, and the relationship between uncovering behaviour and vulnerability measures (e.g., depression). In Study 1b, we fixed the number of images, rather than task time: participants saw the same 90 images (30 per valence). To improve control over participants' behaviour, we introduced a

3s delay between screen presentation screen and when response options appeared. Our secondary, and more exploratory, aim across both studies was to examine *why* people uncover sensitive-content screens. In Study 1a, participants responded to items that covered a range of reasons why they did (or did not) uncover screens. We developed these items in part on Bridgland, Bellet et al.'s (2022) previous work and in part on the existing literature that offers potential explanation for why screens may encourage people to look at sensitive content (e.g., the "forbidden fruit effect"; Weaver, 2011), and why vulnerable people may be particularly susceptible (e.g., the Contrast Avoidance Model; Crouch et al., 2017). In Study 1b, participants responded to an updated version of this questionnaire.

Based on previous literature and theory, in Study 1a we predicted participants would opt to approach, rather than avoid, sensitive content, and that uncovering behaviour (variously operationalised, e.g., uncovered all images or not) would positively correlate with vulnerability characteristics (e.g., depression, PTSD symptom severity, and self-triggering frequency), and negatively correlate with wellbeing. We made no predictions regarding the reasons why people uncover sensitive-content screens since this was an exploratory aim.

Study 1a

Method

The Flinders University Social and Behavioural Research Ethics Committee approved this research, and we preregistered it on OSF (https://osf.io/2fdr7). We used Qualtrics Software (2018). We report all measures, conditions, and data exclusions.[1] The supplementary materials are at the end of the chapter and the data, including a codebook describing all variables, can be found at: https://osf.io/rj987/.

[1] After pre-registration we realised it was not feasible (in terms of time for manual data processing) to calculate two subsidiary ways we proposed to operationalise uncovering behaviour (i.e., time spent viewing the first uncovered image, and the proportion of images participants uncovered *after* the first image they uncovered). Moreover, these variables do not assist in answering our main research questions; therefore, we have not included these variables in our analyses.

Participants

Because the magnitude of a correlation stabilises as it approaches $N = 260$ (with a power of .80, for effect sizes $r = 0.1$ and larger [the smallest effect we would be interested because anything smaller is less likely to be as consequential]), our desired sample was 260 participants (Schönbrodt & Perugini, 2018). To promote data quality and avoid bots/server farmers, we screened out participants who failed a captcha, scored fewer than 8/10 on an English proficiency test (Moeck et al., 2022), and/or selected "Konnect" (a bogus platform included to detect inattentive responses) when asked about social media use. Because we only wanted to recruit Instagram users, we also screened out participants who indicated they do not use Instagram. We recruited 300 participants from the United States using MTurk. Of these, we excluded one participant who failed all three embedded attention checks, one who experienced technical issues, one who did not follow instructions (i.e., rushed through images), and 34 who reported uncovering screens to "fulfil task requirements".[2] Participants received $2.50 USD.

Our final sample of 263 participants, aged 18–71 years ($M = 35.4$, $SD = 8.5$) included 51.7% females, 47.5% males; one participant preferred not to report gender and three identified as non-binary. Our sample was predominantly European American/White (72.6%, $n = 191$); other participants were of African American/Black (8.7%), Asian (7.2%), and Hispanic/Latino (2.7%), or other (6.1%; e.g., multiracial/biracial) descent; 2.7% of participants specified nationality (e.g., American) instead of ethnicity. Most participants (54.4%) reported $45,000–$140,000 in income and tended to be (66.9%) college graduates.

[2] After collecting data from 237 participants, we detected several reasons for uncovering related to "fulfilling task requirements". Because we are interested in understanding behaviour as it typically occurs on Instagram—rather than when people believe such behaviour is required—we excluded these participants. We amended task instructions for the remaining data collection (see OSF addendum: https://osf.io/f8tgx).

Materials[3]

Instagram Task.

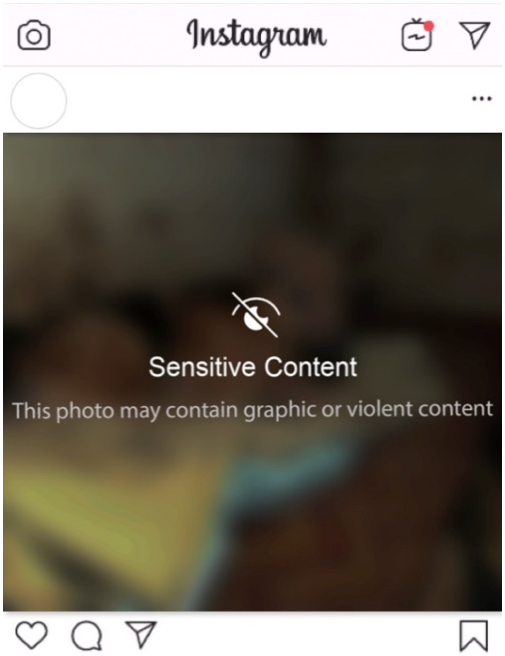
Participants viewed the 70 most negative, positive, and neutral images (210 total) from the Nencki Affective Picture System (NAPS; Marchewka et al., 2014; based on normative ratings: 1 = *negative* to 9 = *positive;* see Appendix A for image codes). We included positive and neutral images to improve the ecological validity of our design; participants are likely to come across negative images *among* positive and neutral images on their own Instagram feed. The content of the images (e.g., people, animals, objects) is commonly found on Instagram and would likely meet the threshold for Instagram to screen it (e.g., many negative images include people/animals that have been injured/are deceased). All images appeared in an Instagram border with non-functional like and comment buttons (Figure 1.1). Consistent with Instagram's sensitive screen format, a warning covered negative images. We randomly assigned participants to one of two warnings: one with wording Instagram introduced in ~2020, "This photo may contain graphic or violent content": the other with Instagram's original 2017 wording, "This photo contains sensitive content which some people may find offensive or disturbing". We included both warnings to assess whether wording influenced uncovering behaviour (and to compare with Bridgland, Bellet et al., 2022 who used original wording), but wording was not a key manipulation in the present study and therefore we made no predictions about it. Other work also suggests that changes to warning wording does not influence outcomes (Bridgland et al., 2019). The task occurred for 5-min; participants viewed images/screens—presented in a randomised order—one at a time. The number of images/screens viewed depended on the time participants spent viewing each image/screen. We made the task time-limited rather than a fixed number of images to reduce

[3] Participants also completed state anxiety/mood measures before and after the task; because these variables are independent of the current aims, these data are reported elsewhere. Moreover, participants completed the Centrality of Events Scale (CES; Berntsen & Rubin, 2006), but it is not pertinent to our aims, so it is not included in analyses (see Supplementary Table S1 for correlations).

the "appeal" of rushing through images, and to control total task time. When participants saw a sensitive screen, they had the option to uncover it (select *See Photo*) and view the negative image underneath, or leave it covered (select *Next Photo*) and move to the next image.

Figure 1.1

Example NAPS Image Modified to Look Like an Instagram Image with Sensitive-Content Overlay



Depression Anxiety Stress Scales-21 (DASS-21; Lovibond & Lovibond, 1995).

Participants rated the degree to which 21 statements (e.g., "I felt downhearted and blue") applied to them over the past week (0 = *never* to 3 = *almost always*). We summed items for each subscale (higher scores indicate higher severity; present study depression: α = .89; anxiety: α = .87; stress: α = .89).

Trauma History Screen (THS; Carlson et al., 2011).

Participants completed the THS to identify/index their most traumatic/stressful event. They indicated if (and how often) they had been exposed to traumatic events (e.g., "A really bad car accident"), then described the event that bothers them most. If the event(s) did not

bother them, or they had not experienced any of the events, we asked them to describe their most stressful experience. We told participants they would refer to their identified event in subsequent questions. Participants also provided: their age at the time of event; whether anyone was hurt or killed (yes/no); whether they felt afraid, helpless, or horrified (yes/no); how long they were bothered by it (1 = *not at all* to 4 = *a month or more*); and how much it bothered them emotionally (1 = *not at all* to 5 = *very much*). The THS has good test-retest reliability (high magnitude stressors, i.e., sudden events that cause extreme distress in most people: r = .77-.93, persisting posttraumatic distress events, i.e., events associated with significant subjective distress that lasts more than a month: r = .73-.95; Carlson et al., 2011).

Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5; Weathers et al., 2013).

Participants rated the degree to which they had been bothered by 20 symptoms—corresponding to the DSM-5 symptom criteria for PTSD (e.g., "Feeling jumpy or easily startled")—over the past month (0 = *not at all* to 4 = *extremely*). The items were indexed to their most stressful/traumatic event (from the THS). We summed items for a total severity score (higher scores indicate higher severity; present study α = .96) and subscales scores: intrusions (α = .93); avoidance (α = .89); negative alterations in cognitions and mood (α = .89); alterations in arousal and reactivity (α = .85).

History of Self-Triggering.

Participants completed two items to assess their history of self-triggering. Participants indicated if they (1) had self-triggered with reminders of their most stressful/traumatic event (yes/no), and if "yes", we asked them (2) how frequently they had self-triggered in the past month (1 = *not at all* to 6 = *every day*; adapted from the Self-Triggering Questionnaire; Bellet, Jones & McNally, 2020).

The Scales of General Well-Being Short Form (SGWB-14; Longo et al., 2018).

Participants rated 14 statements relating to life experiences (e.g., "I have a purpose"; 1= *not at all true* to 5 = *very true*). We summed items; higher scores indicate higher wellbeing (present study: α = .95).

Reasons For Uncovering Questionnaire.

Participants responded to 32 statements regarding possible reasons why they uncovered sensitive-content screens or not *during* the task (e.g., "…my freedom to view the image was restricted"; 0 = *not at all true of me* to 4 = *extremely true of me*; see Table 1.3 for α).

Procedure

We told participants we were investigating Instagram engagement, personality, and life experiences. After consent, all participants completed demographic and Instagram use questions, and items designed to reduce suspicion and demand: participants rated how often they view topics on Instagram (e.g., Fashion). Participants then completed the Instagram task and vulnerability measures (i.e., DASS, THS, PCL-5, STQ, and SGWB-14), counterbalanced to minimise potential order effects. Completing vulnerability measures first may have changed participants' affect prior to the Instagram task, potentially changing uncovering behaviours, whereas completing the Instagram task first (and seeing negative images) could have primed participants to response differently (e.g., more negatively) on the vulnerability measures. Participants then responded to the reasons for uncovering questionnaire, before completing feedback questions (e.g., Have you seen sensitive-content screens on your own Instagram?). To detect poor response quality, we asked participants if they had any technical issues or left the task for an extensive period: 21 (8.0%) participants reported leaving the task (five for < 1 min; 11 for a few minutes, two for > a few minutes; three participants did not report how long). Finally, participants were debriefed.

Statistical Analyses

Consistent with our pre-registration, we used descriptive statistics to examine uncovering behaviour, and correlations to examine relationships between this behaviour and vulnerability measures. Given the high rates of uncovering all images (which we determined during pilot testing), we also classified participants as uncovering *all* images or not. To better understand the *key* reasons why people uncovered sensitive-content screens, we ran principal components analyses (PCA) on the reasons for uncovering questionnaire. We assessed the suitability of PCA prior to analysis, and found the data were likely factorisable (see supplementary materials). We ran three PCAs because several items interfered with 'simple structure' in the first two (Thurstone, 1947); we reran the third PCA without these items. We also ran the third PCA without participants who did not uncover any screens (n = 14), but the factors did not change so we report the PCA including all participants.

Results and Discussion

Preliminary Analyses

Because participants' progress through the Instagram task was self-paced, we first considered their overall exposure to negative content. On average, participants saw 61.53 (*SD* = 32.96, *Mdn* = 55, range = 9-171) images during the task. As expected, around a third (M = 19.98, SD = 11.11, *Mdn* = 18, range = 2-56) were negative images covered by screens. We next considered the effect of warning wording. There were no statistically significant differences in uncovering behaviour between the two warning wordings (first image behaviour, $\chi^2(1) = 0.0002$, $p = .988$; uncovered all, or not, $\chi^2(1) = 0.09$, $p = .769$; proportion of screens uncovered, $r_s = .07$, $p = .229$).[4] Therefore, we collapsed all analyses across warnings. We also assessed possible task order effects. Participants who completed the PCL-5 before the Instagram task (M = 19.9, SD = 16.5) had higher PTSD symptom severity than

[4] The proportion of screens uncovered was not normally distributed (skewness ± SE: -0.825 ± 0.150; kurtosis ± SE: -1.048 ± 0.299): visual inspection of the histogram revealed a negatively skewed distribution (even after transformation). Thus, we ran non-parametric tests (i.e., Spearman's correlations) for this variable.

those who completed it after the task (M = 15.7, SD =15.6), $t(261) = 2.12$, $p = .035$, $d = 0.26$. It is possible the Instagram task led participants to reduce symptom reporting (Nahleen et al., 2021), or there were existing group differences despite random allocation. However, the differences in PTSD symptom severity by time of measurement did not influence uncovering behaviour. Moreover, there were no significant order effects for depression symptom severity or wellbeing (ts < 1, ds = -0.6-0.09). In terms of uncovering behaviour, there were no overall differences between participants who completed vulnerability measures before vs. after the Instagram task (first image behaviour, $\chi^2(1) = 0.97$, $p = .326$; uncovered all, or not, $\chi^2(1) = 0.45$, $p = .504$; proportion of screens uncovered, $r_s = -.05$, $p = .466$). Therefore, we collapsed time of measures (before vs. after) for all analyses.

Participant Characteristics

Because sensitive-content screens are designed primarily to protect vulnerable users, we examined our sample for vulnerability characteristics (Table 1.1). Overall, 85.9% (n = 226) of participants reported experiencing one or more high magnitude stressors, and 59.3% (n = 156) reported experiencing actual or threatened death/injury (Carlson et al., 2011). The most common events were the sudden death of a close family member/friend (57.4%; n = 151). Further, 19.4% of the sample met criteria for probable PTSD according to the conservative PCL-5 cut-off (> 33; Bovin et al., 2016), and 20.5% (n = 54) of participants indicated they had self-triggered with reminders of their most stressful/traumatic experience; of these, 32.0% (n = 16) had self-triggered at least 2-3 times over the past month. Moreover, 17.5% (n = 46) of participants were experiencing severe to extremely severe distress associated with symptoms related to depression according to the DASS-21 cut-off (>11; Lovibond & Lovibond, 1995). We next examined Instagram use. Most participants reported using Instagram every day in the last 7 days (56.7%; n = 197) and on average, using it for more than one hour a day (in the last 30 days; 72.6%; n = 191). Additionally, 34.2% (n = 90) of participants reported they have seen sensitive-content screens on their Instagram feed.

Hypothesis Testing

Uncovering Behaviour

To address our primary aim, we examined the frequency of uncovering the first sensitive-content screen and then all screens viewed within the 5-min task. Consistent with Bridgland, Bellet et al. (2022), most participants (i.e., 84.4%; n = 222) uncovered the first screen they saw. Of these participants, 136 (51.7% of the total sample) repeatedly uncovered *all* subsequent screens, whereas 86 (32.7% of the total sample) left at least one subsequent screen covered. On average, participants uncovered 12.44 screens (SD = 9.05) during the task. Of the 41 participants who left the first screen covered, only 14 (5.3% of the total sample) left all subsequent screens covered.[5] Thus, people opt to approach, rather than avoid, sensitive content, and often do so repeatedly despite the screen.

Table 1.1

Means (and Standard Deviations) for Vulnerability Measures

Measure	Scale (Range)	*M* (*SD*)
DASS-21	Depression (0–21)	5.0 (5.2)
	Anxiety (0–21)	3.4 (4.2)
	Stress (0–21)	5.1 (4.5)
SGWB-14	(14–70)	48.1 (12.6)
PCL-5	Intrusions (0–20)	4.3 (4.8)
	Avoidance (0–8)	2.4 (2.4)
	Negative Cognition/Mood (0–28)	5.6 (5.9)
	Hyperarousal (0–24)	5.5 (4.8)
	Total (0–80)	17.9 (16.2)

Note. n = 263. DASS-21 = Depression Anxiety Stress Scales-21; SGWB-14 = Scales of General Well-Being; PCL-5 = Posttraumatic Stress Disorder Checklist.

[5] We included these participants within the "did not uncover all" category for all subsequent analyses.

Uncovering Behaviour and Vulnerability Characteristics

To see if vulnerable people were particularly susceptible to uncovering screens, we next examined whether vulnerability characteristics were associated with uncovering behaviour (Table 1.2).[6] Against predictions, we found no statistically significant relationships between the vulnerability measures (i.e., depression, anxiety, stress, PTSD symptom severity, self-triggering frequency, or wellbeing) and uncovering behaviour. Therefore, vulnerable people were no more susceptible to engaging with sensitive content (than people lower on these measures); however, these results also suggest vulnerable people were also *no more likely* to avoid such content. After collecting the data, we wondered whether vulnerable people engaged with less content overall (by viewing images at a slower pace)—a more passive type of avoidance. However, we found no evidence to support this possibility: there were no statistically significant relationships between the number of images viewed and vulnerability measures.

Reasons for Uncovering

To address our secondary aim, we examined our PCA. Five components had eigenvalues > 1, explaining 27.1%, 23.7%, 6.7%, 4.9%, and 3.6% of the total variance, respectively. Visual inspection of the scree plot indicated five components should be retained: a five-component solution was also interpretable/exhibited 'simple structure' (Thurstone, 1947; Supplementary Table S1.2). Therefore, we retained five factors (Table 1.3): one related to avoidance and four related to approach behaviours.[7] *Avoidance behaviour* included eight items[8]; although items were conceptually different (e.g., "I do not enjoy taking risks" vs. "I do not like viewing distressing material"), they were not distinct enough to load

[6] We reran all vulnerability analyses *within* the "did not uncover all" subset of participants, but the pattern of results did not change.
[7] To see if vulnerable people endorse certain reasons for uncovering, we ran exploratory correlations between vulnerability measures and the reason factors. We found some significant correlations, but there were several discrepancies between Studies 1 and 2. Because a full discussion is beyond the scope of this manuscript, we report these correlations and discuss possible explanations for the discrepancies in the supplementary materials.
[8] We removed one item ("I was trying to forget my past negative experiences") due to conceptual differences.

onto separate factors. Conversely, the remaining factors were characterised by *uncovering* sensitive content but were distinct.[9] *Information seeking behaviour* was characterised by desire to obtain information about the image/alleviate uncertainty and curiosity (e.g., "I wanted to know why it was covered"). *Thrill-seeking behaviour* was characterised by enjoyment in exhilarating experiences (e.g., "I was excited to see what might lie beneath the screen"). *Negative affect driven behaviour* [10] was characterised by making sense of negative experiences/regulating negative affective states (e.g., "I was trying to remind myself of past negative experiences"). Finally, *positive affect driven behaviour* was characterised by positive emotion states (e.g., "I was content"). Therefore, although participants often approached sensitive content, the reasons underpinning such behaviour seemed to differ.

[9] We thank the editor and an anonymous review for their suggestions in naming these factors.
[10] We note that affect driven behaviour may encompass efforts to increase, decrease or maintain affect, as well as savour (or ruminate on) affect (Gross et al., 2015). Thus, people may engage in affect driven behaviour for a range of reasons (Tamir et al., 2016).

Table 1.2

Correlations Between Uncovering Behaviour and Vulnerability Measures

Measure	Scale	Type of Uncovering Behaviour				
		First image r_{pb} [95% CI]	Proportion uncovered r_s [95% CI]	Uncovered all (or not) r_{pb} [95% CI]	Total images viewed r_s [95% CI]	Total screens viewed r_s [95% CI]
DASS-21	Depression	-.01 [-.13, .11]	-.04 [-.16, .08]	-.06 [-.18, .06]	.07 [-.05, .19]	.07 [-.05, .19]
	Anxiety	.002 [-.12, .12]	.03 [-.09, .15]	-.001 [-.12, .12]	.07 [-.05, .19]	.07 [-.05, .19]
	Stress	-.01 [-.13, .11]	.01 [-.11, .13]	-.04 [-.16, .08]	.03 [-.09, .15]	.02 [-.10, .14]
SGWB-14		.003 [-.12, .12]	.04 [-.08, .16]	.07 [-.05, .19]	-.10 [-.22, .02]	-.12 [-.24, .001]
PCL-5	Intrusions	.02 [-.10, .14]	.03 [-.09, .15]	.07 [-.05, .19]	.11 [-.01, .23]	.08 [-.04, .20]
	Avoidance	-.01 [-.13, .11]	.02 [-.10, .14]	.04 [-.08, .16]	.03 [-.09, .15]	.001 [-.12, .12]
	Neg Cog/Mood	-.03 [-.15, .09]	-.03 [-.15, .09]	.01 [-.11, .13]	.08 [-.04, .20]	.06 [-.06, .18]
	Hyperarousal	-.01 [-.13, .11]	-.03 [-.15, .09]	.01 [-.11, .13]	.07 [-.05, .19]	.07 [-.05, .19]
	Total	-.09 [-.21, .03]	-.01 [-.13, .11]	-.01 [-.13, .11]	.10 [-.02, .22]	.08 [-.04, .20]
STQ	Freq. of ST	.18 [-.10, .44]	.15 [-.13, .41]	.13 [-.15, .39]	-.13 [-.39, .15]	-.15 [.41, .13]

Note. n = 263, except frequency of ST (= self-triggering) n = 50. DASS-21 = Depression Anxiety Stress Scales-21; SGWB-14 = Scales of General Well-Being; PCL-5 = Posttraumatic Stress Disorder Checklist. STQ = Self-Triggering Questionnaire. Neg Cog/Mood = Negative Cognition/Mood

Table 1.3

Means (and Standard Deviations) for Reason Factors

Reasons	*M* (*SD*)	*α*
Avoidance behaviour	0.91 (1.02)	.91
Negative affect driven behaviour	0.27 (0.58)	.91
Positive affect driven behaviour	0.46 (0.74)	.83
Information seeking behaviour	1.60 (0.96)	.85
Thrill-seeking behaviour	1.19 (0.93)	.77

Note. Range = 0-4.

Study 1b

In Study 1a, participants moved through images at their own pace for 5-min. This approach was ecologically valid because it allowed participants to decide when to move onto the next image, and we could equalise time on task across participants. However, we could not control the number of images/screens participants saw, or the number of behavioural choices participants had to make, meaning there was substantial variability. Thus, participants' behaviour may have changed due to the number of images/screens they saw. Moreover, although we explicitly instructed participants not to, some participants rushed through the task: the average time to uncover (or move to the next image) was < 1s for 8% of participants (M = 8.3, SD = 15.9; collapsed across image valence: M = 4.0, SD = 3.6). Therefore, some participants' responses may have reflected inattention, rather than their typical Instagram behaviour. In Study 1b we fixed the number of images, so participants made the same number of behavioural choices; we also introduced an image-response delay to limit participants' ability to rush through images.[11] We also updated our reasons for

[11]We were interested in whether pre-existing individual characteristics (e.g., intolerance to uncertainty) relate to uncovering behaviour (or reasons for uncovering). However, we found limited evidence of meaningful relationships, so we report these analyses in the supplementary materials. It is possible that individual characteristics are important, but that the relationships are more *nuanced* than we captured here (e.g., they may interact with state and contextual factors).

uncovering questionnaire: we excluded three items that interfered with 'simple structure' in the previous PCA and included six additional items based on participants' qualitative responses to why they uncovered screens in Study 1a (see data on OSF for more information: https://osf.io/fhysr). We planned to examine the same vulnerability characteristics as in Study 1a.

Based on Study 1a's findings, we predicted most participants would uncover the first sensitive-content screen they saw, and many participants would continue to uncover most, if not all, subsequent screens. We had competing hypotheses regarding the relationship between vulnerability characteristics and uncovering behaviour. Based on previous literature and theory, we predicted people higher in vulnerability characteristics (e.g., depression; post-traumatic stress symptoms, history of self-triggering) would be more likely to uncover screens (i.e., a positive relationship between uncovering and vulnerabilities). However, based on Study 1a, we predicted people higher in vulnerability characteristics would uncover a similar number of screens than people lower in vulnerability characteristics (i.e., no relationship between uncovering and vulnerabilities). Again, we made no predictions regarding the reasons why people uncovered screens since this was still an exploratory interest, but we expected several factors, like those in Study 1a, to emerge.

Method

We preregistered this study (https://osf.io/r7hf6), and report all measures, conditions, and exclusions. The supplementary materials are at the end of the chapter and the data, including a codebook describing all variables, can be found at: https://osf.io/rj987/.

Participants

As in Study 1a, we recruited participants online through MTurk and used the same screening measures. As per the addendum to our pre-registration (https://osf.io/46zdb), 101 participants were exited from the study after the Instagram task because they endorsed one or more of the following demand items: I only uncovered the sensitive-content screens because

(1) I thought I was supposed to uncover the screens, (2) I thought the study might have hidden requirements, or (3) I thought there would be a penalty for not uncovering. A further 85 participants elected to discontinue the study (when given the opportunity to complete additional questionnaires for bonus reimbursement or to discontinue). These participants, regardless of whether they were exited or elected to discontinue, received a payment of $0.70USD each. Of the participants who elected to continue, we excluded one participant who experienced technical issues. Our analyses focused on the remaining 264 participants (though we included the additional Instagram task data from participants who only completed this task [n = 85] when reporting the descriptive information from this task and demographics). Participants who completed both parts of the study received $2.50 USD.

Participants—who completed the Instagram task (N = 349)—were 18–72 years (M = 36.0, SD = 10.0) and mostly female (59.6%; 39.3% male; one participant did not report gender and three identified as non-binary). Our sample was predominantly European American/White (69.6%, n = 243); other participants were of African American/Black (10.0%), Asian (6.3%), and Hispanic/Latino (5.4%), or other (5.7%; e.g., multiracial/biracial) descent; 2.9% of participants specified nationality (e.g., American). Most participants (49.0%) reported between $45,000–$140,000 income and tended to be (52.7%) college graduates.

Measures

Instagram Task.

We selected the 30 most negative, positive, and neutral NAPS images (90 total) and presented them one at a time (in a randomised order). After 3s, the response options appeared; for neutral and positive images, participants selected *Next Photo* to move to the next image, and for sensitive-content screens, participants had the option select *See Photo* and view the image underneath or select *Next Photo* and move to the next image.

As in Study 1a, participants completed measures of Instagram use and vulnerability characteristics; specifically, depression, anxiety, and stress symptoms (DASS-21; Study 1b: depression, α = .94; anxiety, α = .88; stress, α = .90), wellbeing (SGWB-14: α = .94), PTSD symptoms (PCL-5: α = .95), and history of self-triggering. Participants who uncovered at least one sensitive-content screen also completed our updated reasons for uncovering questionnaire (see Table 1.6 for α). To reducing participant load, we used a single-item assessment of trauma exposure (vs. THS used in Study 1a). Participants reported their most stressful/traumatic event, and indicated (yes/no), if, during that event, they were directly or indirectly exposed to: actual or threatened death/injury, or actual or threatened sexual violence.

Procedure

We used the same cover story as Study 1a. After consent, all participants completed demographic and Instagram use questions, as well as items designed to reduce suspicion and demand. Participants then completed the Instagram task. Next, all participants completed feedback questions, along with demand items: participants who endorsed (one or more) demand items were excluded and finished here with full debrief; all other participants had the opportunity to continue or discontinue. Participants who opted to continue completed the vulnerability measures, and *if* they uncovered at least one sensitive-content screen they also completed the reasons for uncovering questionnaire.[12] Finally, we asked participants if they had any technical issues or stopped the task for any extensive period: six (2.3%) participants reported leaving the task (one for a few minutes, three for > a few minutes, and two did not report how long). We fully debriefed the remaining participants.

[12] Participants also completed individual difference measures here. Because there were no meaningful order effects in Study 1a, we opted not to counterbalance the order of the vulnerability/individual difference measures and the Instagram task, and instead presented the measures *after* the Instagram task so we could exit participants who endorsed demand items at this point.

Statistical Analyses

We used the same statistical approach as in Study 1a. Here, the number of screens uncovered was not normally distributed (skewness ± SE: 0.428 ± 0.131; kurtosis ± SE: 1.491 ± 0.260): visual inspection of the histogram revealed a multimodal distribution (even after transformation). Therefore, we ran non-parametric tests (i.e., Spearman's correlations) for this variable. We also dichotomised uncovering behaviour (uncovered all or not) so we could compare between studies. To see if similar reasons factors emerged in this study, we ran another PCA on the reasons for uncovering questionnaire using items from the final factors in Study 1a. Although participants endorsed the other, new items to varying degrees (M = 0.7-2.3; Supplementary Table S1.3), we realised after collecting data (prior to running the PCA) that it was likely that we had not included enough items for each new theme to properly explore them as potential factors. Again, we assessed the suitability of PCA and found the data were likely factorisable (see supplementary materials). We ran two PCA because several items interfered with 'simple structure' in the first (Thurstone, 1947); we reran the second PCA without these items).

Results and Discussion

Preliminary Analyses

Participant Characteristics.

We first examined our sample for vulnerability characteristics (Table 1.4). Overall, 76.5% (n = 202) of participants self-reported experiencing actual or threatened death/injury; the most common event was the sudden death of a close family member/friend. Further, 23.5% (n = 62) of participants met criteria for probable PTSD (Bovin et al., 2016) and 27.7% (n = 73) of participants indicated they had self-triggered with reminders of their most stressful/traumatic experience; of these, 46.6% (n = 34) had self-triggered at least 2-3 times over the past month. Moreover, 23.5% (n = 62) of participants were experiencing severe to extremely severe distress associated with symptoms related to depression (Lovibond &

Lovibond, 1995). We next examined Instagram use. Most participants reported using Instagram every day in the last 7 days (56.4%; n = 197), and on average, using it for more than one hour a day (in the last 30 days; 67.0%; n = 234). Additionally, 57.0% of participants (n = 199) reported they have seen sensitive-content screens on their Instagram feed.

Hypothesis Testing

Uncovering Behaviour.

Recall we were interested in the frequency of uncovering behaviour. For the next analyses, we included all participants who completed the Instagram task (N = 349). Consistent with Study 1a, most participants (i.e., 86.0%; n = 300) uncovered the first screen they came across. Of these participants, 62 (17.5% of the total sample) continued to uncover every subsequent screen (i.e., the remaining 29), but most (n = 287; 82.2% of the total sample) left at least one subsequent screen covered. On average, participants uncovered 13.14 screens (out of 30; SD = 11.42) during the task. Of the 49 participants who left the first screen covered, only 17 (4.9% of the total sample) left all subsequent screens covered. Thus, although behaviour on the first screen was consistent with Study 1a, there was a significant difference between the percentages of uncovering over the entire task in Studies 1 (51.7%) and 2 (17.5%), χ^2 (1) = 79.0, p < .001, φ = -.36; here, avoidance was higher over the entire task.

Table 1.4

Means (and Standard Deviations) for Vulnerability Measures

Measure	Scale (Range)	*M* (*SD*)
DASS-21	Depression (0–21)	6.1 (5.9)
	Anxiety (0–21)	4.4 (4.6)
	Stress (0–21)	6.8 (5.0)
SGWB-14	(14–70)	46.1 (12.4)
PCL-5	Intrusions (0–20)	4.9 (4.3)
	Avoidance (0–8)	2.8 (2.4)
	Negative Cognition/Mood (0–28)	6.8 (6.0)
	Hyperarousal (0–24)	5.9 (5.5)
	Total (0–80)	20.5 (16.4)

Note. n = 264. DASS-21 = Depression Anxiety Stress Scales-21; SGWB-14 = Scales of General Well-Being**;** PCL-5 = Posttraumatic Stress Disorder Checklist**.**

It is possible that participants rushed through the task in Study 1a to complete it as quickly as possible—indeed, such behaviour was the reason for methodological changes in Study 1b. Although we cannot rule out this possibility, there are other potential reasons why uncovering behaviour differed between Studies 1 and 2. A second possibility is there were systematic differences between samples. However, the data do not support this explanation: as shown in Tables 1.1 and 1.4, the samples are comparable on characteristics we measured, albeit with slightly higher scores in Study 1b. Though this slight elevation in vulnerabilities may stem from the *ongoing* impacts of COVID-19, both studies were collected during the pandemic. A third—and more likely—possibility relates to the nature of the images. In Study 1a, we used a pool of 70 negative images, from which people saw a subset of varying size and content. In Study 1b, all participants saw the *most negative* 30 of these 70 images. Thus, the images participants uncovered in Study 1b were likely to be more negative, which may

have made participants less likely to uncover them, especially images after those they uncovered initially. Because the images were more negative, we wondered whether participants avoided images based on what they could identify about the blurred image. However, no images appeared "unique" in their uncovering: the number of people who uncovered each screen was relatively consistent (M = 131.6, SD = 11.7, range = 109-151). A fourth possibility relates to the number of screens participants saw (Study 1a: M = 20; Study 1b: 30). Perhaps there was greater opportunity for people's curiosity to 'wear' off over time and for uncovering behaviour to decline in Study 1b (Day, 1982). The data pattern in Study 1b supports this possibility; the percentage of participants who uncovered each screen was initially high (51.9% -86% over the first five screens), but steadily decreased until it plateaued over the final 10 screens (just below 30%; see supplementary materials, Figure S1.1). Finally, a combination of these factors may explain the discrepancy in uncovering behaviour across studies.

Uncovering Behaviour and Vulnerability Characteristics.

Like Study 1a, we aimed to see if vulnerable people are particularly susceptible to uncovering screens. Therefore, we next examined whether vulnerability characteristics were associated with uncovering behaviour (Table 1.5). Consistent with Study 1a, there were no relationships between the vulnerability measures (i.e., depression, anxiety, stress, overall PTSD, self-triggering frequency, or wellbeing) and uncovering behaviours, with one exception: participants who uncovered *all* sensitive-content screens during the task reported less avoidance of their trauma-related thoughts, feelings, or external reminders after the trauma; a negative correlation between the avoidance subscale of the PCL-5 and whether participants uncovered all sensitive-content screens (or not) during the task. This finding is unsurprising given uncovering screens could be considered as an approach behaviour, rather than avoidance; however, it is possible this effect reflects a Type 1 error given the high number of correlations. Taken together, there is no evidence to suggest vulnerable people are

more susceptible to approaching sensitive content: people appear to behave similarly *irrespective* of vulnerabilities.

Reasons for Uncovering.

Next, we examined our PCA. Six components had eigenvalues > one, explaining 24.6%, 17.4%, 8.9%, 6.7%, 4.4% and 3.9% of the total variance, respectively. However, visual inspection of the scree plot indicated four components should be retained (explaining 57.6% of the total variance): a four-component solution was also interpretable/exhibited 'simple structure' (Thurstone, 1947; Supplementary Table S1. 4). Therefore, we retained four factors. Like Study 1a, the factors from were *avoidance behaviour*, *information seeking behaviour*, *negative affect driven behaviour,* and *positive affect driven behaviour* (Table 1.6).[13] The thrill-seeking behaviour factor collapsed into the information seeking behaviour factor in Study 1b. Items within each factor were mostly consistent with Study 1a (see supplementary materials for items). Thus, four reasons seemingly underpinned participants' decisions to uncover sensitive-content screens or not.

[13] See supplementary materials for exploratory correlations between reasons factors and vulnerability measures.

Table 1.5

Correlations Between Uncovering Behaviour and Vulnerability Measures

Measure	Scale	Type of Uncovering Behaviour		
		First image r_{pb} [95% CI]	Total screens uncovered r_s [95% CI]	Uncovered all (or not) r_{pb} [95% CI]
DASS-21	Depression	.07 [-.05, .19]	.11 [-.01, .23]	.03 [-.09, .15]
	Anxiety	.02 [-.10, .14]	.06 [-.06, .18]	-.04 [-.16, .08]
	Stress	.06 [-.06, .18]	.08 [-.04, .20]	-.003 [-.12, .12]
SGWB-14		-.12 [-.24, .001]	-.09 [-.21, .03]	-.06 [-.18, .06]
PCL-5	Intrusions	.01 [-.11, .13]	-.04 [-.16, .08]	-.09 [-.21, .03]
	Avoidance	-.04 [-.16, .08]	-.02 [-.14, .10]	-.13* [-.25, -.01]
	Neg Cog/Mood	.12 [-.001, .24]	.09 [-.03, .21]	-.001 [-.12, .12]
	Hyperarousal	.09 [-.03, .21]	.06 [-.06, .18]	-.01 [-.13, .11]
	Total	.07 [-.05, .19]	.04 [-.08, .16]	-.05 [-.17, .07]
STQ	Freq. of ST	-.001 [-.23, .23]	.15 [.03, .27]	.08 [-.15, .30]

Note. n = 263, except frequency of ST (=self-triggering) n = 73. DASS-21 = Depression Anxiety Stress Scales-21; SGWB-14 = Scales of General Well-Being; PCL-5 = Posttraumatic Stress Disorder Checklist. STQ = Self-Triggering Questionnaire. * p <.05

Table 1.6

Means (and Standard Deviations) for Reason for Uncovering

Reasons	*M* (*SD*)	α
Avoidance behaviour	1.52 (1.19)	.89
Negative affect driven behaviour	0.14 (0.37)	.87
Positive affect driven behaviour	0.48 (0.71)	.73
Information seeking behaviour	1.67 (0.95)	.87

Note. Range = 0-4.

General Discussion

One of the reasons Instagram introduced sensitive-content screens was to deter people—particularly vulnerable people—from viewing sensitive content. However, across two studies we found most participants—who were predominantly young (~36 years) European American/White female Instagram users—opted to uncover the first sensitive screen they discovered, consistent with Bridgland, Bellet et al. (2022). Many people were also willing to *repeatedly* uncover screens, even when the images beneath were likely distressing for them. We found no evidence vulnerable people were more likely to avoid sensitive content: rather, people similarly uncovered screens irrespective of their vulnerabilities. However, we note the order effect of PTSD symptom severity in Study 1a may indicate task order (in some way) influenced participants' reporting of vulnerabilities (even though there was no impact on uncovering behaviour). Information seeking behaviour and negative affect driven behaviour appeared to be the most important reasons why people uncovered screens: they were the most strongly endorsed (information seeking behaviour more so than negative affect driven behaviour, noting our non-clinical sample).

Because some participants viewed objectively fewer sensitive images (i.e., by opting not to uncover some images), one interpretation of our uncovering data is that sensitive-content screens are "effective"—because they helped a small fraction of people to avoid a small fraction of images. However, another interpretation of our data—which we consider to be more parsimonious—is that they offer *too little* benefit to be considered "effective", especially given the possible consequences of repeated exposure to sensitive content (e.g., distress/PTSD-like symptoms) outside of a therapeutic setting.

Thus, overall, our results demonstrate that sensitive-content screens may be ineffective in deterring people from viewing sensitive content. Our data align with research demonstrating trigger warnings may be ineffective at prompting avoidance (e.g., Kimble et al., 2021), may make content more attractive (i.e., "forbidden fruit effect"; Weaver, 2011),

and with the broader finding people often *intentionally* expose themselves to negative and potentially distressing content (Oosterwijk, 2017). Indeed, the general and non-specific nature of the "Sensitive Content" warning (i.e., it has no link to content type) may foster curiosity and encourage people to engage with screened negative content (Loewenstein, 1994).

Our data also have important implications. Repeated exposure to sensitive content may become clinically consequential by affecting large numbers of people, and by having cascading effects on users' other social media behaviour (e.g., increasing self-triggering; Anvari et al., 2022; Funder & Ozer, 2019). Therefore, these findings may have specific implications for Instagram and other social media platforms that employ similar deterrence tools (e.g., TikTok)—though we acknowledge that these platforms evolve quickly, and our findings may not be relevant to future platforms if the user experience changes significantly.

Our findings also suggest vulnerable people are no more susceptible (than people lower in vulnerability characteristics) to uncover sensitive-content screens, but they also appear no more likely to *avoid* sensitive content. This result aligns with recent research (Bridgland, Bellet et al., 2022; Study 1b) demonstrating vulnerability measures are *not* related to uncovering behaviour. However, these findings are at odds with previous research showing people with depression and PTSD symptomology often seek out negative content (e.g., Millgram et al., 2015; Bellet, Jones & McNally, 2020). Potentially there is no relationship between the vulnerabilites we measured and uncovering behaviour. However, there are several other explanations. First, perhaps the warning included in the study advertisement and consent (e.g., "A small minority of people also experience distressing memories …") filtered out people who were most likely to seek out distressing content (Bethlehem, 2010). This warning might also have influenced the rate of uncovering altogether: participants may have been more/less motivated to view sensitive content than they would be in "real life" without additional warnings. It is also possible that the non-

specific nature of the "Sensitive Content" warning prevented people with vulnerabilities from using the screens to avoid content specifically relevant to their trauma that could be triggering (e.g., people with PTSD from a fire related event avoiding images related to fire). However, much like the sensitive-content screens tested here, user generated trigger warnings are often vague and non-specific. Therefore, many vulnerable people make decisions about approaching or avoiding potentially triggering content, daily, with little information. A logical extension of this work—which would have implications for Instagram's sensitive-content screen format—would be to vary the warnings to include nature of the content (e.g., "this image contains violence") and examine whether behaviour changes uniquely for people with personally relevant trauma.

Second, although the PCL-5 and DASS-21 cut-offs suggested a subset of our sample were experiencing probable PTSD and significant distress related to symptoms of depression, we did not sample a clinical population. It is possible that despite meeting cut-offs, some of these participants would not meet criteria for a formal diagnosis using a structured clinical interview. However, bearing this limitation in mind, we note MTurk has been identified as an excellent source for studying clinical and subclinical populations (Shapiro et al., 2013), and our estimates compare with those in previous work using the same recruitment platform (e.g., van Stolk-Cooke et al., 2018). Nonetheless, future research could pre-screen participants to recruit a specific clinical population (e.g., people with PTSD).

Our study has several other limitations. First, although we included the Instagram logo and "like"/comment buttons to replicate Instagram, other features (e.g., captions) were missing. Relatedly, one of the primary purposes of sensitive-content screens is to manage non-graphic self-injury related content. Therefore, although the content we used was negative in nature and would likely be screened by Instagram, we did not include images related explicitly to self-injury. Future studies could include such features and content. Second, we examined whether sensitive-content screens deter people from viewing sensitive content, but

we did not examine how well screens *prepare* people for upcoming sensitive content—an equally important aim of such screens. It is possible that sensitive-content screens may influence people's subsequent emotional reactions to sensitive content, especially when people are given an explicit choice to view the content or not. Indeed, any benefits of preparation or choice on emotional reactions (e.g., reductions in distress) may bear the consequences of screens failing to deter users altogether (though we note trigger warning literature suggests otherwise, e.g., Bridgland et al., 2019; Bridgland, Barnard et al., 2022; Sanson, et al., 2019). Nonetheless, future research should examine how well screens prompt preparation by assessing the subsequent effects on people's emotional reactions, perhaps using a paradigm where participants have a choice whether to view the content or not. Third, participants may have endorsed reasons for uncovering that appear intuitive in retrospect but were not reflective of their motivations during the task. It is also possible that participants selectively reported reasons to rationalise their behaviour and avoid cognitive dissonance (Festinger, 1957). Future research could examine uncovering behaviour after manipulating proposed reasons (e.g., by varying image blur to induce different levels of uncertainty). Future research could also focus on the minority of people who do not choose to uncover sensitive-content screens; this work may reveal unique information with respect to factors that 1) make someone less likely to uncover such screens and 2) set these people apart from the (majority of) people who uncover such screens. Finally, we acknowledge that our sample was not demographically diverse, meaning that our results may not generalise to other populations that vary in such variables (e.g., age, education, socioeconomic status etc.).

Across two studies we examined whether sensitive-content screens deter people from viewing sensitive content. Our findings suggest that sensitive-content screens may be ineffective in deterring people—including people with mental health vulnerabilities—from viewing such content. Furthermore, four distinct reasons appear to underpin people's

decisions to uncover sensitive-content screens. Social media platforms may need to adapt their sensitive-content screens to deter people from viewing sensitive content.

Supplementary Materials

Table S1.1

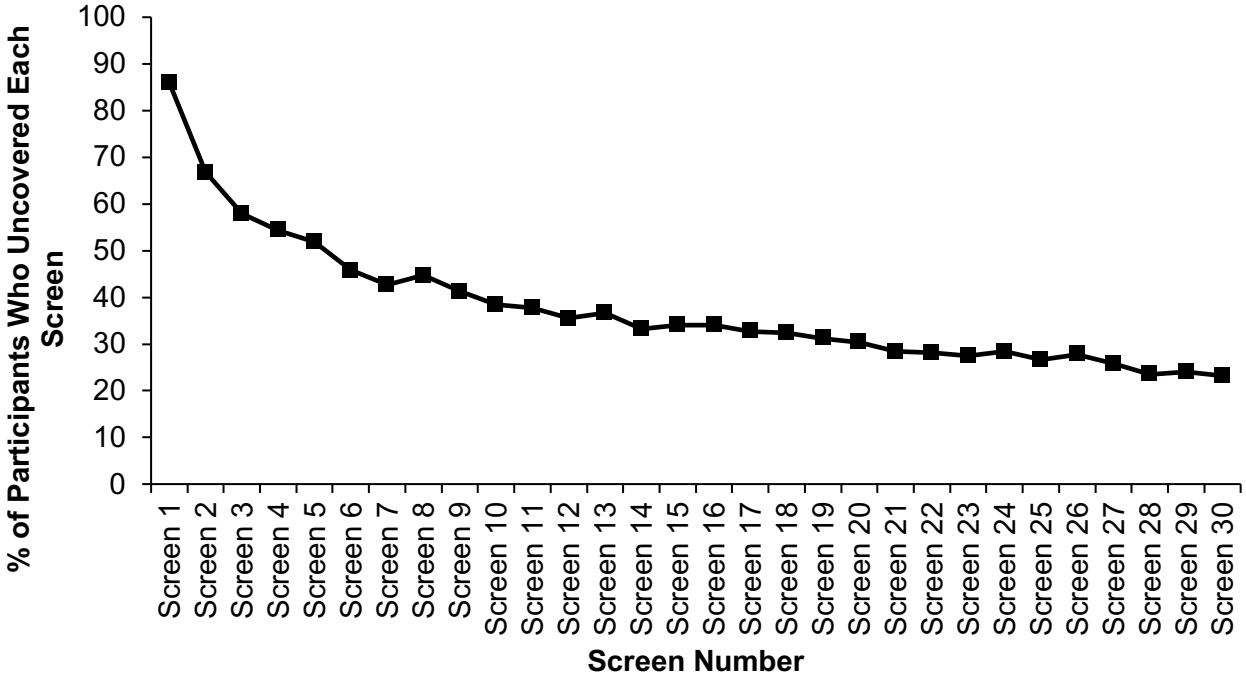
Study 1a: Correlations Between Uncovering Behaviour and Centrality of Events

		Type of Uncovering Behaviour				
Measure	Scale	First image r_{pb} [95% CI]	Proportion uncovered r_s [95% CI]	Uncovered all (or not) r_{pb} [95% CI]	Total images viewed r_s [95% CI]	Total screens viewed r_s [95% CI]
CES		-.05 [-.17, .07]	.02 [-.10, .14]	.02 [-.10, .14]	-.05 [-.17, .07]	-.04 [-.16, .08]

Note. n = 263, CES = Centrality of Events.

Figure S1.1

Study 1a: The Percentage of Participants who Uncovered Each Sensitive-Content Screen as a Function of the Screen Number (i.e., When the Image was Presented)



Reasons for Uncovering Sensitive-Content Screens

Study 1a: PCA

We assessed the suitability of PCA prior to analysis (Field, 2013), and the data were likely factorisable. Inspection of the correlation matrix showed that all variables had at least one correlation coefficient greater than 0.4. The overall Kaiser-Meyer-Olkin (KMO) measure was .91 with individual KMO measures all greater than .84, indicating 'meritorious' to 'marvelous' classifications (Kaiser, 1974). Bartlett's Test of Sphericity was statistically significant (p <.001).

The first PCA revealed five components, however the Direct Oblimin rotation used to aid interpretability revealed two items ("I uncovered the screened image(s) because I enjoy having new and varied experiences", and "…I wanted to have an experience that matched my positive mood") loaded onto two components. Therefore, we reran a second PCA without these items. Again, there was five components, however the second Direct Oblimin rotation revealed another item ("I uncovered the screened image(s) because I was curious") loaded onto two components. Therefore, we reran the third PCA without this item (Table S3).

Study 1b: PCA

We assessed the suitability of PCA prior to analysis (Field, 2013), and the data were likely factorisable. Inspection of the correlation matrix showed that nearly all variables had at least one correlation coefficient greater than 0.4. The overall Kaiser-Meyer-Olkin (KMO) measure was .86 with individual KMO measures all greater than .75, indicating 'middling' to 'marvelous' classifications (Kaiser, 1974). Bartlett's Test of Sphericity was statistically significant (p <.001).

We first ran a PCA, including only those items from the final factors in Study 1a. The first PCA revealed six components, however the Direct Oblimin rotation used to aid interpretability revealed two items ("I uncovered the screened image(s) because it was thrilling/exhilarating to do so" and "…I was excited to see what might lie beneath the screen"

loaded onto two components, interfering with 'simple structure' (Thurstone, 1947). Therefore, we reran a second PCA without these items (Table S4).

The two components we did not retain from the second PCA appeared to be a second avoidance behaviour factor—tied to disinterest, rather than affect—and a past experiences factor. Although participants endorsed the items in these components, endorsement was low overall (M = 0.3-1.2; range = 0-4) and explained little variance (3.9% and 4.1%). Future research should explore these themes further.

Table S1.2

Study 1a: Factor Loadings and Communalities Based on the Third PCA with Direct Oblimin Rotation

Item	Rotated component coefficients					
	C1	C2	C3	C4	C5	C
I uncovered the screened image(s) because…						
I was trying to remind myself of past negative experiences.	**.878**	-.032	-.003	-.037	-.056	.790
I was trying to make sense of my past negative experiences.	**.841**	-.029	.194	-0.29	.299	.628
I wanted to have an experience that matched my negative mood.	**.832**	-.026	-.042	.085	-.030	.735
I was trying to prevent the memories of my past negative experiences fading.	**.741**	.084	.030	.012	-.141	.711
I was feeling down and blue.	**.735**	.094	-.013	-.071	-.120	.699
I was sad.	**.724**	-.008	-.095	.082	-.114	.620
I was unhappy.	**.718**	-.042	-.059	.010	-.167	.624
I did not uncover the screened image(s) because…						
I make an effort to avoid distressing and graphic material.	-.102	**.923**	-.028	-.051	.104	.869
I do not like viewing distressing or graphic material.	-.087	**.919**	-.008	-.009	.125	.823
I thought the material underneath the screen would make me upset	-.103	**.900**	.036	-.060	.114	.799

Item	Rotated component coefficients					
	C1	C2	C3	C4	C5	C
I trust they were covered for my own good.	-.035	**.848**	-.038	.004	.032	.720
I don't enjoy taking risks.	.105	**.776**	.036	.007	.027	.627
I was uninterested.	-.006	**.647**	-.065	.051	-.112	.452
I thought I knew what the image was.	.051	**.583**	.107	-.097	-.181	.438
I thought it would remind me of a past negative experience.	.352	**.520**	-.035	.078	-.147	.562
(I uncovered the screened image(s) because) I was trying to forget my past negative experiences.	.374	**.408**	-.130	.049	-.242	.550
I uncovered the screened image(s) because…						
I was excited to see what might lie beneath the screen.	.007	.001	**.850**	-.105	-.173	.745
I was eager to learn what the image was.	-.102	-.101	**.716**	.214	.003	.712
It was thrilling/exhilarating to do so.	.217	-.005	**.596**	.020	-.226	.601
I uncovered the screened image(s) because my freedom to view the image was restricted.	.286	.017	**.422**	.267	.018	.477
I was uncomfortable when I didn't know what the image was.	.079	.059	-.089	**.838**	.034	.658
I do not enjoy ambiguity.	.029	-.073	-.198	**.805**	-.116	.650
I was frustrated that I couldn't see the image	.066	-.040	.171	**.679**	-.052	.647
I wanted to reduce uncertainty associated with the covered image	-.013	-.013	.295	**.611**	.139	.572
I wanted to act with my own free will.	-.088	-.130	.342	**.499**	-.149	.605
I wanted to know why it was covered.	-.124	-.216	.367	**.478**	.011	.623
I was feeling good.	.005	.059	.114	.089	**-.833**	.798
I was content.	.024	-.043	.158	.067	**-.797**	.756
I was happy.	.350	-.056	.095	-.076	**-.594**	.658

Note. Based on $n = 263$. Major loadings are bolded. C1-C5 = Component 1 to Component 5. C = Communalities.

Table S1.3

Study 1b: Mean Endorsement for New Items Not Included in PCA

Item	M (SD)
I did not uncover the screened image(s) because I was afraid of what I might see.	1.9 (1.6)
I uncovered the screened image(s) because I did not think the images would be negative.	0.9 (1.1)
I uncovered the screened image(s) because I wanted to build a realistic understanding of the world.	1.1 (1.3)
I uncovered the screened image(s) because I was interested to see what the image was.	2.3 (1.3)
I uncovered the screened image(s) because I know I can cope with negative content.	1.8 (1.4)
I uncovered the screened image(s) because I want to prepare for negative events through the experience of others.	0.7 (1.1)

Table S1.4

Study 1b: Factor Loadings and Communalities Based on the Second PCA with Direct Oblimin Rotation

Item	Rotated component coefficients						
	C1	C2	C3	C4	C5	C6	C
I did not uncover the screened image(s) because…							
I make an effort to avoid distressing and graphic material.	**-.820**	-.123	-.118	-.003	.152	-.072	.830
I thought the material underneath the screen would make me upset	**-.798**	-.089	-.022	-.076	.164	-.102	.784
I trust they were covered for my own good.	**-.791**	-.061	-.082	.044	-.035	-.124	.659
I do not like viewing distressing or graphic material.	**-.764**	-.099	-.031	-.137	.254	.075	.819
I don't enjoy taking risks.	**-.695**	.159	-.001	.047	-.052	.121	.502
I uncovered the screened image(s) because…							
I was unhappy.	-.112	**.825**	.066	.124	-.040	.251	.742

Item	Rotated component coefficients						
	C1	C2	C3	C4	C5	C6	C
I wanted to have an experience that matched my negative mood.	.063	**.795**	-.017	.060	-.036	-.065	.675
I was feeling down and blue.	-.138	**.777**	.131	.039	.004	.158	.646
I was sad.	-.008	**.748**	.043	-.076	.042	-.155	.630
I was trying to remind myself of past negative experiences.	.218	**.707**	-.067	-.013	.091	-.266	.654
I was trying to prevent the memories of my past negative experiences fading.	.130	**.659**	-.011	.008	.090	-.280	.598
I was frustrated that I couldn't see the image	.073	.098	**.781**	.115	.046	.024	.705
I was uncomfortable when I didn't know what the image was.	-.230	.106	**.774**	-.130	-.266	.028	.680
I do not enjoy ambiguity.	.085	.135	**.715**	-.015	.017	.069	.565
I wanted to know why it was covered.	.242	-.048	**.713**	-.069	.164	.111	.605
I wanted to reduce uncertainty associated with the covered image	-.082	-.072	**.683**	-.127	-.012	-.362	.598
I was eager to learn what the image was.	.092	-.077	**.599**	.361	-.056	.015	.632
I uncovered the screened image(s) because my freedom to view the image was restricted.	.135	-.008	**.522**	.306	.101	.071	.544
I wanted to act with my own free will.	.235	-.086	**.491**	.365	-.042	-.145	.620
I was feeling good.	-.063	.118	-.018	**.858**	.010	-.015	.768
I was content.	.089	-.223	.122	**.776**	.036	-.116	.668
I was happy.	-.077	.211	-.132	**.741**	-.023	.076	.625
(I did not uncover the screened images because…) I was uninterested.	-.058	.004	-.083	.047	**.828**	.232	.760
(I did not uncover the screened images because…) I thought I knew what the image was.	-.194	.081	.100	-.020	**.709**	-215	.677

Item	Rotated component coefficients						
	C1	C2	C3	C4	C5	C6	C
I was trying to make sense of my past negative experiences.	-.001	.344	-.056	.096	-.075	**-.652**	.650
I thought it would remind me of a past negative experience.	-.371	.155	.066	.040	.050	**-.510**	.505

Note. Based on $n = 264$. Major loadings are bolded. C1-C6 = Component 1 to Component 6. C = Communalities.

Study 1a: Final Factors and Items

Factor 1: Negative affect driven behaviour

1. I uncovered the screened image(s) because I was trying to make sense of my past negative experiences.
2. I uncovered the screened image(s) because I was sad.
3. I uncovered the screened image(s) because I wanted to have an experience that matched my negative mood.
4. I uncovered the screened image(s) because I was trying to remind myself of past negative experiences.
5. I uncovered the screened image(s) because I was trying to prevent the memories of my past negative experiences fading.
6. I uncovered the screened image(s) because I was unhappy.
7. I uncovered the screened image(s) because I was feeling down and blue.

Factor 2: Avoidance behaviour

1. I did not uncover the screened image(s) because I was uninterested.
2. I did not uncover the screened image(s) because I trust they were covered for my own good.
3. I did not uncover the screened image(s) because I thought I knew what the image was.
4. I did not uncover the screened image(s) because I thought it would remind me of a past negative experience.
5. I did not uncover the screened image(s) because I don’t enjoy taking risks.
6. I did not uncover the screened image(s) because I do not like viewing distressing or graphic material.
7. I did not uncover the screened image(s) because I make an effort to avoid distressing and graphic material.
8. I did not uncover the screened image(s) because I thought the material underneath the screen would make me upset.

Factor 3: Information seeking behaviour

1. I uncovered the screened image(s) because I wanted to reduce uncertainty associated with the covered image.
2. I uncovered the screened image(s) because I was frustrated that I couldn't see the image.
3. I uncovered the screened image(s) because I do not enjoy ambiguity.
4. I uncovered the screened image(s) because I was uncomfortable when I didn't know what the image was.
5. I uncovered the screened image(s) because I wanted to know why it was covered.
6. I uncovered the screened image(s) because I wanted to act with my own free will.

Factor 4: Thrill-seeking behaviour

1. I uncovered the screened image(s) because I was excited to see what might lie beneath the screen.
2. I uncovered the screened image(s) because it was thrilling/exhilarating to do so.
3. I uncovered the screened image(s) because I was eager to learn what the image was.
4. I uncovered the screened image(s) because my freedom to view the image was restricted.

Factor 5: Positive affect driven behaviour

1. I uncovered the screened image(s) because I was content.
2. I uncovered the screened image(s) because I was feeling good.
3. I uncovered the screened image(s) because I was happy.

Study 1b: Final Factors and Items

Factor 1: Avoidance behaviour

1. I did not uncover the screened image(s) because I thought the material underneath the screen would make me upset.
2. I did not uncover the screened image(s) because I make an effort to avoid distressing and graphic material.
3. I did not uncover the screened image(s) because I trust they were covered for my own good.
4. I did not uncover the screened image(s) because I do not like viewing distressing or graphic material.
5. I did not uncover the screened image(s) because I don't enjoy taking risks.

Factor 2: Negative affect driven behaviour

1. I uncovered the screened image(s) because I was unhappy.
2. I uncovered the screened image(s) because I wanted to have an experience that matched my negative mood.
3. I uncovered the screened image(s) because I was feeling down and blue.
4. I uncovered the screened image(s) because I was sad.

5. I uncovered the screened image(s) because I was trying to remind myself of past negative experiences.
6. I uncovered the screened image(s) because I was trying to prevent the memories of my past negative experiences fading.

Factor 3: Information seeking behaviour

1. I uncovered the screened image(s) because I was frustrated that I couldn't see the image.
2. I uncovered the screened image(s) because I do not enjoy ambiguity.
3. I uncovered the screened image(s) because I was uncomfortable when I didn't know what the image was.
4. I uncovered the screened image(s) because I wanted to know why it was covered.
5. I uncovered the screened image(s) because I wanted to reduce uncertainty associated with the covered image.
6. I uncovered the screened image(s) because I was eager to learn what the image was.
7. I uncovered the screened image(s) because my freedom to view the image was restricted.
8. I uncovered the screened image(s) because I wanted to act with my own free will.

Factor 4: Positive affect driven behaviour

1. I uncovered the screened image(s) because I was content.
2. I uncovered the screened image(s) because I was feeling good.
3. I uncovered the screened image(s) because I was happy.

Reasons for Uncovering Sensitive-Content Screens and Vulnerability Measures

To see if vulnerable people endorse certain reasons for their uncovering behaviour, we ran a series of correlations between vulnerability measures and the three *approach-based* reason factors (separately for each study; see Tables S1.5 & S1.6). In Study 1a, the higher people's depression the more likely they were to endorse negative and positive affect driven behaviour, and information seeking behaviour. Comparatively, in Study 1b, higher depression was only related to negative affect driven behaviour and information seeking behaviour. In Study 1a, there was no relationship between people's overall PTSD symptomology and their reasons for uncovering screens. However, in Study 1b, the higher people's PTSD symptomology the more likely they were to endorse negative and positive affect driven behaviour, and information seeking behaviour. In Study 1a, for the people who had self-triggered (n = 49), the higher the frequency of self-triggering behaviours the more likely they

were to endorse negative affect driven behaviour. However, in Study 1b, there were no relationships between self-triggering and any of the reasons. Finally, in Study 1a, the higher people's wellbeing the less likely they were to endorse negative affect driven behaviour, whereas in Study 1b, the higher people's wellbeing the more likely they were to endorse positive affect driven behaviour.

There are several discrepancies between Studies 1a and 1b. Possibly, methodological changes between the two studies (e.g., the number/nature of images) influenced participants' reasons for uncovering screens, alongside their actual uncovering behaviour. For example, in Study 1b there may have been greater opportunity for people to gain information/for their curiosity to 'wear' off. However, our data do not support this explanation: as shown in Tables 3 and 4, the mean level of endorsement for the three approach-based factors are comparable across studies. Alternatively, existing trait vulnerabilities (e.g., depression) may interact with state mood factors (e.g., low mood) *and* contextual motivations (e.g., being alone at night vs. with others during the day) to influence a person's uncovering decision. Therefore, even though behaviour may be similar at surface-level, possible interactions between underlying factors may mean reasons for uncovering are unique to each person/situation and thus more *nuanced* than captured here. Future research should explore this possibility.

Table S1.5

Study 1a: Correlations Between Reason for Uncovering and Vulnerability Measures

		Reasons				
Measure	Scales	Negative affect driven behaviour	Positive affect driven behaviour	Information seeking behaviour	Thrill-seeking behaviour	Avoidance behaviour
DASS-21	Depression	.30** [.19, .41]	.13* [.01, .25]	.16** [.04, .28]	.16** [.04, .28]	.20** [.08, .31]
	Anxiety	.43** [.33, .52]	.23** [.11, .34]	.19** [.07, .30]	.21** [.09, .32]	.18** [.06, .30]
	Stress	.25** [.13, .36]	.14* [.02, .26]	.23** [.11, .34]	.14** [.02, .26]	.17** [.05, .29]
SGWB-14		-.14* [-.26, -.02]	.08 [-.04, .20]	.01 [-.11, .13]	.04 [-.08, .16]	-.13* [-.25, -.01]
PCL-5	Intrusions	.26** [.14, .37]	.12 [-.01, .24]	.23** [.11, .34]	.20** [.08, .31]	.10 [-.02, .22]
	Avoidance	.23** [.11, .34]	.07 [-.05, .19]	.18** [.06, .30]	.13* [.01, .25]	.09 [-.03, .21]
	Negative Cognitions/Mood	.29** [.18, .40]	.07 [-.05, .19]	.17** [.05, .29]	.13* [.01, .25]	.16** [.04, .28]
	Hyperarousal	.33** [.22, .43]	.14* [.02, .26]	.22** [.10, .33]	.18* [.06, .30]	.12 [-.001, .24]
	Total PTSD	-.07 [-.19, .05]	-.06 [-.18, .06]	.03 [-.09, .15]	-.02 [-.14, .10]	-.11 [-.23, .01]
STQ	Frequency of self-triggering	.39** [.13, .60]	.28 [.02, .52]	.15 [-.13, .41]	.31* [.04, .54]	.12 [-.16, .39]
CES		.02 [-.10, .14]	-.09 [-.21, .03]	.15** [.03, .27]	.01 [-.11, .13]	-.01 [-.17, .07]

Note. N = 263, except for freq. of ST where n = 50. DASS-21 = Depression Anxiety Stress Scales-21; SGWB-14 = Scales of General Well-Being; PCL-5 = Posttraumatic Stress Disorder Checklist. STQ = Self-Triggering Questionnaire; CES = Centrality of Events. * p <.05, ** p <.01.

Table S1.6

Study 1b: Correlations Between Reason for Uncovering and Vulnerability Measures

		Reasons			
Measure	Scales	Negative affect driven behaviour	Positive affect driven behaviour	Information seeking behaviour	Avoidance behaviour
DASS-21	Depression	.31** [.20, .42]	.04 [-.08, .16]	.13* [.01, .25]	.05 [-.07, .17]
	Anxiety	.38** [.27, .48]	.08 [-.04, .20]	.20** [.08, .31]	-.01 [-.13, .11]
	Stress	.28** [.17, .39]	.09 [-.03, .21]	.24** [.12, .35]	.05 [-.07, .17]
SGWB-14		-.08 [-.20, .04]	.24** [.12, .35]	.05 [-.07, .17]	-.02 [-.14, .10]
PCL-5	Intrusions	.23** [.11, .34]	.10 [-.02, .22]	.20** [.08, .31]	.04 [-.08, .16]
	Avoidance	.08 [-.04, .20]	.31 [.20, .42]	.12 [-.001, .24]	.13* [.01, .25]
	Negative Cognition/Mood	.29** [.18, .40]	.09 [-.03, .21]	.16* [.04, .28]	.05 [-.07, .17]
	Hyperarousal	.32** [.21, 42]	.17** [.05, .29]	.17** [.05, .29]	.05 [-.07, .17]
	Total PTSD	.28** [.17, .39]	.13* [.01, .25]	.18** [.06, .29]	.07 [-.05, .19]
STQ	Frequency of self-triggering	.09 [-.03, 21]	.003 [-.12, .12]	.14 [.02, .26]	-.04 [-16, .08]

Note. $N = 263$, except for frequency of self-triggering where n = 50. DASS-21 = Depression Anxiety Stress Scales-21; SGWB-14 = Scales of General Well-Being; PCL-5 = Posttraumatic Stress Disorder Checklist. STQ = Self-Triggering Questionnaire; CES = Centrality of Events. * p <.05, ** p <.01.

Reasons for Uncovering Sensitive-Content Screens in Real Life

Aside from assessing participant responses to the statements aligned with what the literature predicts, we asked all participants (in Studies 1 and 1b) to think about their decision to approach sensitive content (or not) in real life (i.e., outside the Instagram task) and to respond to a series of statements (using the same 5-point scale as above). The majority of participants indicated (i.e., that it was moderately to extremely true) that they would be more likely to uncover sensitive-content screens on their own Instagram feed if they thought they would be interested in the subject matter (Study 1a: 72.6; Study 1b: 73.0%), if a caption caught their attention/was interesting to them (Study 1a: 71%; Study 1b: 71.5%), if comments caught their attention/were interesting to them (Study 1a: 66%; Study 1b: 68.2%), if they knew the posting person/account (Study 1a: 65.2%; Study 1b: 70.0%), or if they knew what the image actually was (Study 1a: 58.1%; Study 1b: 59.9%). Less frequently endorsed factors included mood (both good [Study 1a: 30.9%; Study 1b: 30.7%] and bad [Study 1a: 15.2%; Study 1b: 11.0%]) and being in the presence of others (Study 1a: 14.9%; Study 1b: 14.8%) or not (Study 1a: 48.8%; Study 1b: 41.2%). Therefore, there appears to be additional *situational* factors at play in real life which may influence a person's decision to uncover sensitive-content screens (or not). Future research should examine these factors.

Individual Characteristics

We were also interested in whether certain pre-existing individual characteristics (e.g., intolerance to uncertainty) would be related to uncovering behaviour (and/or reasons for uncovering). It is possible that people with negative beliefs about uncertainty and its implications, as well as people who have an unwillingness to remain in contact with private experiences (e.g., feelings of anxiety due to uncertainty), or people who have higher trait curiosity, may be more susceptible to uncovering sensitive screens. The same may be true for people who try to up-regulate negative emotions (i.e., increase the intensity, duration, and/or quality of those emotions) and/or down-regulate positive emotions (i.e., decrease intensity,

duration and/or quality; Tamir, 2009). Therefore, in Study 1b we measured these individual characteristics. We predicted that curiosity, intolerance to uncertainty, experiential avoidance, and emotion regulation difficulties (i.e., difficulty down-regulating negative and up-regulating positive emotions) would be positively related to uncovering behaviour.

Measures

Intolerance to Uncertainty Scale Short Form (IUS-12; Carleton et al., 2007).

Participants completed the IUS-12 to assess intolerance to uncertainty. Participants rated to what extent they agree with seven items assessing prospective anxiety (e.g., "It frustrates me not having all the information I need") and five items assessing inhibitory anxiety (e.g., "When it's time to act, uncertainty paralyses me") on a 5-point scale (1 = *not at all characteristic of me* to 5 = *entirely characteristic of me*). We summed prospective anxiety (7-35; Study 1b: α = .88;), inhibitory anxiety (5-25; α = .89) and total IUS-12 scores (12-60; α = .92).

The Five-Dimensional Curiosity Scale Revised (5DCR; Kashdan et al., 2020).

Participants completed the 24-item 5DCR to assess five dimensions of trait curiosity. Participants responded to a series of statements (e.g., "I view challenging situations as an opportunity to grow and learn") on a 7-point scale (from 1 = *does not describe me at all* to 7 = *completely describes me).* We averaged item scores for each dimension: joyous exploration (pleasurable experience of finding the world intriguing; Study 1b: α = .87), deprivation sensitivity (anxiety and frustration of being aware of information you do not know and want to know; α = .89), stress tolerance (dispositional tendency to handle the anxiety that arises when confronting new experiences; α = .88), overt social (open interest in other people; α = .88), covert social (interest in discovering details about other people in indirect and secretive ways; α = .89), and thrill-seeking (willingness to accept risks to acquire new experiences; α = .86).

Perth Emotion Regulation Competency Inventory (PERCI; Preece et al., 2018).

This measure assesses people's ability to regulate their own negative and positive emotions. Participants responded to a series of statements (e.g., "When I'm feeling bad, I have strong urges to do risky things") on a 7-point scale (from 1 = *strongly disagree* to 7 = *strongly agree*), with higher scores indicating more emotion regulation difficulties. We calculated eight subscale (e.g., negative inhibiting behaviour) and five composite (e.g., general emotion regulation) scores to assess domain specific as well as overall competencies. All subscales and compositive scores have good to excellent internal consistency (Study 1b: α = .84-.95).

The Acceptance and Actions Questionnaire-II (AAQ-II ; Bond et al., 2011).

This measure assesses psychological flexibility and trait experiential avoidance, which is defined as a tendency to avoid private experiences (e.g., bodily sensations, emotions, thoughts etc.) typically associated with anxiety. Participants responded to a series of statements (e.g., "I worry about not being able to control my worries and feelings") on a 7-point scale (1 = *never true* to 7 = *always true*). We summed items, with possible scores ranging from 7 to 49 (high scores indicate less flexibility; Study 1b: α = .94).

Results and Discussion

Uncovering Behaviour and Individual Characteristics.

We were interested in whether people with certain individual characteristics (Table S1.7) are likely to uncover sensitive-content screens. Therefore, we examined the relationship between the individual characteristic measures and uncovering behaviour across a series of Spearman's rho and point biserial correlations (Table S1.8). We predicted that curiosity, intolerance to uncertainty, experiential avoidance, and emotion regulation difficulties would be positively related to uncovering behaviour. Largely, our predictions were unsubstantiated: most correlations were small and/or not statistically significant (first image behaviour, r_{pb}s = -.08-.12; uncovered all, or not, r_{pb}s = -.09-.12; number of sensitive screens uncovered, r_ss = -

.09-.11). There are a few exceptions, though we acknowledge that the number of analyses may have inflated the likelihood of finding significant effects.

People higher in the desire to seek out new knowledge and information, and the subsequent joy of learning, and people higher in anxiety and frustration when aware of information that they do not know and want to know, were less likely to uncover all sensitive screens during the Instagram task. That is, the joyous exploration and deprivation sensitivity subscales of the 5DCR were *negatively* correlated with whether participants uncovered all sensitive screens, or not, during the Instagram task, $r_{pb}(262) = -.13$, $p = .04$, 95% CI [-.25, -.01], and $r_{pb}(262) = -.14$, $p = .02$, 95% CI [-.26, -.02], respectively. The deprivation sensitivity subscale was also negatively correlated with the total number of sensitive screens participants uncovered, $r_s(262) = -.17$, $p = .006$, 95% CI [-.29, -.05]. Moreover, people who have greater difficulties controlling behaviours relating to negative emotions—in terms of activating non-dominant behavioural response tendencies (e.g., "When I'm feeling bad, I can't get motivated to do important things")—uncovered more sensitive screens during the Instagram task. That is, the negative activating behaviour subscale of the PERCI was positively correlated with the total number of sensitive screens participants uncovered, $r_s(262) = .12$, $p = .049$, 95% CI [-.001, .24].

The negative relationship between the 5DCR subscales and uncovering behaviour is at odds with our predictions, but there are several explanations for the pattern. First, it is possible that the negative nature of the warning curtails information seeking behaviours that would otherwise arise in the face of such curiosity inducing stimuli. However, we know that people are especially willing to engage with stimuli if the consequences are uncertain and negative in nature ("Pandora effect"; Hsee & Ruan, 2016). Second—although the 5DCR is a trait measure of curiosity—it is possible that participants' state affect following the Instagram task influenced their responses. For example, people with higher trait joyous exploration and deprivation sensitivity may have uncovered more sensitive screens during the Instagram task

(as predicted) such that they resolved lingering state curiosity and subsequently reported lower 5DCR scores. As for the positive relationship between the PERCI subscale and uncovering behaviour, it is possible that the task made people feel bad, and to move on with their life, they had to resolve the "bad feeling" by uncovering sensitive screens. Notably, these correlations are all relatively weak (with low predictive value): therefore, there is little evidence to suggest that people with these individual characteristics are more likely to approach sensitive content. Indeed, people appear to behave similarly, irrespective of these individual characteristics.

Table S1.7

Study 1b: Means (and Standard Deviations) for Individual Characteristic Measures

Measure	Scale (Range)	*M* (*SD*)
IUS-12	Prospective anxiety (7-35)	21.9 (6.3)
	Inhibitory anxiety (5-25)	12.2 (5.1)
	Total (12-60)	34.1 (10.5)
5DCR	Joyous exploration (1-7)	5.1 (1.3)
	Deprivation sensitivity (1-7)	4.4 (1.5)
	Stress tolerance (1-7)	4.5 (1.5)
	Thrill-seeking (1-7)	2.9 (1.5)
	Overt social (1-7)	4.8 (1.4)
	Covert social (1-7)	4.3 (1.5)
AAQ-II	(7-49)	22.5 (10.4)
PERCI	Negative controlling experience (4-28)	12.9 (6.1)
	Negative inhibiting behaviour (4-28)	10.1 (5.9)
	Negative activating behaviour (4-28)	15.7 (7.4)
	Negative tolerating emotions (4-28)	14.2 (5.8)
	Positive controlling experience (4-28)	11.5 (5.9)
	Positive inhibiting behaviour (4-28)	7.4 (4.7)
	Positive activating behaviour (4-28)	7.9 (4.6)
	Positive tolerating emotions (4-28)	6.1 (4.2)
	Negative emotion regulation (16-112)	52.8 (20.3)
	Positive emotion regulation (16-112)	32.9 (15.6)
	General facilitating hedonic goals (20-140)	64.3 (24.4)
	Positive containing emotions (12-84)	21.5 (11.9)
	General emotion regulation (32-224)	85.8 (31.9)

Note. n = 264. IUS-12 = Intolerance to Uncertainty Scale Short Form; 5DCR = The Five-Dimensional Curiosity Scale Revised; AAQ-II = The Acceptance and Actions Questionnaire-II; PERCI = Perth Emotion Regulation Competency Inventory.

Table S1.8

Study 1b: Correlations Between Uncovering Behaviour and Individual Characteristic Measures

Measure	Scale	Type of Uncovering Behaviour		
		First image behaviour	Total screens uncovered	Uncovered all (or not)
		r_{pb} [95% CI]	r_s [95% CI]	r_{pb} [95% CI]
IUS-12	Prospective anxiety	.004 [-.12, .13]	.06 [-.06, .18]	.07 [-.05, .19]
	Inhibitory anxiety	.03 [-.09, .15]	.08 [-.04, .20]	.05 [-.07, .17]
	Total	.01 [-.11, .13]	.07 [-.05, .19]	.07 [-.05, .19]
5DCR	Joyous exploration	.001 [-.12, .12]	-.09 [-.21, .03]	-.13* [.01, .24]
	Deprivation sensitivity	.05 [-.07, .17]	-.17** [.05, .29]	-.14* [.02, .26]
	Stress tolerance	-.08 [-.20, .04]	-.08 [-.20, .04]	-.03 [-.15, .09]
	Thrill-seeking	.01 [-.11, .13]	-.02 [-.14, .10]	-.01 [-.13, .11]
	Overt social	-.06 [-.18, .06]	.03 [-.09, .15]	.03 [-.09, .15]
	Covert social	.06 [-.06, .18]	.02 [-.10, .14]	.01 [-.11, .13]
AAQ-II		.11 [-.01, .23]	.06 [-.06, .18]	.02 [-.10, .14]
PERCI	Negative controlling experience	.11 [-.01, .23]	.07 [-.05, .19]	.06 [-.06, .18]
	Negative inhibiting behaviour	.05 [-.07, .17]	.09 [-.03, .21]	.09 [-.03, .21]
	Negative activating behaviour	.12 [-.001, .24]	.12* [-.001, .24]	.06 [-.06, .18]
	Negative tolerating emotions	.01 [-.11, .13]	.04 [-.08, .16]	.12 [-.001, .24]
	Positive controlling experience	.02 [-.10, .14]	.02 [-.10, .14]	.05 [-.07, .17]
	Positive inhibiting behaviour	.03 [-.09, .15]	-.01 [-.13, .11]	-.08 [-.20, .04]
	Positive activating behaviour	.01 [-.11, .13]	-.01 [-.13, .11]	-.11 [-.23, .01]
	Positive tolerating emotions	.05 [-.07, .17]	.04 [-.08, .16]	-.04 [-.16, .08]
	Negative emotion regulation	.10 [-.02, .22]	.11 [-.01, .23]	.10 [-.02, .22]
	Positive emotion regulation	.03 [-.09, .15]	.03 [-.09, .15]	-.05 [-.17, .07]
	General facilitating hedonic goals	.08 [-.04, .20]	.11 [-.01, .23]	.10 [-.02, .22]
	Positive containing emotions	.03 [-.09, .15]	.01 [-.11, .13]	-.09 [-.21, .03]
	General emotion regulation	.08 [-.04, .20]	.01 [-.11, .13]	.04 [-.08, .16]

Note. n = 264. IUS-12 = Intolerance to Uncertainty Scale Short Form; 5DCR = The Five-Dimensional Curiosity Scale Revised; AAQ-II = The Acceptance and Actions Questionnaire-II; PERCI = Perth Emotion Regulation Competency Inventory. * p <.05, ** p <.01.

Reasons for Uncovering Sensitive-Content Screens and Individual Characteristics

As detailed in the main paper, we found four key reasons for uncovering behaviour—information seeking behaviour, positive and negative affect driven behaviour, and avoidance behaviour. To see if people with certain pre-existing individual characteristics endorse certain reasons for their uncovering behaviour, we ran a series of correlations between our individual characteristic measures and the three approach-based reason factors (Table S1.9). We found that most of the characteristics were related to information seeking behaviour (*r*s = .15-.26) and negative affect driven behaviour (*r*s = -.02-.61), and a few were related to positive affect driven behaviour (*r*s = .13-.27; Table S1.9). Again however, we acknowledge that the number of analyses here may have inflated the likelihood of finding significant effects.

Taken together, we found limited evidence of any meaningful relationships between these individual characteristics and uncovering behaviour *or* reasons for uncovering. Indeed, it is possible that meaningful relationships do exist with respect to individual characteristics, yet like vulnerabilities, the relationships with uncovering behaviour and people's reasons uncovering are more complex. For example, existing individual characteristics (e.g., a preference to down-regulate positive emotions) may interact with state mood factors (e.g., low mood) *and* contextual motivations (e.g., being alone in the middle of the night vs. with others during the day) to influence a person's uncovering decision. Therefore, even though overall behaviour (i.e., to uncover or keep covered) may be the same at surface-level, the possible interactions between these underlying factors may mean reasons for uncovering sensitive screens are unique to each person/situation and thus more *nuanced* than captured here. Future research should explore this possibility.

Table S1.9

Study 1b: Correlations Between Reason Factors and Individual Characteristic Measures

Scales	Reasons			
	Negative affect driven behaviour	Positive affect driven behaviour	Information seeking behaviour	Avoidance behaviour
Prospective anxiety	.05 [-.07, .17]	-.04 [-.16, .09]	.19** [.06, .30]	.12 [-.01, .24]
Inhibitory anxiety	.21** [.09, .32]	-.03 [-.15, .10]	.07 [-.05, .19]	.19** [.06, .30]
IUS Total	.13* [.01, .25]	-.04 [-.16, .09]	.15* [.02, .27]	.16* [.04, .28]
Joyous exploration	-.13* [.24, -.01]	.07 [-.05, .20]	.16* [.04, .28]	-.04 [-.16, .09]
Deprivation sensitivity	-.02 [-.15, .10]	.11 [-.01, .23]	.21** [.09, .33]	.01 [-.11, .14]
Stress tolerance	-.16** [-.28, -.04]	.01 [-.11, .14]	-.04 [-.16, .09]	-.18** [-.29, -.05]
Thrill-seeking	.19** [.07, .31]	.27** [.15, .38]	.23** [.11, .34]	-.21** [-.33, -.09]
Overt social	-.14* [-.26, -.01]	.13* [.01, .25]	.18** [.06, .30]	.06 [-.07, .18]
Covert social	.02 [-10, .14]	.11 [-.02, 23]	.26** [.14, .37]	-.10 [-.22, .02]
AAQ-II	.25** [.13, .36]	-.06 [-.18, .07]	.09 [-.04, .21]	.14* [.02, .26]
Neg controlling experience	.25** [.14, .37]	-.04 [-.17, .08]	.15* [.03, .27]	.06 [-.06, .19]
Neg inhibiting behaviour	.38** [.27, .48]	.07 [-.05, .20]	.19** [.07, .31]	-.03 [-.15, .10]
Neg activating behaviour	.18** [.06, .30]	-.04 [-.17, .08]	.19** [.07, .31]	.12 [-.01, .24]
Neg tolerating emotions	.06 [-.07, .18]	.04 [-.09, .16]	.16* [.04, .28]	.08 [-.05, .20]
Pos controlling experience	.28** [.16, .39]	-.001 [-.13, .12]	.08 [-.05, .20]	.12 [-.01, .24]

Scales	Reasons			
	Negative affect driven behaviour	Positive affect driven behaviour	Information seeking behaviour	Avoidance behaviour
Pos inhibiting behaviour	.46** [.35, .55]	.27** [.15, .38]	.22** [.10, .33]	-.09 [-.21, .04]
Pos activating behaviour	.39** [.28, .49]	.17** [.04, .29]	.15* [.02, .26]	-.03 [-.15, .10]
Pos tolerating emotions	.61** [.53, .68]	.18** [.06, .30]	.08 [-.04, .20]	-.04 [-.17, .08]
Neg emotion regulation	.27** [.15, .38]	.004 [-.12, .13]	.22** [.10, .33]	.08 [-.05, .20]
Pos emotion regulation	.52** [.42, .61]	.18** [.06, .30]	.16* [.04, .28]	-.003 [-.13, .12]
General facilitating hedonic goals	.29** [.17, .40]	.003 [-.12, .13]	.20** [.08, .32]	.09 [-.03, .21]
Pos containing emotions	.54** [.45, .63]	.23** [.11, .35]	.17** [.05, .29]	-.06 [-.18, .06]
General emotion regulation	.43** [.32, .52]	.09 [-.04, .21]	.22** [.09, .33]	.05 [-.08, .17]

Note. n = 250-252. AAQ-II = The Acceptance and Action Questionnaire. Neg = Negative. Pos = Positive. * p <.05, ** p <.01.

4 Investigating the Role of Information Seeking Behaviour in the Decision to Uncover Sensitive-Content Screens

Chapter 4 is published as:

Simister, E. T., Bridgland, V. M. E., Williamson, P., & Takarangi, M. K. T. (2023). Mind the Information-Gap: Instagram's Sensitive-Content Screens are more likely to deter people from viewing potentially distressing content when they provide information about the content. *Media Psychology, 26,* 660-679. https://doi.org/10.1080/15213269.2023.2211774

Authors Contributions: I developed the study design with the guidance of MKTT and VMEB. I collected the data, performed the data analysis and interpretation (with assistance from PW during the revision process), and drafted the manuscript. MKTT and VMEB contributed equally by making critical revisions to the manuscript. All authors approved the final version of the manuscript for submission.

Abstract

Instagram's sensitive-content screens seek to minimise engagement with negative content by blurring sensitive images and providing a warning. However, the very design of sensitive-content screens may elicit uncertainty/curiosity and prompt information-seeking behaviours: congruent with the information-gap hypothesis. To test this idea experimentally, we presented participants with screened negative images accompanied by a brief, detailed, or no content description, during a simulated Instagram task. Participants viewed screens one at a time and uncovered at their discretion. In line with our predictions, people uncovered screens irrespective of description type, but did so most often with no description. Most participants indicated that knowing *what* the sensitive content contained bolstered their ability to make an informed decision. These results have implications; information provided

alongside sensitive-content screens can influence engagement and therefore should be considered as part of sensitive-content guidelines.

Introduction

Instagram's sensitive-content screens blur images and provide a warning (e.g., about upcoming graphic or violent content). The idea is that such screens should *minimise* engagement with the content (Mosseri, 2019a, 2019b). Yet, our previous work suggests sensitive-content screens may promote (rather than minimise) engagement. Specifically, we found that many people deliberately *and* repeatedly exposed themselves to potentially distressing graphic material (by uncovering screens; Bridgland, Bellet et al., 2022; Simister, Bridgland & Takarangi, 2023; Studies 1a and 1b). An important question then, is *why* sensitive-content screens seem to promote engagement? In our previous work we asked participants about their reasons for engaging with screened negative content. Participants consistently and most strongly endorsed items related to (what we termed) *information-seeking* reasons (e.g., "…I wanted to know why it was covered"; Simister, Bridgland & Takarangi, 2023; Studies 1a and 1b). Thus, here our primary aim was to *experimentally* examine the role of information-seeking in participants' decisions to engage with screened negative content. We draw on broader psychological theories to better understand uncovering behaviour in the online context and provide the foundations for a theoretical framework—which is currently non-existent. Like in our previous work, we presented participants with screened negative images in a simulated Instagram task, but each screen was accompanied by either a brief, detailed, or no content description. We measured the frequency of uncovering according to content description. As a secondary aim, we examined whether people's intolerance to uncertainty influenced their uncovering behaviour under our different content description conditions. Finally, on an exploratory basis, we examined participants' views on how content descriptions influenced their uncovering decisions.

Sensitive-Content Screens may Foster Uncertainty and Curiosity

A hallmark feature of sensitive-content screens—and similar warning systems found on TikTok, Facebook, Twitter, Reddit and Buzzfeed, among others—is the obfuscation of images via an image processing technique called Gaussian Blur. The resulting image has reduced noise (e.g., variations in brightness or color) and detail, which makes it difficult for people to determine exactly what the image is and likely increases their uncertainty, and subsequently their curiosity about what the image might possibly be. Indeed, several lines of established theoretical work are consistent with the idea that uncertainty fosters curiosity (e.g., Berlyne, 1954; Day, 1982; Loewenstein, 1994). Though these theories use seemingly different terminology, the underlying concepts are similar. That is, curiosity arises in response to arousal (the level of which differs from person to person) which can be triggered by novelty, incongruity, complexity, and/or uncertainty. In one example, Campion et al. (2009) presented participants with stories that were either missing information or not; participants were consistently more curious about the stories that were missing information compared with those that were not. Therefore—although the *exact* amount of arousal required to foster curiosity is unknown—it is possible that the very design of sensitive-content screens elicits uncertainty and makes people feel curious. Of course, a descriptive warning that accompanies blurred images could hypothetically satiate such curiosity. However, Instagram's current warning (i.e., "Sensitive Content: This photo may contain graphic or violent content") provides no information about the nature of the blurred image, for example, how it may be "graphic" or "violent". Thus, it seems unlikely that the current warning would satiate such curiosity.

Curiosity Prompts Information-Seeking

We know that curiosity is associated with information-seeking, which, broadly speaking, is characterised by exploration and approach-driven behaviour (e.g., the move

towards unknown information; Day, 1982). According to the information-gap hypothesis (Loewenstein, 1994), when people perceive a gap in knowledge—that is, when what *they want to know* exceeds their current level of knowledge—they experience feelings of deprivation. These feelings are aversive and motivate people to obtain information to eliminate, or at least reduce, their perceived gap in knowledge (Loewenstein, 1994). Thus, the eventual resolution of curiosity is rewarding. Indeed, functional resonance imaging studies have shown that the relief of curiosity (through the provision of information) activates brain regions related to reward processing (Jempa et al., 2012). Though in some situations this relationship between curiosity and information-seeking lends itself to positive outcomes (e.g., in educational settings where curiosity predicts academic performance; Von Stumm et al., 2011), there are other situations in which information-seeking arising from curiosity may result in negative outcomes. For example, some people are morbidly curious, such that they deliberately expose themselves to information that may be distressing (e.g., images that portray death; Oosterwijk, 2017). Indeed, it seems that some people seek to resolve curiosity even if—and in some cases, because—the consequences are uncertain, but expected to be negative in nature (e.g., electric shocks; the "pandora effect"; Hsee & Ruan, 2016). Such negative content may offer stronger informational gain than positive or neutral information because of its unique (and sometimes, socially deviant) nature (Oosterwijk, 2017). Although people may come to regret such decisions—perhaps if/when they experience negative consequences—the desire to resolve curiosity (under uncertain conditions) is seemingly more important than regret aversion (Van Dijk & Zeelenberg, 2007).

In the case of sensitive-content screens, curiosity may pose a similar risk; people may seek out further information about the blurred content despite their readiness—or lack thereof—to see such negative material. That is, people likely engage with screened negative content—despite the potential for distress—to get more information about what is

beneath/why it is covered, and to ultimately reduce their feelings of deprivation. Our previous work—which investigated what people do when encountering screened images—aligns with these ideas. In this research, we asked Instagram users what they would do when encountering a screened image: ~80% of participants said they would uncover it (Bridgland, Bellet et al., 2022). The same proportion (~85%) uncovered a screened image at the first opportunity when interacting with a mock Instagram feed, and 52.7% of these participants uncovered every screened image (Simister, Bridgland & Takarangi, 2023; Study 1a).

Can we Reduce Information-Seeking Behaviour by Providing Content Descriptions with Sensitive-Content Screens?

Given previous theoretical and empirical work, we hypothesised that if we could reduce feelings of curiosity about the screened negative content, we could therefore also reduce information-seeking behaviour (in the form of uncovering). Accordingly, our primary research aim was to investigate if providing brief or detailed information about the screened negative images would reduce uncovering behaviour. Specifically, we predicted participants would uncover sensitive-content screens with an accompanying content description (brief or detailed) less frequently than sensitive-content screens without a content description (H1) because the additional information would help satiate their curiosity for the content without them needing to uncover it. However, we made no specific predictions regarding behaviour for brief vs. detailed content descriptions because it is unclear exactly *how much* information is necessary to reduce information-seeking behaviour.

Does Intolerance to Uncertainty Change Information-Seeking Behaviour?

Intolerance to uncertainty, which refers to negative beliefs about uncertainty and its implications (e.g., “uncertainty keeps me from living a full life”; Carleton, Mulvogue et al., 2012), may be an important individual difference to consider here. Although intolerance to uncertainty is a transdiagnostic characteristic associated with avoidance in some clinical

populations (e.g., panic disorder; Carleton et al., 2013), difficulties tolerating uncertainty occur along a continuum in the general population (e.g., from low to high intolerance; Carleton, Weeks et al., 2012). People with elevated intolerance to uncertainty find it difficult to cope in uncertain situations (Buhr & Dugas, 2002). Therefore, it is possible that people with elevated intolerance to uncertainty experience an even greater sense of curiosity in the face of sensitive-content screens (compared with people lower on intolerance) and may then be more driven to seek out information. We examined this possibility as a secondary aim, with the following hypothesis (H2): We predicted that intolerance to uncertainty would moderate the effect of condition. Specifically, we predicted that the expected difference in uncovering behaviour between content description conditions (as per H1) would increase as intolerance to uncertainty increased (i.e., the effect of condition on uncovering behaviour would be stronger at higher levels of intolerance to uncertainty).

Study 2

Method

The Flinders University Social and Behavioural Research Ethics Committee approved this study, and we pre-registered it on the Open Science Framework (https://osf.io/4wjq6). We used Qualtrics Software (2018) to conduct the study. We have reported all measures, conditions, and data exclusions. The supplementary materials are at the end of the chapter and the data, including a codebook describing all variables, can be found at: https://osf.io/ewa59/.

Participants

Our desired sample size was 199 participants, determined by a priori power analysis for a two-tailed, paired t-test (based on our planned contrasts; using G*Power; Faul et al., 2007) with an alpha of 0.05, power of .80, and effect size of $d = 0.2$ (the largest sample size we could achieve with the resources we had available for this study). We note that this

sample size was also adequate to reliably identify a small within-person predictor effect, with an alpha of 0.05, power of .80, and effect size of $d = 0.2$ (Murayama et al. 2022). We recruited participants—with previous experience of >1,000 tasks—from the United States using Amazon's Mechanical Turk (MTurk) through Cloud Research. To avoid bots/server farmers, we screened out participants who failed a captcha and/or scored less than 8/10 on an English proficiency test (Moeck et al., 2022). Because we only wanted to recruit Instagram users, we also screened out participants who did not select Instagram and/or selected "Konnect" (a bogus platform included to detect inattentive responses) when asked about social media use. Participants who were screened out were ineligible to continue with the survey. In total, 264 participants completed the survey. Of these participants, we excluded data from 62 (per our pre-registered plan): four participants reported that they did not read the content descriptions, 19 participants did not report the attention check word (i.e., giraffe), and 39 participants reported uncovering screens to "fulfil task requirements" (e.g., because they thought they *had* to uncover screens). Participants received a payment of $2.00 USD.

Our final sample of 202 participants, aged 20-76 years ($M = 36.81$, $SD = 10.51$) included 63.4% females ($n = 128$), 35.1% males ($n = 71$), 1.0% reported as non-binary ($n = 2$), and one participant preferred not to report their gender. Our sample was predominantly European American/White (71.7%; $n = 145$); other participants were of African American/Black (8.4%; $n = 17$), Asian (5.4%; $n = 11$), Latinx (5.0%; $n = 10$), Native American (1.0%; $n = 2$), or other (5.9%; $n = 12$; e.g., mixed race) descent; four (2.0%) participants specified nationality (e.g., American), and one participant reported their gender. Most participants (48.5%; $n = 98$) reported an income between $45,000-$140,000 and were predominantly (52.5%; $n = 106$) college graduates (see Table S2.1 for details). Aside from all being Instagram users, most participants reported using Facebook (89.1%; $n = 180$), YouTube (88.6%; $n = 179$), Reddit (71.3%; $n = 144$), Twitter (64.9%; $n = 131$), and TikTok

(56.4%; *n* = 114); some users also reported using Snapchat (39.6%; *n* = 80), Pinterest (35.6%; *n* = 72), WhatsApp (21.3%; *n* = 43), and Tumblr (16.3%; *n* = 33).

In terms of participants' Instagram use, most participants (53.5%; *n* = 108) reported they had used Instagram every day over the past week (followed by 5 days = 17.3%; *n* = 35, 4 days = 8.9%; *n* = 18, 3 days = 7.9%; *n* = 16, 2 days = 7.4%; *n* = 15, 1 day or less = 3.0%; *n* = 6, and 6 days = 2.0%; *n* = 4). Most participants (79.2%; *n* = 160) reported they had used Instagram for one hour or more on an average day in the last 30 days (< half an hour = 20.8%; *n* = 42; 1 hour = 40.1%; *n* = 81, 2-3 hours = 24.8%; *n* = 50, 4-5 hours = 6.9%; *n* = 14, > 6 hours = 7.4%; *n* = 15). Additionally, most participants (68.8%; *n* = 139) reported they have seen sensitive-content screens covering content on their own Instagram. Overall, there was a preference for limiting exposure to sensitive content: 48.5% (*n* = 98) of participants reported they had/would select Instagram's default "limit" sensitive content option—which allows some, but not all, sensitive content to appear on Instagram—and 26.2% (*n* = 53) of participants reported they wanted to see even less sensitive content such that they had/would select the "limit even more" option. A small subset of our sample (25.5%; *n* = 51) reported they wanted to see more sensitive content (that is, more than the default setting typically shows), such that they had/would select the "allow" sensitive content option. Notably, only 12.9% (*n* = 26) of participants had used this feature to date, meaning 87.1% (*n* = 176) of participants were yet to "control" the amount of sensitive content on their Instagram, even though some participants' preferences did not align with the default setting.

Materials and Procedure

We told participants we were collecting information about social media engagement. Following informed consent procedures, we asked participants to indicate how many days of the last 7 days, and for how many hours on average each day, they used Instagram (over the last 30 days; Bridgland, Bellet et al., 2022). We also asked participants how often (not at all,

sometimes, often, very often) they viewed different types of images (e.g., portraits, animals) on Instagram, to reduce suspicion about the true nature of our study.

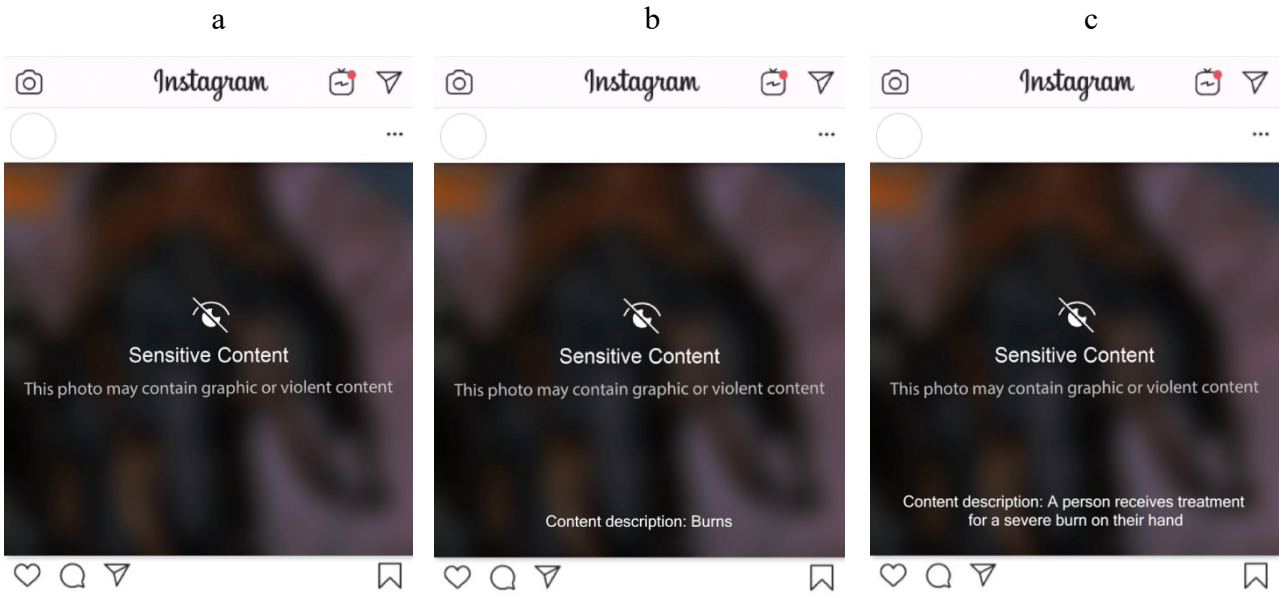
Participants then completed our *simulated Instagram task*. This task included the 30 most negative, positive, and neutral images (90 total) from the Nencki Affective Picture System (NAPS; Marchewka et al., 2014; based on normative ratings: 1 = *negative* to 9 = *positive*). We included positive and neutral images to improve the ecological validity of our design, because participants are likely to come across negative images *among* positive and neutral images on their own Instagram feed. The content of the images (e.g., people, animals, objects) is commonly found on Instagram and the negative content would likely meet the threshold for Instagram to screen them (e.g., some of the negative images include people/animals that have been injured/are deceased). All images appeared in an Instagram border with non-functional like and comment buttons. Consistent with Instagram's formatting, negative images had a warning ("Sensitive Content: This photo may contain graphic or violent content") obfuscating the image.

We manipulated content description condition within subjects: 10 of the screens had no content description (Figure 2.1a), 10 of the screens had a brief content description (e.g., "Burns"; Figure 2.1b), and 10 of the screens had a detailed content description (e.g., "A person receives treatment for a severe burn on their hand"; Figure 2.1c, see Appendix B for all content descriptions). We developed content descriptions from pilot study data: MTurk participants (N = 55) viewed the negative images and described them in one sentence. We then modified the descriptions to match them on style and word length across images (brief: range = 1-3, M = 2.5, SD = 0.7; detailed: range = 11-15, M = 12.8, SD = 1.5). We created three sets of 10 negative images matched on valence and arousal ratings (set 1: valence M = 1.8, SD = 0.2, arousal M = 7.3, SD = 0.5; set 2: valence M = 1.9, SD = 0.2, arousal M = 7.4, SD = 0.2; set 3: valence M = 1.9, SD = 0.2, arousal M = 7.3, SD = 0.4). We ran a series of

one-way ANOVAs that revealed no significant differences between sets in valence, $F(2, 29) = 0.29$, $p = .752$, $\eta^2 = .021$, or arousal, $F(2, 29) = 0.30$, $p = .745$, $\eta^2 = .022$. We counterbalanced sets across participants, meaning each set of 10 screens appeared equally often with no content description, brief content descriptions, or detailed content descriptions. Additionally, to determine whether people read the content descriptions, we included attention check descriptions (i.e., "If you are reading this, remember the word giraffe, you will report it later") on three additional negative images (that were not part of the other image sets). These images were all negatively valenced ($M = 2.5$, $SD = 0.02$), albeit to a lesser extent than the main negative images.

Figure 2.1

Example NAPS Image Modified to Look Like Instagram Images with a Sensitive-Content Screen Overlay and (a) No Content Description, (b) Brief Content Description, and (c) Brief Content Description



Participants viewed images/screens one at a time in a randomised order. After 3s the response options appeared;[14] for neutral and positive images, participants selected *Next Photo* to move onto the next image, and for negative images participants had the option to uncover the screen (select *See Photo*) and view the negative image underneath, or leave it covered (select *Next Photo*) and move to the next image. We also allowed participants to uncover the attention check screens so that coming across them during the task was consistent with the overall task experience, but these responses were not included in our uncovering variable.

At the completion of the simulated Instagram task, participants completed the intolerance to uncertainty measure (IUS-12; Carleton et al., 2007). Participants rated to what extent they agree with seven items assessing prospective anxiety (e.g., "*It frustrates me not having all the information I need*") and five items assessing inhibitory anxiety (e.g., "*When it's time to act, uncertainty paralyses me*"; 1 = *not at all characteristic of me* to 5 = *entirely characteristic of me*). We summed all items to create a total IUS-12 score (12-60; current study: α = .92).[15] Participants also indicated if they have seen sensitive-content screens before (yes/no), and their preferences for sensitive content on their own Instagram (i.e., to *allow*, *limit*, or *limit even more* sensitive content). We also asked participants if they read the content descriptions when they were presented (yes/no), and if they indicated they did, to explain whether (yes/no, and how) content descriptions influenced their decision to uncover screens or not (participants who did not read content descriptions were excluded). At this time, we also asked participants to report the word we asked them to remember during the task (i.e., "*giraffe*").

[14] We had a 3s delay between the presentation of images/screens and when response options (e.g., "See Photo/Next Photo") appeared (like Simister, Bridgland & Takarangi, 2023; Study 1b), so we had some control over how quickly participants moved through images.

[15] We also summed items into subscales (prospective anxiety: α = 87; inhibitory anxiety: α = .90), but because we made no predictions about them, and because the analyses for these subscales did not show any patterns of results that diverge from the findings for the full scale, all analyses relating to these subscales are in the supplementary materials.

Then, to detect participants who uncovered screens because of reasons associated with demand effects (i.e., because they thought they had to, rather than because they wanted to), we asked participants to respond to a series of true/false statements (e.g., "I thought I was supposed to uncover the screens"). To detect poor response quality, we also asked participants if they stopped the task for any extensive period (and when/for how long), or if they experienced any technical issues. Finally, participants completed demographics. We then fully debriefed participants.

Results

Main Analyses

Uncovering Behaviour.

Overall, participants uncovered 28.6% of the total screens in the Instagram task (M = 8.6, SD = 9.0). Only 4.5% (n = 9) of participants uncovered every screen, but most participants (90.1%; n = 182) uncovered *at least* one or more screens. That is, only 9.9% (n = 20) of participants left all screens covered. Taken together, it seems people do engage with screened sensitive content.

Content Descriptions.

We next addressed our primary aim and H1: to determine whether the number of sensitive-content screens participants uncovered differed depending on content description condition. We pre-registered a one-way repeated measures ANOVA, but because the pre-requisites for this analysis were not met (i.e., the data was not normally distributed), we ran a negative binominal linear mixed model instead, following the advice of an anonymous reviewer. We included random intercepts and slopes in the model. The fixed effects for the model confirmed a significant effect of content description condition on the number of screens uncovered, $F(2, 603) = 6.95$, $p = .001$.

To compare between the conditions, we ran an initial model using dummy variables with the no content description condition as the reference condition. See Table S2.1 for the model coefficients and their associated inferential statistics. The no content description condition (Estimate ± SE: 2.50 ± 1.26) was significantly different from both the brief (Estimate ± SE: 1.77 ± 0.90) and the detailed (Estimate ± SE: 1.59 ± 0.80) content description conditions: participants uncovered screens significantly *more* often when there was no content description, compared to when the screens appeared with either brief or detailed content descriptions. Therefore, the mere presence of content descriptions—1 to 15 words in length—reduced uncovering behaviour.

We next examined whether there was a difference in the number of sensitive-content screens participants uncovered according to the level of information in the content descriptions. We re-ran the binominal linear mixed model with detailed content description as the reference group to compare uncovering behaviour between brief and detailed conditions (Table S2.1). The brief content description condition was not significantly different from the detailed content description condition: participants uncovered a *similar* number of screens, irrespective of whether the screens appeared with brief or detailed content descriptions. Therefore, the level of information in the content descriptions did not appear to matter—neither brief nor detailed was more optimal than the other in reducing uncovering behaviour.

Intolerance to Uncertainty.

We next addressed our secondary aim and H2: to determine whether intolerance to uncertainty moderated the effect of content description condition on the number of screens uncovered. Here, we re-ran the negative binominal linear mixed model but included mean centered intolerance to uncertainty (Aiken & West, 1993) and the interaction term between content description and intolerance to uncertainty as fixed effects. As per the original model, there was a significant effect of content description; but, the effect of intolerance to

uncertainty on the number of screens uncovered was not significant, $F(1, 600) = 0.23$, $p = .634$, nor was the interaction between content description and intolerance to uncertainty, $F(2, 600) = 1.34$, $p = .263$ (see Table S2.1 for the coefficients). Therefore, contrary to our predictions, intolerance to uncertainty did not moderate the relationship between the level of information provided and uncovering behaviour.

Pre-Registered Exploratory Analyses

The Influence of Content Descriptions.

We were also interested in understanding participants' views on whether content descriptions influenced their decisions to uncover screened images, irrespective of their actual uncovering behaviour. Recall, we first asked participants to respond to a yes/no question regarding the influence of content descriptions: 89.1% ($n = 180$) of participants indicated that content descriptions influenced their decision to uncover screened images, whereas 10.9% ($n = 22$) of participants indicated they did not. We then asked participants who responded "yes" to explain *how* content descriptions influenced their decision, and participants who responded "no", *why* they did not[16], using open-text responses. One coder (the first author) analysed these responses using NVivo (Bazeley & Jackson, 2013) and through an iterative process, developed themes. A second coder—blind to the original coding—coded responses into the already identified themes. Agreement between coders was good (73.8%); the discrepancies were resolved via discussion.[17]

The overwhelming theme that emerged from the data was that content descriptions helped people make an *informed decision*, specifically about whether they should engage with screened sensitive content (79.2%; $n = 160$). Notably, *within* this theme there were several key subthemes. Participants reported that content descriptions helped them avoid

[16] The sample for this subset of data was small ($n = 22$), precluding us from drawing strong conclusions; nonetheless, we report themes in the supplementary material (Supplementary Table S2.4) for completeness.
[17] Four responses were unclear and were not coded into a theme, but the % reported here still refers to the total sample. See supplementary materials (Supplementary Table S2.5) for a full breakdown of the total sample.

certain types of content (e.g., animal/child abuse; "If it described animal abuse, I did not want to see the image"; 34.7%; *n* = 70), and more specifically, avoid content they thought might be disturbing/distressing (e.g., "I had no desire to view an image if I knew the content would be disturbing to me"; 9.9%; *n* = 20), or, that they believed they would not cope with (e.g., "If it were something I did not think I could handle I would avoid clicking on it"; 9.9%; *n* = 20). Another key theme that emerged from the data was the idea that content descriptions influenced people's levels of curiosity. Some people said content descriptions satiated their curiosity, such that they felt less inclined to uncover screened content (e.g., "Knowing what was within the pictures took care of the curiosity I felt"; 3.5%; *n* = 7), whereas others—albeit a minority—said the descriptions *increased* their curiosity, such that they felt more inclined to uncover screened content (e.g., "I sometimes was curious about how it would look"; 1.0%; *n* = 2). The final key theme that emerged from the data was that—unlike the other themes—a decision had already been made, but that the content descriptions *affirmed* people's decision to avoid screened content ("They gave me confirmation that I was making the right decision to not view the image" 3.5%; *n* = 7). For these participants, content descriptions were less influential on their subsequent behaviour, yet still appeared useful.

Discussion

Content Descriptions Reduce Information-Seeking Behaviour

Overall, our findings demonstrate that people engage with screened sensitive content irrespective of description type (albeit substantially less than in our previous work; Simister, Bridgland & Takarangi, 2023; Study 1a), but—in line with our primary predictions—information-seeking seems to play a role in engagement. We found people uncovered screens most often when they had the least amount of information available to them, that is, when we presented screens as they typically appear on Instagram—with a (non-specific) "Sensitive Content" warning, but without a content description. Importantly, we found people uncovered

screens less often when we presented them with information about the nature of the content, in the form of content descriptions, alongside Instagram's typical warning. Furthermore, *both* brief and detailed content-related information (1-15 words in length) minimised uncovering behaviour. Indeed, it is possible that upon seeing brief content descriptions people generated their own ideas about what might be beneath the screens, such that the detailed counterparts offered little additional information. Similarly, our brief content descriptions captured the most negative aspect of the proceeding image (e.g., "Burns"), so it is possible that the "additional information" offered by detailed content descriptions (e.g., "A person receives treatment for a severe burn on their hand") was only marginally more informative.

Intolerance to Uncertainty Does Not Appear to Change Information-Seeking Behaviour

Against our secondary predictions, the pattern of results remained the same irrespective of participants' ability to tolerate uncertainty. That is, people with varying levels of intolerance to uncertainty uncovered a similar number of sensitive-content screens with and without content descriptions. With hindsight, we see that it is possible that the construct(s) captured by the IUS-12 (i.e., the inability to cope with ambiguity in everyday life context) may not relate to behaviour, or changes in the availability of information, in this specific *online* context—which could explain why we did not observe the predicted effect. Alternatively, it is possible that changes in *state* arousal (e.g., increases in uncertainty)—possibly resulting from changes in the availability of information, or the simulated Instagram task itself—are just more influential on imminent uncovering behaviour than a trait characteristic.

Content Descriptions Help Participants Make Informed Decisions

Our exploratory qualitative analyses also revealed several key themes for *how* content descriptions influenced participants' uncovering decision. Most participants reported that content descriptions helped them make an informed decision in respect to whether they

should engage with screened sensitive content; other participants said content descriptions made them feel more or less curious, or affirmed their decision to avoid content. Taken together, the presence of content descriptions appears to minimise engagement with negative content, but this shift in behaviour may arise *because* people have more information and are therefore better positioned to make informed decisions. Although clinically, the effect of such descriptions is likely to be small for a single occasion of exposure to negative content, such reductions in exposure to negative images may accumulate and have larger emotional consequences with repetition, by affecting large numbers of people, and by having cascading effects on users' other social media behaviour (e.g., reducing self-triggering; Anvari et al., 2022; Funder & Ozer, 2019).

Methodological Implications

Hypothetically then, it may be appropriate to include content descriptions (either brief or detailed) preceding negative content—whether creating these descriptions is the responsibility of the user posting the content deemed "sensitive", or an additional responsibility for the algorithm/Instagram's moderators who screen the content. However, one potential limitation of the present research is that—due to the within-subjects design—we could not measure levels of anxiety or negative affect caused by each condition for example, before and after completing the Instagram task. Therefore, we currently do not know if providing content-related information—especially written information that is negative by its very nature—causes people to experience similar or more distress (than without the information) even if they decide *not* to view the content. Put another way, perhaps content descriptions merely shift one issue (i.e., whereby people feel distressed viewing negative content) to another (e.g., whereby people feel distressed reading *about* the negative content).

Another more troubling possibility is that the content descriptions may *enhance* how negative a person feels if they do decide to uncover the screen and view the subsequent

image. That is, people's distress may be higher than it would have otherwise been (i.e., when viewing negative content without preceding content descriptions). However, we know that while viewing more detailed trigger warning messages about the possible content of images (e.g., "torture, maltreatment, and death") induces anticipatory anxiety and negative affect, it does not seem to enhance how negative participants rate subsequent images (Bridgland et al., 2019). Indeed, a recent meta-analysis demonstrates that trigger warnings have a largely trivial effect on emotional responses towards warned of content (Bridgland et al., 2023). Recent data from our lab also shows that sensitive-content screens—in the format they appear on Instagram—increase anxiety and negative affect (Takarangi et al., 2023), even without giving people sufficient information about the preceding content so they can avoid it if necessary.

Interestingly, participants in the current study appeared to appreciate the additional description information, reporting that knowing *what* the sensitive content was helped them avoid certain content that they anticipated would be distressing/too difficult to cope with. Therefore, content descriptions may have benefits related to regulating current and anticipated affective states (e.g., by avoiding potential distress; a separate but related reason for behaviour identified in our previous work; Simister, Bridgland & Takarangi, 2023; Studies 1a and 1b), but also increases in autonomy. However, we know that in some cases people are seemingly poor judges of what is "good" for them (e.g., avoiding anxiety-provoking situations can increase anxiety; Barlow, 2021). Future research should address this limitation and assess anxiety and negative affect following exposure to screens (and the subsequent content) with and without brief and detailed content descriptions. It may be that brief(er) content descriptions still increase people's ability to make an informed decision, and minimise engagement with negative content, all while avoiding considerable increases in anxiety and negative affect. This idea parallels with a known conundrum within the broader literature on informed consent: balancing concerns over non-maleficence (i.e., not providing

too much information that may induce nocebo effects) and the right to autonomy (i.e., providing enough information to make an informed decision; Stirling et al., 2022).

Theoretical Contributions

Our findings have theoretical contributions. Currently, there is no existing theoretical framework for understanding why people behave the way they do online in relation to screened negative content; but here, we have done some of the foundational work towards developing such a framework. We now know that (one reason) people engage with negative content is because sensitive-content screens seemingly encourage—or, at a minimum, do not discourage—information-seeking behaviour (in the form of uncovering). Here, we theorise that sensitive-content screens elicit uncertainty due to their ambiguous nature, and subsequently make people feel curious. This argument is not purely speculation: we presented a separate group of pilot participants ($N = 50$) screened negative images with and without content description (using stimuli and descriptions from the present study), and asked them to rate their curiosity.[18] In line with our theory, participants were *most* curious about images without a content description ($M = 2.19$, $SD = 1.13$), and significantly more so than images with a brief ($M = 1.65$, $SD = 0.92$; $t(45) = 3.67$, $p < .001$, $d = 0.54$) or detailed ($M = 1.70$, $SD = 0.93$; $t(44) = 2.68$, $p = .010$, $d = 0.40$) content description. Furthermore, we theorise that to satiate curiosity (and reduce feelings of deprivation that are likely to arise; Loewenstein, 1994), people seek out information by uncovering screened negative content. In support of our theory, we observed this behaviour in the present study per the pattern of curiosity ratings above. Indeed, in this case, it seems that brief and detailed content-related

[18] We asked one group of participants ($n = 30$) "How curious are you about the content of this image" on a scale of 1 = *very slightly or not at all* to 5 = *extremely.* We asked another group of participants ($n = 20$) "How curious are you to know what the image beneath this screen depicts?" using the same scale. Given the similarity of these questions, to maximise statistical power we collapsed the data across groups, though we note six participants missed ratings for at least one of the conditions (there were 10 images per condition)—meaning usable data for our paired t-test was reduced (to $n = 45$ and $n = 46$). Nonetheless, this sample size, per a sensitivity analysis for a two-tailed, paired t-test (using G*Power; Faul et al., 2007) was still adequate to reliable identify a medium-sized effect ($d = 0.42$), with an alpha of 0.05 and power of .80.

information (1-15 words in length) can satiate curiosity, even though the image itself may provide more information.

Notably, endorsement of the idea that content descriptions satiated participants' curiosity was lower than we had anticipated (based on participants' endorsement of curiosity-related reasons in our previous work; Simister, Bridgland & Takarangi, 2023; Studies 1a and 1b). However, it is likely that, despite emerging as separate themes here, participants' responses related to making informed decisions were closely associated with changes in curiosity. Indeed, most participants provided insight into how they *used* the additional information to make behavioural decisions, moving beyond merely explaining how the presence of descriptions changed their level of curiosity.

Importantly, our contribution here is only the beginning of the work required to develop a comprehensive theoretical framework for understanding why people behave the way they do online in relation to screened negative content. Uncovering behaviour may be influenced by complex interactions between momentary state factors (e.g., mood), existing trait vulnerabilities (e.g., depression, or heightened psychological reactance; see also Ringold, 2002 for related work on the "forbidden fruit" effect) *and* situational demands (e.g., the online presence of others). For example, previous work has shown that people with depression often seek to maintain negative mood states (e.g., by listening to sad music; Millgram et al., 2015); therefore seeking out screened negative content may serve as another means by which people with depression regulate their affect. Additionally, given the social nature of online platforms, it is possible that this seemingly maladaptive behaviour (of seeking out screened content to maintain negative mood states) may be validated and/or encouraged by other users in vulnerable online communities (e.g., non-suicidal self-injury communities; Fulcher et al., 2020). Therefore, future research should examine some of these factors to expand on our foundational work.

Implications for Social Media Platforms

Our findings also have practical implications for the use of sensitive-content screens in their current format. Instagram has previously argued that sensitive-content screens help protect users by minimising engagement with negative content, but here we have demonstrated that screens seemingly do the opposite. One solution to counteract uncovering behaviour—and which aligns with Instagram's recent aims to give people more choice over what they see—is to provide content-related information alongside the "Sensitive Content" warning. Here we have demonstrated that including such information not only shifts behaviour (by minimising engagement with negative content), but also—and perhaps more importantly—bolsters people's ability to make informed decisions with respect to which content they want to engage with (or not). Though we acknowledge that irrespective of providing such information, some people will still decide to engage with negative content, arguably, they will be more informed when they do so—which is a benefit that advocates of warnings (and Instagram) seem to hold in high regard. Finally, although we focused specifically on Instagram, our findings have similar implications for other social media platforms that use similar warning initiatives (e.g., TikTok, Facebook, Twitter, Reddit and Buzzfeed).

Limitations

Our study has other limitations. First, despite our efforts to exclude participants who appeared not to follow task instructions, we cannot know for sure that all participants read *every* content description. However, the pattern of our results is consistent with the idea that the manipulation affected how participants responded (it also followed the expected pattern according to theory), and participants provided insightful responses with respect to the influence of such descriptions, suggesting our manipulation was effective. Second, because a third of the total number of images were covered by a sensitive-content screen, and we

manipulated content descriptions within-person, it may have been obvious to participants what we were interested in. Thus, the external validity of the experiment may be limited. Relatedly, it is possible there may have been carryover effects: participants' behaviour on screens without content descriptions may have been influenced, at least partly, by the presence of content descriptions on other screens. Indeed, this issue seems likely given uncovering behaviour was lower within our no content description condition (participants uncovered 35.2% of these screens) compared with our previous work, where screens were also presented without content descriptions (Simister, Bridgland & Takarangi, 2023; Study 1a: 70.7%; Study 1b: 43.8%; though we note subtle differences in methodology across these studies). However, if anything, this limitation suggests that the difference between the content description conditions in the present study may have been larger *without* carryover effects, such that our results may be a conservative estimate of the true effect of content descriptions. To test this possibility, future research could compare descriptions between-subjects. Third, it is possible some participants were just more or less interested in viewing certain content, such that it was not the descriptions per se that changed their behaviour, but rather their level of interest—a separate but closely related concept to curiosity (see Litman, 2005). If this were the case then we might refine our theories with respect to the mechanisms at play (i.e., uncertainty may prompt interest along with curiosity), but the main conclusions would remain the same. Fourth, because participants completed intolerance to uncertainty measures after the simulated Instagram task, it is possible that exposure to images in the task/the task itself influenced their responses to this measure. However, intolerance to uncertainty is, by definition, a dispositional characteristic (Buhr & Dugas, 2009) and the items we used to measure it reflect a trait rather than situational variable (e.g., "A small, unforeseen event can spoil everything, even with the best of planning"). Furthermore, giving participants the opportunity to reflect on their beliefs about uncertainty first may have

influenced their affective state (e.g., by making participants more apprehensive about upcoming uncertainty) and changed their subsequent uncovering behaviour. Therefore, on balance, we preferred to risk the possibility of the experimental stimuli influencing our intolerance to uncertainty data rather than jeopardise our ability to address our primary research aim. Nonetheless, future research could measure intolerance to uncertainty prior to the Instagram task (or counterbalance the order of measures) to address this limitation.

Additionally, because sensitive-content screens were originally designed to protect vulnerable users, future research could explore how content descriptions—like the ones we used here, but also content descriptions for personally relevant content—influence behaviour (and affect) within certain populations (e.g., people with post-traumatic stress disorder). Finally, because we only included content descriptions up to 15 words in length, we do not know if increasing the number of words would further minimise information-seeking behaviour or whether there would be a marginal or boundary effect as the number of words increase (considering the additional information may only be marginally more informative). Indeed, it is possible that more words may increase people's anxiety (Day, 1982), but offer little additional benefit in terms of minimising uncovering behaviour.

Conclusions

Taken together, these data provide preliminary evidence that sensitive-content screens in their current format promote (rather than minimise) engagement with negative content by prompting information-seeking behaviours. Content descriptions provided alongside Instagram's typical warning seemingly minimise un-informed engagement by facilitating informed decision making. Therefore, content descriptions should be considered as part of sensitive content guidelines on Instagram—and more generally speaking, across other social media platforms that use similar initiatives.

Supplementary Materials

Table S2.1

Demographic Characteristics for Study 2

Variable	% (*n*)
Household income	
<$20,000	16.3% (33)
$20,000 - $45,000	29.2% (59)
$45,000 - $140,000	48.5% (98)
$140,000 - $150,000	2.0% (4)
$150,000 - $200,000	2.5% (5)
>$200,000	1.5% (3)
Education	
Less than high school graduate	0.0% (0)
High school graduate	13.4% (27)
Some college	34.2% (69)
College Graduate	52.5% (106)

Intolerance to Uncertainty at the Subscale Level

As per our pre-registration plan, we re-run our main analyses for intolerance to uncertainty at the subscale level (though we note we pre-registered hierarchical regressions but report negative binominal linear mixed models here for consistency with the main analyses). We included mean centered anxiety (prospective and inhibitory in their respective models; Aiken & West, 1993) and the interaction term between content description and (prospective or inhibitory) anxiety as fixed effects. To compare between the conditions, we ran an initial model using dummy variables with the no content description condition as the reference condition. See Tables S2.2 and S2.3 for the model coefficients and their associated inferential statistics. The effects of prospective anxiety, $F(1, 600) = 1.21$, $p = .273$, and inhibitory anxiety, $F(1, 600) = 0.08$, $p = .777$, on the number of screens uncovered were not

significant; nor were the interactions between content description and prospective anxiety, $F(2, 600) = 1.24$, $p = .289$, and inhibitory anxiety, $F(2, 600) = 1.04$, $p = .355$. Therefore, neither prospective nor inhibitory anxiety moderated the relationship between the level of information provided and uncovering behaviour.

Table S2.2

Coefficient Estimates from a Series of Negative Binomial Linear Mixed Models Testing the Effect of Content Description, Prospective Anxiety and the Interaction Between Content Description and Prospective Anxiety on Uncovering Behaviour

Predictor	*b*	*SE*	*t*	*p*
(Intercept)	0.92	0.51	1.81	.071
Brief (vs. No Description)	-0.35	0.13	-2.74	.007
Detailed (vs. No Description)	-0.46	0.13	-3.60	< .001
Brief (vs. Detailed) *	0.12	0.13	0.87	.385
IUS_P	-0.01	0.02	-0.27	.791
Brief vs. No Description x IUS_P	0.03	0.01	1.33	.185
Detailed vs. No Description x IUS_P	0.03	0.02	1.39	.166
Brief vs. Detailed x IUS_P *	-0.002	0.02	-0.09	.931

*Note. b*s are unstandardised regression coefficients. IUS_P = centred prospective anxiety. * Indicates coefficients from a second model run using dummy variables with detailed description as the reference group

Table S2.3

Coefficient Estimates from a Series of Negative Binomial Linear Mixed Models Testing the Effect of Content Description, Inhibitory Anxiety and the Interaction Between Content Description and Inhibitory Anxiety on Uncovering Behaviour

Predictor	*b*	*SE*	*t*	*p*
(Intercept)	0.91	0.51	1.80	.072
Brief (vs. No Description)	-0.34	0.13	-2.70	.008
Detailed (vs. No Description)	-0.45	0.13	-3.51	< .001
Brief (vs. Detailed) *	0.11	0.13	0.87	.387
IUS_I	-0.03	0.02	-1.20	.232
Brief vs. No Description x IUS_I	0.03	0.03	1.13	.260
Detailed vs. No Description x IUS_I	0.03	0.03	1.32	.186
Brief vs. Detailed x IUS_I *	-0.01	0.03	-0.20	.839

*Note. b*s are unstandardised regression coefficients. IUS_I = centred inhibitory anxiety.
* Indicates coefficients from a second model run using dummy variables with detailed description as the reference group

The Influence of Content Descriptions

Why Content Descriptions Did Not Influence Participants' Decisions

The sample for this subset of data was small (n = 22), precluding us from drawing strong conclusions; nonetheless, we report themes here for completeness (Table 2.5). There were two seemingly opposing themes that emerged from the data here: people reported that content descriptions did not influence their decisions because, *irrespective* of the descriptions, they either wanted to view negative images (because of curiosity or otherwise; e.g., "*I was curious about all of them anyway*"; 5.0%; n = 10), or they did not want to view images (e.g., "*If there was a screen, I assumed I did not want to see the image*"; 5.9%; n = 12). Interestingly, upon closer examination of the qualitative data, we found that the nine

participants who uncovered every screen in the present study all indicated that the content descriptions did not influence their decisions, and that they wanted to view negative images irrespective of the descriptions. Although is it possible these participants selectively reported reasons to rationalise their behaviour and avoid cognitive dissonance (Festinger, 1957), it seems plausible that there is something unique about this small group of participants. Indeed, although we cannot reliably determine statistical significance due to a large discrepancy in sample size, the pattern of means suggests that people in this group were less tolerant of uncertainty (M = 42.4, SD = 10.5), compared with the rest of the sample (M = 35.2, SD = 9.9). Future research could focus recruitment *within* this subsample to examine the uniqueness of this population more closely.

Table S2.4

How Content Descriptions Influenced Participants Decisions

Theme	% (n)
Helped make informed decision	79.2% (160)
(General) Descriptions made me not want to look	8.9% (18)
(General) Had insight into what was coming	15.8% (32)
Able to avoid certain content	34.7% (70)
Able to avoid potentially disturbing or distressing content	9.9% (20)
Able to make decision based on perceived ability to cope with the content	9.9% (20)
Curiosity	4.5% (9)
Description increased curiosity	1.0% (2)
Description satisfied curiosity	3.5% (7)
Affirmed decision to avoid content	3.5% (7)
Unclear responses	2.0% (4)

Table S2.5

Why Content Descriptions Did Not Influence Participants' Decisions

Theme	% (*n*)
Did not want to view negative images (irrespective of description)	5.9% (12)
Wanted to view negative images (irrespective of description)	3.0% (6)
Wanted to view negative images (curious)	2.0% (4)

5 Investigating Whether Adding Content-Related Information to Sensitive-Content Screens Creates an Emotional Cost

Chapter 5 is submitted for publication:

Simister, E. T., Bridgland, V. M. E., & Takarangi, M. K. T. (2024). Adding brief content-related information to sensitive-content screens does not exacerbate screen- or image-related distress.

Authors Contributions: I developed the study design with the guidance of MKTT and VMEB. I collected the data, performed the data analysis and interpretation, and drafted the manuscript. MKTT and VMEB contributed equally by making critical revisions to the manuscript. All authors approved the final version of the manuscript for submission.

Abstract

Content descriptions presented on sensitive-content screens reduce how often people view negative images. But does this reduction in exposure come at an emotional cost? Across two experiments, we investigated this possibility. In Study 3a, we compared participants' change in state anxiety when exposed to sensitive-content screens with and without brief and detailed content descriptions. State anxiety was similar for participants who saw screens with and without brief content descriptions, but we found larger increases in state anxiety for detailed content descriptions. Therefore, detailed content descriptions negatively impact how people feel when they view sensitive-content screens. In Study 3b, we presented participants with a single sensitive-content screen, either with or without a brief content description, and gave them the opportunity to uncover it. Participants who uncovered the screen viewed the negative image and then rated their distress. Most participants uncovered the screen and, irrespective of condition, reported similar image-related distress. Taken together, brief descriptions do not negatively impact how people feel when they view sensitive-content screens or the forewarned content. Therefore, brief content descriptions do not create an

emotional cost. Social media platforms should move beyond merely warning about upcoming content and provide brief content descriptions indicating *what* the content depicts.

Introduction

Social media platforms—including Instagram and TikTok—use sensitive-content screens, a form of *trigger warning* (Bridgland et al., 2023), to minimise users' engagement with negative and potentially distressing content. But our previous research found that many people still engage with this content (Bridgland, Bellet et al., 2022) and do so repeatedly, partly because sensitive-content screens prompt information-seeking behaviour (Simister, Bridgland & Takarangi, 2023; Studies 1a and 1b). Content-related information presented alongside Instagram's typical warning—in the form of brief and detailed content descriptions (1-15 words in length)—can reduce people's uncovering behaviour (Simister, Bridgland, Williamson & Takarangi, 2023; Study 2). However, we do not know if this reduction in uncovering behaviour comes at an emotional *cost*; for example, perhaps merely reading about the content increases people's level of anxiety, and/or increases their distress if they decide to uncover the screens and view the forewarned content. We addressed these possibilities here. Specifically, in Study 3a, we investigated whether just viewing sensitive-content screens—alongside other neutral and positive images—is more anxiety provoking if they are presented with brief or detailed content descriptions. In Study 3b, we examined whether participants report content as more distressing when the preceding sensitive-content screen appears with a brief (vs. no) content description.

Sensitive-content screens in their current format are ambiguous; the warning accompanying the screens (i.e., "*Sensitive Content: This photo may contain graphic or violent content*") provides no specific information about the nature of the photo, for example, how it may be "graphic" or "violent". Therefore, it is perhaps unsurprising that some people repeatedly engage with such content despite the presence of screens (Simister, Bridgland &

Takarangi, 2023; Studies 1a and 1b). Indeed, people's tendency to uncover screened content fits with related research showing people find restricted content attractive ("forbidden fruit effect"; Weaver, 2011) and seek to resolve curiosity even when the consequences of doing so are uncertain but expected to be negative (e.g., electric shocks; "pandora effect"; Hsee & Ruan, 2016). To reduce ambiguity, we previously presented participants with sensitive-content screens that included content-related information, in the form of brief or detailed content descriptions (1-15 words in length), alongside the typical warning (Simister, Bridgland, Williamson & Takarangi, 2023; Study 2; Study 2). Participants uncovered these screens *less* often than screens without content-related information. Thus, perhaps including brief or detailed content descriptions on sensitive-content screens could work as a harm minimisation strategy. Indeed, strategies that lead people to uncover sensitive-content screens *less* often would reduce people's exposure to negative and potentially distressing content. This reduction may have an immediate emotional benefit (in terms of distress reduction), but also, cascading effects on other behaviours (e.g., reduction in distress driven self-harm behaviours; see Hetrick et al., 2020). However, there may be an emotional *cost* to viewing sensitive-content screens with content descriptions. Here, we theorise two opposing—yet not mutually exclusive—ways people may emotionally respond to viewing sensitive-content screens with content-related information.

On the one hand, reducing the ambiguity of sensitive-content screens by including content-related information may reduce people's experience of uncertainty/curiosity (Berlyne, 1954; Day, 1982; Loewenstein, 1994). Indeed, pilot participants reported being more certain ($N = 66$) and less curious ($N = 50$) about sensitive-content screens presented with brief and detailed content descriptions (as used in Simister, Bridgland, Williamson &

Takarangi, 2023; Study 2; Study 2), compared to screens without descriptions.[19] Perhaps more importantly though, when people *have* content-related information they are likely to experience a decrease in the negative emotional states that often accompany uncertainty/curiosity (e.g., anxiety, deprivation; Loewenstein, 1994). Therefore, content-related information on sensitive-content screens may not only reduce people's uncovering behaviour but it may also provide people with an emotional benefit (e.g., in terms of reducing anxiety) at the point of viewing such screens. Additionally, to the extent that detailed content descriptions offer greater reductions in uncertainty/curiosity compared to brief content descriptions, they may provide the greatest emotional benefit.

On the other hand, reading about the content may lead people to imagine it and/or imagine their own potential reactions to it. Imagining potential events in the future—or 'mental time travel' (Berntsen & Jacobsen, 2008)—may be *as* anxiety provoking or distressing as viewing the content itself would be (see Blackwell, 2019; 2021 for review). In fact, many people already experience anticipatory anxiety at the time of viewing sensitive-content screens (Takarangi et al., 2023), and traditional trigger warnings (for review, see Bridgland et al., 2023). Therefore, adding content-related information on sensitive-content screens may exacerbate such anticipatory anxiety *because* the details in and of themselves may be aversive, and/or people have more details to imagine compared to when there is no content-related information, per the typical sensitive-content screens. Thus, content-related information on sensitive-content screens may come at an emotional cost for people at the point of viewing such screens—irrespective of whether they decide to view the forewarned content. Additionally, to the extent that detailed content descriptions are more aversive,

[19] Brief content descriptions (certainty: $M = 3.1$; $SD = 1.2$; $d = 1.3$; curiosity: $M = 1.6$; $SD = 0.9$; $d = 1.1$); detailed content descriptions (certainty: $M = 3.6$; $SD = 1.2$; $d = 1.4$; curiosity: $M = 1.7$; $SD = 0.9$; $d = 1.1$); no content descriptions (certainty: $M = 1.9$; $SD = 0.9$; curiosity: $M = 2.2$; $SD = 1.1$; on scales of 1 = very slightly or not at all certain/curious to 5 = extremely certain/curious).

and/or prompt people to imagine more details compared to brief content descriptions, they may create the greatest emotional cost.

To investigate these possibilities, in our first experiment we assessed participants' state anxiety pre and post a passive image-viewing task: participants viewed multiple sensitive-content screens (without an option to uncover) amongst neutral and positive images for 5s each (total task time = 5 min; adapted from Takarangi et al., 2023). We randomly allocated participants to a content description condition (sensitive-content screens appeared with a brief or detailed content description) or the control condition (screens appeared without content descriptions). With regards to the *presence* of content-related information on sensitive-content screens, we predicted an interaction between condition (with vs. without content descriptions) and time (pre- to post image-viewing task) on state anxiety (Hypothesis 1), but we had competing predictions regarding the pattern of the relationship. If content-related information reduces people's uncertainty/curiosity, then we would expect participants who see sensitive-content screens with content descriptions to show smaller increases in state anxiety compared with participants who see sensitive-content screens without content descriptions. But, if content-related information exacerbates anticipatory anxiety, we would expect the opposite pattern. With regards to the *level* of content-related information on sensitive-content screens, we also predicted an interaction between condition (brief vs. detailed content descriptions) and time on state anxiety (Hypothesis 2). Again, we had competing predictions regarding the pattern of the relationship, based on the same theoretical ideas. Relative to screens with brief content descriptions, detailed content descriptions should elicit smaller increases in state anxiety to the extent they offer greater reductions in uncertainty/curiosity, or greater increases in state anxiety to the extent they are more aversive, and/or prompt people to imagine more details.

Open Practices Statement

The Flinders University Human Research Ethics Committee approved this research and we pre-registered it on the OSF (Study 3a: https://osf.io/jgn3f; Study 3b: https://osf.io/q9crk). We have reported all measures, conditions, and data exclusions. The supplementary materials are at the end of the chapter and the data, including codebooks describing all variables, can be found at: https://osf.io/vh42c/. We analysed data using SPSS (Version 25), therefore there is no separate analysis code.

Study 3a

Method

Participants

At 80% power, to detect *d* of at least 0.4 (the smallest effect we were interested in based on practical significance and financial constraints), Brysbaert (2019) recommends a minimum of *n* = 100 participants per group for a between-subjects design. Therefore, we aimed to collect 300 participants. We recruited participants from the United States using Amazon's Mechanical Turk (MTurk) via Cloud Research. To promote data quality and minimise bots/server farmers, we screened out participants who failed a captcha, scored less than 8/10 on an English proficiency test (Moeck et al., 2022), selected "Konnect" (a bogus platform included to detect inattentive responses) when asked about social media use, and/or indicated they do not use Instagram (since we only wanted to recruit Instagram users). Of 357 participants who completed the survey and received a payment of $1.50 USD, we excluded 49 as per our pre-registration: 31 failed to achieve 6/8 on forced choice questions about the positive and neutral image content[20]; six failed an embedded attention check; six reported leaving during the image-viewing task; and six did not pass the cultural check (we showed

[20] Since participants passively viewed images, we gave participants an 8-item forced choice test about the content of the neutral and positive images (e.g., Which of the following did you see? [select one]; "mountain", "waterfall") to make sure they were paying attention during the task.

participants a picture of an eggplant and asked them what it is called [we expected participants from the United States to answer: “eggplant”]). Thus, our final sample comprised 308 participants (no content description condition: n = 105; brief content description condition: n = 103; detailed content description condition: n = 100).

Participants were aged 20-69 years (M = 38.8, SD = 10.3) and included 58.1% women (n = 179) and 39.9% men (n = 123); three participants identified as non-binary (1.0%), and three participants preferred not to report their gender (1.0%). Our sample was predominantly European American/White (73.4%; n = 226); other participants were of African American/Black (10.7%; n = 33), Hispanic (5.8%; n = 18), Asian (4.5%; n = 14), Middle Eastern (0.6%; n = 2), or mixed-race (3.2%; n = 10) descent; four (1.3%) participants specified nationality (e.g., USA) when given the option to self-describe their ethnicity, and one participant (0.3%) preferred not to provide their ethnicity. Most participants (50.6%; n = 156) reported income between $45,000-$140,000 and were predominantly (59.4%; n = 183) college graduates (Supplementary Table S3.1). Moreover, most participants (52.3%; n = 161) reported they had used Instagram every day over the past week, and for one hour or more on an average day in the last 30 days (70.8%; n = 218; Supplementary Table S3.2). Most participants also reported they have seen sensitive-content screens on their own Instagram feed (71.8%; n = 221). In addition to Instagram, most participants reported using YouTube (88.0%; n = 271), Facebook (82.5%; n = 254), Twitter (71.8%; n = 221), Reddit (61.4%; n = 189), and TikTok (53.2%; n = 164) on a regular basis.

Materials and Procedure

We adapted the procedure from our prior work (Takarangi et al., 2023). Participants signed up for a “social media engagement” study. After providing informed consent, participants indicated how many days of the last 7, and for how many hours on average each day, they used Instagram (over the last 30 days; Bridgland, Bellet et al., 2022). To reduce

suspicion about the true nature of our experiment, we also asked participants to report how often (not at all, sometimes, often, very often) they view different types of images (e.g., portraits, animals) on Instagram. Participants then completed the short-form *State-Trait Anxiety Inventory* (STAI-6; Marteau & Bekker, 1992).[21] Participants rated their current feelings across 6-items (e.g., *I am worried*; 1 = *not at all* to 4 = *very much*). The STAI-6 had good internal consistency (current experiment: α = .91 [Time 1], .92 [Time 2]).

Next, participants completed our *image-viewing task*. Participants viewed the 20 most neutral, positive, and negative images (60 total) from the Nencki Affective Picture System (NAPS; Marchewka et al., 2014; valence ratings: 1 = *negative* to 9 = *positive*), which include content commonly found on Instagram (e.g., people, animals, landscapes). All images appeared in an Instagram border with non-functional like and comment buttons for 5s each (total task time = 5 min). Consistent with Instagram's sensitive-content screen format (as of June 2023), negative images were blurred and accompanied by a warning ("*Sensitive Content: This photo may contain graphic or violent content*"). Here, participants did not have the option to uncover sensitive-content screens to view negative images. We randomly assigned participants to either the *no*, *brief,* or *detailed* content description condition (as used in Simister, Bridgland, Williamson & Takarangi, 2023; Study 2). In the *no* content description condition, we presented sensitive-content screens as they typically appear on Instagram (Figure 2.1a). In the *brief* content description condition, we presented sensitive-content screens with brief descriptions of the negative images (e.g., "Burns"; Figure 2.1b). In the *detailed* content description condition, we presented screens with detailed descriptions of the negative images (e.g., "A person receives treatment for a severe burn on their hand";

[21] Participants also completed the negative subscale of the *Positive and Negative Affect Schedule* (PANAS; current study*:* α = .94 [Time 1], .93 [Time 2]*;* Watson et al., 1988; along with four positive adjectives to make our focus on negative adjectives less obvious) here and immediately after the image-viewing task, but because this variable is not pertinent to testing our hypotheses, these data appear in the supplementary materials.

Figure 2.1c).[22] Prior to beginning the task, we told participants that the images would appear for a fixed duration, and to promote attention to the images, we told participants we would ask them questions about the images at the end of the task. We also told *all* participants that they may see screens where a negative image had been covered.

After the image-viewing task, participants repeated the STAI, and so we could assess our competing theories regarding the role of imagination, and uncertainty/curiosity, participants rated: how vividly they imagined the content of the screened image (1 = *perfectly clear and as vivid as normal vision* to 5 = *no image at all*; adapted from the Vividness of Visual Imagery Questionnaire [VVIQ; Marks, 1973]), as well as how uncertain (1 = *no at all uncertain* to 5 = *extremely uncertain*) and curious (1 = *no at all curious* to 5 = *extremely curious*) they felt about the content of the screened image. Then, we asked participants to indicate whether they have seen sensitive-content screens on their Instagram feed (yes/no), if they had left the image-viewing task for any extensive period (yes/no; if so, when/for how long), or experienced any technical issues (yes/no). Finally, participants completed demographics. We then debriefed participants.

Results and Discussion

Preliminary Analyses

First, we compared Instagram use (Supplementary Table S3.3), previous exposure to sensitive-content screens (Supplementary Table S3.4), and demographics (including age, gender, income, and education; Supplementary Tables S3.5 & S3.6) between conditions; all patterns were comparable across conditions.

[22] We developed content descriptions from pilot study data: MTurk participants (N = 55) viewed the negative images and described them in one sentence. We then modified the descriptions to match them on style and word length across images (brief: range = 1-3, M = 2.6, SD = 0.7; detailed: range = 11-15, M = 13.1, SD = 1.5; see Appendix B for all content descriptions).

Hypothesis Testing

Next, we turned to our primary research aims. Recall, we were first interested in whether the *presence* of content-related information on sensitive-content screens would have an emotional cost, relative to sensitive-content screens without content-related information. Specifically, we predicted an interaction between condition (with vs. without content descriptions) and time (T1 [pre-image-viewing task] vs. T2 [post-image-viewing task]) on state anxiety (Hypothesis 1); we also had competing predictions regarding the pattern of the relationship. To test our predictions, we ran a 2 (condition: content descriptions, no content descriptions) x 2 (time: T1, T2) mixed ANOVA on participants' STAI scores (Table 3.1).[23] Overall, state anxiety was higher (i.e., more negative) at T2 compared with T1; a main effect of time for state anxiety, $F(1, 306) = 75.69$, $p < .001$, $\eta_p^2 = .20$. There was *no* difference in state anxiety between participants who saw sensitive-content screens with vs. without content descriptions; a nonsignificant main effect of condition for state anxiety, $F(1, 306) = 0.28$, $p = .60$, $\eta_p^2 = .001$. But, as predicted, the effect of time on state anxiety depended on whether participants saw sensitive-content screens with or without content descriptions; a significant interaction between condition and time for state anxiety, $F(1, 306) = 4.65$, $p = .03$, $\eta_p^2 = .02$. Specifically, in line with the idea that content-related information may exacerbate anticipatory anxiety *because* the details in and of themselves are aversive, and/or people have more to imagine, participants who saw sensitive-content screens with content descriptions showed larger increases in state anxiety (from T1 to T2) compared with participants who saw sensitive-content screens without content descriptions. Thus, there appears to be an emotional cost associated with the presence of content-related information.

[23] We also ran this analysis for the negative affect scale of the PANAS (as pre-registered) and report these analyses in full in the supplementary materials. Notably, the interaction was not significant; participants who saw sensitive-content screens with content descriptions showed similar increases in negative affect (from T1 to T2) compared with participants who saw sensitive-content screens without content descriptions.

However, recall we were also interested in whether the *level* of content-related information on sensitive-content screens would influence the emotional cost. Specifically, we predicted an interaction between condition (brief vs. detailed content descriptions) and time (T1 vs. T2) on state anxiety (Hypothesis 2); again, we also had competing predictions regarding the pattern of the relationship. To test this hypothesis, we ran another 2 (condition: brief content description, detailed content description) x 2 (time: T1, T2) mixed ANOVA on STAI scores (Table 3.1).[24] Like the previous ANOVA, there was a main effect of time, $F(1, 201) = 88.24$, $p = <.001$, $\eta_p^2 = .88$, and a nonsignificant main effect of condition, $F(1, 201) = 0.39$, $p = .53$, $\eta_p^2 = .002$, for state anxiety; the pattern of the data was also the same. As predicted, the effect of time on state anxiety depended on whether participants saw sensitive-content screens with brief or detailed content descriptions; a significant interaction between condition and time for state anxiety, $F(1, 201) = 4.29$, $p = .04$, $\eta_p^2 = .02$. Specifically, in line with the idea that *detailed* content-related information may be more aversive/evoke more elaborated mental imagery, and thus exacerbate anxiety further, participants who saw sensitive-content screens with detailed content descriptions showed larger increases in state anxiety (from T1 to T2) compared with participants who saw sensitive-content screens with brief content descriptions. Thus, although people uncover sensitive-content screens *less* often when they include brief *or* detailed content descriptions (Simister, Bridgland, Williamson & Takarangi, 2023; Study 2), detailed content-related information creates a larger emotional cost for people at the point of viewing such screens.

[24] We also ran this analysis for the negative affect scale of the PANAS (as pre-registered) and report these analyses in full in the supplementary materials. Notably, the interaction showed the same pattern as state anxiety.

Table 3.1

Means (and Standard Deviations) for State Anxiety, by Condition and Time

Condition	Time		
	Pre-task *M* (*SD*)	Post-task *M* (*SD*)	Total *M* (*SD*)
No Content Description	10.4 (4.1)	11.7 (4.3)	11.1 (4.1)
Content Description	10.2 (4.3)	12.4 (4.9)	11.3 (4.1)
Total	10.3 (4.2)	12.2 (4.7)	
Brief Content Description	10.3 (4.3)	12.0 (4.9)	11.2 (4.3)
Detailed Content Description	10.2 (4.3)	12.9 (4.8)	11.5 (4.3)
Total	10.2 (4.3)	12.4 (4.9)	

Note. Possible scores for State Anxiety range from 6 to 24.

When looking at the descriptive statistics in Table 1, we noticed that the brief content description condition appeared to have a similar change in state anxiety (from T1 to T2) to the no content description condition. Therefore, we suspected the detailed content description condition was driving the effect we found when comparing the no content description condition with the content description condition (which had the brief and detailed conditions collapsed within it). To examine this possibility, we ran another 2 x 2 mixed ANOVA on the STAI, specifically comparing the change from T1 to T2 for the brief content description and no content description conditions; this analysis was not pre-registered. Like the previous ANOVAs, there was a main effect of time, $F(1, 206) = 41.59$, $p = <.001$, $\eta_p^2 = .17$, and a nonsignificant main effect of condition, $F(1, 206) = 0.02$, $p = .89$, $\eta_p^2 = .0001$, for state anxiety; the pattern of the data was also the same. However, the effect of time on state anxiety did not depend on whether participants saw sensitive-content screens with brief content descriptions or without; a nonsignificant interaction between condition and time for

state anxiety, $F(1, 206) = 0.70$, $p = .40$, $\eta_p^2 = .003$. Therefore, participants who saw sensitive-content screens with brief content descriptions showed similar increases in state anxiety (from T1 to T2) compared with participants who saw sensitive-content screens without content descriptions. Thus, not only do people uncover sensitive-content screens *less* often when they include brief content descriptions (Simister, Bridgland, Williamson & Takarangi, 2023; Study 2)—which may provide an immediate and ongoing emotional benefit—but there appears to be no additional emotional cost associated with viewing sensitive-content screens with brief content-related information.

Additional Pre-Registered Analyses

Finally, we wondered whether participants' responses to post-task questions about how *vividly* they imagined the content of screened images, as well as how *uncertain* and *curious* they felt about the content differed by condition. We ran a series of one-way ANOVAs to examine this possibility (Supplementary Table S3.7). We found no differences in imagination, uncertainty or curiosity between participants who saw sensitive-content screens with brief, detailed, or no content descriptions. Notably, participants in the detailed content description condition did not report having more vivid imagery, compared to participants who saw less content-related information—which is inconsistent with what we would predict based on our theory regarding imagination. In fact, all participants reportedly having moderately clear and vivid imagery. However, it is possible participants had difficulty providing *one* retrospective vividness rating for what were *multiple* episodes of imagination—interrupted also by other neutral and positive images. Indeed, although people can reliably evaluate the vividness of single episodes of imagination (Pearson et al., 2011), such awareness may not translate to the present task. It is also possible that imagination may not be the driving mechanism behind the detailed content description effect. Put differently,

people may have a negative reaction (e.g., experience state anxiety) to the content-related information in and of itself, *without* imagining the content.

Taken together, it may be suitable to include *brief* content descriptions on sensitive-content screens on Instagram as a harm minimisation strategy. However, it is possible that brief content descriptions *enhance* how negative (or distressed) a person feels if they decide to uncover them and view the forewarned images. This emotional cost is especially concerning given that we know people uncover sensitive-content screens at a high rate. Therefore, we explored this possibility in Study 3b.

Study 3b

We regularly receive information about upcoming emotionally unpleasant content—whether in the form of sensitive-content screens or warnings in other contexts (e.g., film content; Bridgland et al., 2023). Such anticipatory information is intended to mitigate potential negative impact; for example, by giving people an opportunity to "brace" for the worst when anticipating their reaction to the content (e.g., "I am preparing myself for the worst"; Sweeny & Shepperd, 2010). But there is mixed evidence for the impact of anticipatory information on subsequent emotional responses. On the one hand, some evidence suggests that anticipatory information, specifically regarding the valence of upcoming content (e.g., that an image may be negative) directs people's attention towards the content, and increases its unpleasantness (e.g., Lin et al., 2012; for an overview see Shafir & Sheppes, 2020). In fact, when anticipatory information enhances attention towards *negative* content, it is effortful for people to decrease their emotional responses (e.g., distress) to that content (Shafir & Sheppes, 2020). Thus, people may experience *more* negative emotions when content is preceded by anticipatory information (vs. not). In the present experiment then, the presence of anticipatory information, in the form of a brief content description, may enhance

how negative (or distressed) a person feels if they decide to uncover the screen and view the negative image.

On the other hand, research on traditional trigger warnings suggests that forewarning has a trivial effect on people's emotional responses towards the forewarned content (Bridgland et al., 2023). Indeed, although trigger warnings about upcoming content (e.g., *"torture, maltreatment, and death*") induce anticipatory anxiety and negative affect (e.g., Bridgland et al., 2019), they do not seem to enhance how negative people find the forewarned content, or how negative people feel while viewing it, relative to content without such anticipatory information (e.g., Boysen et al., 2021; Sanson et al., 2019). Therefore, in the present study, the presence of a brief content description may have minimal (to no) impact on how negative (or distressed) a person feels if they decide to uncover the screen and view the negative image.

Taken together, the literature suggests competing possibilities for how brief content descriptions on sensitive-content screens might affect people's reactions to negative images—*once* they have decided to uncover the screens. To test these possibilities, we used a brief image-viewing task, with a single sensitive-content screen (Bridgland, Bellet et al., 2022; Study 2): participants were randomly allocated to see the sensitive-content screen with or without a brief content description. Participants had the option to uncover the screened image or not; if they decided to uncover it, they saw the negative image before rating their distress. Based on our previous findings (Simister, Bridgland, Williamson, Takarangi, 2023), we predicted that a greater proportion of participants in the no content description condition would uncover the sensitive-content screen than in the brief content description condition (Hypothesis 1). But, *within* the sub-set of participants who decide to uncover sensitive-content screens, we had competing predictions about the effect of content description condition on participants' distress. If anticipatory information, in the form of a brief content

descriptions, enhances how negative (or distressed) people feel, then distress after viewing the negative image will be higher for participants who see the sensitive-content screen *with* vs. without a brief content description (Hypothesis 2a). However, if brief content descriptions have a largely trivial effect on emotional responses towards forewarned content—like trigger warning messages do—then distress after viewing the negative image will be similar whether participants see the sensitive-content screen with *or* without brief content descriptions (Hypothesis 2b).

Method

Participants

We powered for an independent samples t-test to test our competing hypotheses (Hypotheses 2a and b)—which were our main hypotheses of interest. We followed Brysbaert (2019) as in Study 3a (n = 100 participants per group), but because the conditions were quasi-experimental, we aimed to collect until we had *at least* 100 per condition (i.e., 100 participants who uncovered the sensitive-content screen with a brief content description, and 100 participants who uncovered the sensitive-content screen with no content description). We used the same recruitment and screening procedures as in Study 3a. Of 245 participants who completed the survey and received a $1.00 USD payment, we excluded seven as per our pre-registration: three did not pass the cultural check; three reported leaving during the image-viewing task; one experienced a technical issue that interfered with making a distress rating. Thus, our final sample comprised 238 participants (no content description condition: n = 115; brief content description condition: n = 123).

Participants were aged 19-74 years (M = 35.6, SD = 8.8) and included 70.2% women (n = 167) and 29.0% men (n = 69); two participants identified as non-binary (0.8%). Our sample was predominantly European American/White (70.2%; n = 167); other participants were of African American/Black (13.0%; n = 31), Hispanic (9.2%; n = 22), Asian (1.7%; n =

4), Indigenous (1.3%; *n* = 3), Pacific Islander (0.4%; *n* = 1), or mixed-race (4.2%; *n* = 10) descent. Most participants (56.5%; *n* = 134) reported income between $45,000-$140,000 and were predominantly (58.4%; *n* = 139) college graduates (Supplementary Table S3.1). Moreover, most participants (52.9%; *n* = 126) reported they had used Instagram every day over the past week, and for one hour or more on an average day in the last 30 days (71.4%; *n* = 170; Supplementary Table S2). Most participants also reported they have seen sensitive-content screens on their own Instagram feed (75.2%; *n* = 179). In addition to Instagram, most participants reported using YouTube (84.0%; *n* = 200), Facebook (81.9%; *n* = 195), Reddit (61.8%; *n* = 147), TikTok (60.5%; *n* = 144), and Twitter (54.2%; *n* = 129) on a regular basis.

Materials and Procedure

The cover story and the initial phase of the experiment was the same as in Study 3a: after providing informed consent, participants completed Instagram use questions and rated how often they view different types of images (e.g., portraits). Next, participants completed our *brief image-viewing task*. Participants viewed a set of 5 neutral and 5 positive images sourced from the NAPS (Marchewka et al., 2014)—randomly selected from one of four sets of 10 images (matched on valence and arousal ratings)[25]—in a randomised order. As in Study 3a, all images appeared in an Instagram border with non-functional like and comment buttons. There was a 3 sec delay between the presentation of images and when the '*Next Photo*' button appeared, so that we had better control over participants rushing through the images (Simister, Bridgland & Takarangi, 2023; Study 1b). Participants then viewed a single sensitive-content screen—randomised from a pool of the 20 most negative images from the NAPS (as used in Study 3a). Like Study 3a prior to beginning the task, we told *all*

[25] We created four sets of 5 neutral and 5 positive images matched on overall valence and arousal ratings (set 1: valence $M = 6.5$, $SD = 1.6$, arousal $M = 4.6$, $SD = 1.3$; set 2: valence $M = 6.6$, $SD = 1.6$, arousal $M = 4.8$, $SD = 0.8$; set 3: valence $M = 6.6$, $SD = 1.6$, arousal $M = 3.7$, $SD = 1.4$; set 4: valence $M = 6.6$, $SD = 1.7$, arousal $M = 3.8$, $SD = 1.1$). We ran a series of one-way ANOVAs that revealed no significant differences between sets in valence, $F(3, 36) = 0.02$, $p = 1.00$, $\eta^2 = .0002$, or arousal, $F(3, 36) = 2.37$, $p = .09$, $\eta^2 = .165$.

participants that they may see screens where a negative image has been covered. Half of the participants were randomly allocated to the *no content description* condition: the sensitive-content screen appeared without a content description. The other half of the participants were randomly allocated to the *brief content description* condition: the sensitive-content screen appeared with a brief description (e.g., "Deceased person"). When participants saw the sensitive screen—irrespective of condition—they had the option to uncover it (after the 3 sec delay; select *See Photo*) and view the negative image or go on to the next image (select *Next Photo*). The image-viewing task ended here for participants who did not uncover the sensitive-content screen. Participants who uncovered the sensitive-content screen could view the negative image for as long or as little as they wanted to—the *Next Photo* button appeared automatically (unlike for previous images in the image-viewing task). Once participants selected *Next Photo,* they were asked to respond to the following question: "*How distressed do you feel right now?*" (0 = *not at all distressed*, to 100 = *extremely distressed*). Once participants made their rating, the image-viewing task automatically ended.

After the image-viewing task, participants indicated their familiarity with sensitive-content screens; they also indicated if they had looked away from any negative images, left the image-viewing task for any extensive period, or experienced any technical issues. Finally, participants completed demographics. We then fully debriefed participants.

Results and Discussion

Preliminary Analyses

First, we compared Instagram use (Supplementary Table S3.8), previous exposure to sensitive-content screens (Supplementary Table S3.9), and demographics (including age, gender, income, and education; Supplementary Tables S3.10 & S3.11) between the randomly allocated conditions. All patterns were comparable across conditions, except participants in the brief content description reported using Instagram for slightly longer on an average day

(over the last 30 days), relative to participants in the no content description condition.[26] Per our pre-registration, we re-ran these analyses within the sub-set of participants who uncovered the sensitive-content screen from each condition (uncovered screen with brief content description, n = 101, and uncovered screen with no content description, n = 100; Supplementary Tables S3.12, S3.13, S3.14, & S3.15); however, because most participants uncovered the sensitive-content screen, there is substantial overlap in the participants analysed . Consistent with the differences in the randomly allocated conditions, we found participants in the brief content description who decided to uncover the sensitive-content screen, reported using Instagram for slightly longer on an average day (over the last 30 days), relative to participants in the no content description condition who decided to uncover the sensitive-content screen. We also found participants were less likely to uncover the sensitive-content screen if they saw brief content descriptions (as we predicted) *and* had previous exposure to sensitive-content screens. It is possible that the combination of the brief content description and familiarity with the potentially distressing nature of the forewarned content (i.e., from previous experiences) meant that participants were even more cautious of uncovering sensitive-content screens. To isolate our main effect of interest—the effect of content-related information—we statistically controlled for Instagram use and previous exposure to sensitive-content screens in our analyses related to image-related distress, as per our pre-registration.

Hypothesis Testing

Uncovering Behaviour.

Overall, 84.5% (n = 201) of participants uncovered the sensitive-content screen—consistent with previous rates of uncovering (Bridgland, Bellet et al., 2022; Simister, Bridgland & Takarangi, 2023; Studies 1a and 1b). To examine whether content description

[26] We note this analysis includes ordinal data.

condition influenced uncovering behaviour (Hypothesis 1), we ran a chi-square analysis comparing the proportion of participants in the *no content description* condition who uncovered the sensitive-content screen vs. the proportion of participants in the *brief content description* condition who uncovered the sensitive-content screen. Contrary to our predictions and existing research (Simister, Bridgland, Williamson & Takarangi, 2023; Study 2), we found no significant difference between the two conditions, $\chi^2 = 1.06$, $df = 1$, $p = .303$ (no content description: 87.0%; $n = 100$; brief content description: 82.1%; $n = 101$). Therefore, most participants uncovered the sensitive-content screen and did so irrespective of whether they saw the screen with or without a brief content description.

One explanation for the discrepancy between this finding and existing research is differences in methodology. In our previous study, participants saw 30 sensitive-content screens total (10 per condition; Simister, Bridgland, Williamson & Takarangi, 2023; Study 2), whereas here we measured participant's uncovering behaviour in response to *one* sensitive-content screen. To compare these data more closely, we analysed participants' uncovering behaviour on the first sensitive-content screen they saw during the image-viewing task in our previous study.[27] Consistent with the present study, we found no significant difference in uncovering behaviour between the two conditions on the first trial, $\chi^2 = 1.97$, $df = 1$, $p = .160$ (no content description: 56.6%; $n = 61$; brief content description: 43.5%; $n = 47$).We theorise that the desire to resolve uncertainty/curiosity (Loewenstein, 1994), and/or to test whether the content description matches the forewarned content, is especially strong on the *first* sensitive-content screen, such that the effect of content description condition that we observed over 30 trials is weakened using a single trial. Indeed, our previous work which examined participant's uncovering behaviour for the first and subsequent sensitive-content screens (30

[27] We note there were other task differences (e.g., participants in our previous study could have viewed sensitive-content screens within the first 10 trials, unlike the present study) and variability in sample sizes.

total) found a steady decrease in uncovering behaviour as trials progressed (i.e., from 51.9% - 86% over the first five screens to just below 30% over the final 10 screens; Simister, Bridgland & Takarangi, 2023; Study 1b).

Image-Related Distress.

Overall, participants who decided to uncover the sensitive-content screen were mildly to moderately distressed after viewing the negative image (M = 33.6, SD = 29.5; 0 = *not at all distressed*, to 100 = *extremely distressed*). Recall that *within* the sub-set of participants who decided to uncover the sensitive-content screen, we had competing predictions about the effect of content description condition on participants' distress. We ran an independent samples t-test on the quasi-experimental conditions to test the effect of content description condition. Consistent with the idea that brief content descriptions have a largely trivial effect on emotional responses towards forewarned content (Hypothesis 2b), distress after viewing the negative image was similar (i.e., mild to moderate) irrespective of whether participants saw the sensitive-content screen with (M = 34.1, SD = 31.7) *or* without a brief content description (M = 36.6, SD = 28.8), $t(199) = 0.579$, $p = .56$, $d = 0.08$, 95% CI [-0.20, 0.36]. This finding was robust to our pre-registered sensitivity analyses; the effect of content description condition on participants' distress remained non-significant when we statistically controlled for the time participants spent viewing the negative image, Instagram use, and previous exposure to sensitive-content screens (see supplementary materials). To quantify evidence that our data favoured the null hypothesis (Bayes Factors >1) relative to the alternative hypothesis of an effect of content description, we obtained Bayes Factors (BF_{01}; with the default prior in SPSS) using participants' image-related distress (though we note we did not pre-register this analysis). We followed Wetzels et al.'s (2011) guidelines: anecdotal = 1–3, substantial = 3–10, strong = 10–30, very strong = 30–100, decisive >100. We found

substantial evidence (BF_{01}= 7.70) for no difference between conditions for image-related distress.

General Discussion

Content-related information, in the form of descriptions (1-15 words in length), can reduce how often people view negative and potentially distressing images when added to Instagram's sensitive-content screens (Simister, Bridgland, Williamson & Takarangi, 2023; Study 2). Here, across two studies we examined whether this reduction in negative image exposure comes at an emotional cost. Specifically, we investigated whether just viewing sensitive-content screens—alongside other neutral and positive images—is more anxiety provoking if they are presented with brief or detailed content descriptions (Study 3b), and whether participants report content as more distressing when the preceding sensitive-content screen is presented with vs. without a brief content description (Study 3b).

Overall, we found that exposure to sensitive-content screens, irrespective of whether they were accompanied by content-related information or not, increased people's state anxiety. This finding aligns with other research demonstrating that sensitive-content screens (Takarangi et al., 2023)—as well traditional trigger warnings more generally (Bridgland et al., 2023)—cause anticipatory anxiety. However, our data suggest *detailed* content-related information exacerbates such anticipatory anxiety (relative to brief content descriptions), which is in line with the idea that people find the details in and of themselves aversive, and/or have more to imagine. In fact, the increase in anxiety was similar to that induced by the well-known Trier Social Stress Test (TSST; Degroote et al., 2020).[28] Therefore, detailed content-related information creates an emotional cost for people at the point of viewing such

[28] To compare to studies using the full version of the STAI, we multiplied individual participants' mean STAI scores by 20.

screens—a cost that could arguably outweigh the emotional benefits associated with the reduction in uncovering behaviour.

However, people experienced similar anticipatory anxiety when they viewed sensitive-content screens with *brief* content-related information and sensitive-content screens without content descriptions (i.e., as they typically appear on Instagram). Therefore, although brief content-related information did not mitigate anticipatory anxiety, it also did not exacerbate it like detailed content-related information. Perhaps brief content descriptions reduce the ambiguity of sensitive-content screens—and thus people's experience of uncertainty/curiosity (Berlyne, 1954; Day, 1982; Loewenstein, 1994)—while withholding specific content details content that may otherwise be aversive and/or increase imagination. For example, knowing a preceding image contains "*Burns*" may be informative enough without the additional information offered by a detailed counterpart (e.g., "*A person receives treatment for a severe burn on their hand*").

Additionally, we found no evidence to suggest that sensitive-content screens with brief content descriptions *enhance* how negative (or distressed) a person feels when they decide to uncover the screen and view the forewarned image, relative to sensitive-content screens without content descriptions. In fact, image-related distress levels overall were mild to moderate irrespective of whether people saw sensitive-content screens with or without content-related information. This finding aligns with research demonstrating that traditional trigger warnings, with varying levels of information, have a largely trivial effect on people's emotional responses towards forewarned content (Bridgland et al., 2023).

Taken together, we found differential emotional costs for brief and detailed content-related information. Although people uncover sensitive-content screens *less* often when they include a brief *or* detailed content description (Simister, Bridgland, Williamson & Takarangi, 2023; Study 2), detailed content-related information appears to exacerbate anticipatory

anxiety. However, we found no additional emotional cost associated with viewing sensitive-content screens with brief content-related information *or* with viewing the forewarned image following such screens.

Our findings have practical implications for Instagram and social media platforms alike (e.g., TikTok). These platforms should consider adding brief content-related information (1-3 words in length) to sensitive-content screens—whether creating these descriptions is the responsibility of the user posting the content or for the artificial intelligence/moderators who screen the content. Indeed, brief content-related information can increase people's ability to make an informed uncovering decision—which ultimately may reduce how often people view negative and potentially distressing images (Simister, Bridgland, Williamson & Takarangi, 2023; Study 2)—all while avoiding increases in anticipatory anxiety and image-related distress. Including *brief* content descriptions on sensitive-content screens would therefore provide an immediate and ongoing benefit for users. Furthermore, adopting such a harm minimisation strategy could help balance out Instagram's (self-proclaimed) *need* to create a safe space for people to talk about their experiences (e.g., mental health struggles), and post related non-graphic content online, with *their responsibility* to reduce the potential harm that such content might have on other people (Mosseri, 2019a, 2019b).

Our study has several limitations. First, we used a quasi-experimental design to see how content-related information influenced participants' image-related distress *when* they decided to view the forewarned content (Study 3b)—because people have this choice in real life—but doing so may have created a selection bias. For example, perhaps image-related distress did not differ between content description conditions because people who are more vulnerable (or sensitive) to amplifications in distress caused by the brief content description decided not to uncover the sensitive-content screen. However, this possibility seems unlikely

because past research on vulnerable people (operationalised as people with more severe psychopathological symptoms, e.g., of depression) suggests they are no more likely to avoid uncovering sensitive-content screens (Simister, Bridgland & Takarangi, 2023; Studies 1a and 1b). Nonetheless, future research could employ a forced-choice paradigm, whereby participants are shown a sensitive-content screen with or without brief content-related information, before being shown the forewarned image and making a distress rating.

Second, we examined participants' behaviour and emotional reactions after one screen (Study 3b); therefore, we do not know whether behaviour and emotional reactions change on subsequent screens, or if there is an interplay between content-related information and uncovering behaviour over time. Perhaps when people see sensitive images preceded by brief content-related information, over time they learn to distinguish types of content they are better able to cope with (relative to having no information at all), and accordingly, change their uncovering behaviour in an adaptive manner. Indeed, we know that knowing *what* the sensitive content is helps people avoid certain content that they anticipate will be distressing/too difficult to cope with (Simister, Bridgland, Williamson & Takarangi, 2023; Study 2). Future research could examine participants' behaviour and emotional reactions over a series of sensitive-content screens.

Third, we examined participants' anticipatory anxiety (Study 3a) and image-related distress (Study 3b) to investigate whether content-related information creates an emotional cost. Although existing research on sensitive-content screens (e.g., Takarangi et al., 2023)—and traditional trigger warnings (Bridgland et al., 2023)—has used similar measures of affect, future research could examine a broader cluster of emotional reactions (e.g., intrusions) and outcomes (e.g., the meaning people derive from the content) to expand our understanding of the impact of content-related information.

Across two studies we examined whether providing people with brief and detailed content-related information on sensitive-content screens, in the form of descriptions, comes at an emotional cost. We found some evidence for this possibility with respect to detailed content-related information, which exacerbated people's anticipatory anxiety. However, including *brief* content-related information on sensitive-content screens neither increased people's anticipatory anxiety, nor their image-related distress, relative to screens without content descriptions. Therefore, brief content-related information offers a reduction in negative image exposure without creating an emotional cost. Thus, including brief content descriptions on sensitive-content screens is a harm minimisation strategy that Instagram and other social media platforms should consider.

Supplementary Materials

Table S3.1

Demographic Characteristics for Studies 3a and 3b

Variable	Study 3a % (*n*)	Study 3b % (*n*)
Household income		
<$20,000	11.4% (35)	8.8% (21)
$20,000 - $45,000	22.7% (70)	25.2% (60)
$45,000 - $140,000	50.6% (156)	56.3% (134)
$140,000 - $150,000	4.9% (15)	4.2% (10)
$150,000 - $200,000	5.2% (16)	2.9% (7)
>$200,000	5.2% (16)	2.5% (6)
Education		
Less than high school graduate	1.3% (4)	0.4% (1)
High school graduate	9.7% (30)	9.7% (23)
Some college	29.5% (91)	31.5% (75)
College Graduate	59.4% (183)	58.4% (139)

Table S3.2

Social Media Use for Studies 3a and 3b

Variable	Study 3a % (*n*)	Study 3b % (*n*)
Social media platform		
Facebook	82.5% (254)	81.9% (195)
Instagram	100.0% (308)	100.0% (238)
Twitter	71.8% (221)	54.2% (129)
Snapchat	34.1% (105)	41.2% (98)
WhatsApp	24.7% (76)	15.5% (37)
Tumblr	1.7% (22)	8.0% (19)
YouTube	88.0% (271)	84.0% (200)
TikTok	53.2% (164)	60.5% (144)
Reddit	61.4% (189)	61.8% (147)
Pinterest	39.3% (121)	36.1% (86)
Other (open text): including Discord, LinkedIn, Quora, Mastodon, Next door, Truth, and Twitch.	3.9% (12)	4.6% (11)
In the last 7 days, how many days did you use Instagram?		
Never	0.3% (1)	0.0% (0)
1 day	4.2% (13)	3.4% (8)
2 days	10.4% (32)	9.2% (22)
3 days	10.7% (33)	11.3% (27)
4 days	8.4% (26)	12.2% (29)
5 days	11.0% (34)	6.3% (15)
6 days	2.6% (6)	4.6% (11)
Everyday	52.3% (161)	52.9% (126)
In the last 30 days, on an average day how many hours did you use Instagram?		
Less than half an hour	29.2% (90)	28.6% (68)
1 hour	38.3% (118)	35.7% (85)
2-3 hours	22.7% (70)	25.2% (60)
4-5 hours	4.9% (15)	5.5% (13)
More than 6 hours	4.9% (15)	5.0% (12)

Table S3.3

Study 3a: Descriptive and Inferential Statistics for How Many Days of the Last 7 Days and How Many Hours on Average Each Day (Over the Last 30 Days) Instagram was Used by Condition

Question	No Description	Brief	Detailed			
	M (SD)	*M (SD)*	*M (SD)*	*F*	*df*	*p*
Days	6.2 (2.1)	6.4 (1.9)	6.2 (2.2)	0.54	2, 305	.59
Hours	2.0 (0.8)	2.4 (1.3)	2.2 (1.0)	2.92	2, 305	.06

Note. Days: 6.0 = 5 days and 7.0 = 6 days. Hours: 2.0 = 1 hours and 3.0 = 2-3 hours. We note possible limitations of ordinal data.

Table S3.4

Study 3a: Percentage of Participants Who Indicated They Have Seen a Sensitive-Content Screen Before by Condition

Condition	Seen Screens Before % (*n*)				
	Yes	No	χ^2	*df*	*p*
No Description	70.5% (74)	29.5% (31)			
Brief	70.9% (73)	29.1% (30)	.373	2	.83
Detailed	74.0% (74)	26.0% (26)			

Table S3.5

Study 3a: Descriptive and Inferential Statistics for Gender by Condition

Condition	Gender % (n)				χ^2	*df*	*p*
	Man	Woman	Non-binary	Not reported			
No Description	49.5% (52)	46.7% (49)	1.9% (2)	1.9% (2)	11.17	6	.08
Brief	35.0% (36)	65.0% (67)	0.0% (0)	0.0% (0)			
Detailed	35.0% (36)	63.0% (63)	1.0% (1)	1.0% (1)			

Table S3.6

Study 3a: Descriptive and Inferential Statistics for Age, Income Level and Education Level by Condition

Question	No Description	Brief	Detailed			
	M (SD)	*M (SD)*	*M (SD)*	*F*	*df*	*p*
Age	40.0 (11.2)	37.0 (9.0)	39.6 (10.4)	2.66	2, 305	.07
Income	2.9 (1.1)	2.8 (1.2)	2.8 (1.2)	0.28	2, 305	.76
Education	3.5 (0.8)	3.5 (0.7)	3.5 (0.7)	0.02	2, 305	.98

Note. Income: 2.0 = $20,000-$45,000 and 3.0 = $45,000-$140,000. Education: 3.0 = Some college and 4.0 = College graduate. We note possible limitations of ordinal data for Income and Education.

Table S3.7

Study 3a: Descriptive and Inferential Statistics for Imagination, Uncertainty and Curiosity by Condition

Question	No Description	Brief	Detailed			
	M (SD)	*M (SD)*	*M (SD)*	*F*	*df*	*p*
Imagination	2.9 (1.0)	3.0 (1.1)	3.1 (0.9)	0.97	2, 305	.38
Uncertainty	2.7 (1.2)	2.4 (1.2)	2.5 (1.2)	1.39	2, 305	.09
Curiosity	3.0 (1.3)	2.8 (1.3)	2.7 (1.3)	1.24	2, 305	.29

Note. Imagination: 1 = *perfectly clear and as vivid as normal vision* to 5 = *no image at all*; Uncertainty: 1 = *not at all uncertain* to 5 = *extremely uncertain*; Curiosity: 1 = *not at all curious* to 5 = *extremely curious*.

Table S3.8

Study 3b: Descriptive and Inferential Statistics for How Many Days of the Last 7 Days and How Many Hours on Average Each Day (Over the Last 30 Days) Instagram was Used by Experimental Condition

Question	No Description	Brief			
	M (SD)	*M (SD)*	*t*	*df*	*p*
Days	6.2 (2.0)	6.5 (2.0)	-1.27	236	.20
Hours	1.9 (0.9)	2.5 (1.2)	-3.61	228.9	<.001

Note. Days: 6.0 = 5 days and 7.0 = 6 days. Hours: 2.0 = 1 hours and 3.0 = 2-3 hours. We note possible limitations of ordinal data.

Table S3.9

Study 3b: Percentage of Participants Who Indicated They Have Seen a Sensitive-Content Screen Before by Experimental Condition

Condition	Seen Screens Before % (*n*)		χ^2	*df*	*p*
	Yes	No			
No Description	79.1% (91)	20.9% (24)	1.84	1	.18
Brief	71.5% (88)	28.5% (35)			

Table S3.10

Study 3b: Descriptive and Inferential Statistics for Gender by Experimental Condition

Condition	Gender % (n)			χ^2	*df*	*p*
	Man	Woman	Non-binary			
No Description	34.8% (40)	65.2% (75)	0.0% (0)	5.22	2	.07
Brief	23.6% (29)	74.8% (92)	1.6% (2)			

Table S3.11

Study 3b: Descriptive and Inferential Statistics for Age, Income Level and Education Level by Experimental Condition

Question	No Description	Brief	*t*	*df*	*p*
	M (SD)	*M (SD)*			
Age	35.4 (8.7)	35.8 (9.0)	-.346	236	.73
Income	2.7 (0.9)	2.8 (1.0)	-.135	236	.89
Education	3.5 (0.7)	3.5 (0.6)	-.394	236	.69

Note. Income: 2.0 = $20,000-$45,000 and 3.0 = $45,000-$140,000. Education: 3.0 = Some college and 4.0 = College graduate. We note possible limitations of ordinal data for Income and Education.

Table S3.12

Study 3b: Descriptive and Inferential Statistics for How Many Days of the Last 7 Days and How Many Hours on Average Each Day (Over the Last 30 Days) Instagram was Used by Quasi-Experimental Condition

Question	No Description Uncovered	Brief Uncovered			
	M (SD)	*M (SD)*	*t*	*df*	*p*
Days	6.3 (2.0)	6.4 (2.0)	-.231	199	.81
Hours	2.0 (0.9)	2.5 (1.1)	-3.25	192.6	<.001

Note. Days: 6.0 = 5 days and 7.0 = 6 days. Hours: 2.0 = 1 hours and 3.0 = 2-3 hours. We note possible limitations of ordinal data.

Table S3.13

Study 3b: Percentage of Participants Who Indicated They Have Seen a Sensitive-Content Screen Before by Quasi-Experimental Condition

Condition	Seen Screens Before % (*n*)				
	Yes	No	χ^2	*df*	*p*
No Description Uncovered	83.0% (83)	17.0% (17)	5.88	1	.015
Brief Uncovered	68.3% (69)	31.7% (32)			

Table S3.14

Study 3b: Descriptive and Inferential Statistics for Gender by Quasi-Experimental Condition

Condition	Gender % (n)			χ^2	*df*	*p*
	Man	Woman	Non-binary			
No Description Uncovered	37.0% (37)	63.0% (63)	0.0% (0)	4.98	2	.08
Brief Uncovered	23.8% (24)	75.2% (76)	1.0% (1)			

Table S3.15

Study 3b: Descriptive and Inferential Statistics for Age, Income Level and Education Level by Quasi-Experimental Condition

Question	No Description Uncovered	Brief Uncovered			
	M (SD)	*M (SD)*	*t*	*df*	*p*
Age	35.0 (8.8)	35.8 (9.0)	-.615	199	.54
Income	2.7 (0.9)	2.7 (0.9)	.06	199	.96
Education	3.4 (0.7)	3.4 (0.7)	.145	199	.89

Note. Income: 2.0 = $20,000-$45,000 and 3.0 = $45,000-$140,000. Education: 3.0 = Some college and 4.0 = College graduate. We note possible limitations of ordinal data for Income and Education.

Study 3a: Negative Affect

We also pre-registered running our main analyses for negative affect. Recall, we were first interested in whether the *presence* of content related information on sensitive-content screens would have an emotional cost, relative to sensitive-content screens without content related information. Therefore, we ran a 2 (condition: no content descriptions, content descriptions) x 2 (time: T1, T2) mixed ANOVA on the negative affect scale of the PANAS (Table S17). Overall, negative affect was higher (i.e., more negative) at T2 compared with T1; a main effect of time for negative affect, $F(1, 306) = 47.45$, $p < .001$, $\eta_p^2 = .13$. There was *no* difference in negative affect between participants who saw sensitive-content screens with vs without content descriptions; a nonsignificant main effect of condition for negative affect, $F(1, 306) = 0.06$, $p = .81$, $\eta_p^2 = .0001$. The effect of time on negative affect did not depend on whether participants saw sensitive-content screens with or without content descriptions; a nonsignificant interaction between condition and time for negative affect, $F(1, 306) = .82$, $p = .36$, $\eta_p^2 = .003$. Therefore, participants who saw sensitive-content screens with content descriptions showed similar increases in negative affect (from T1 to T2) compared with participants who saw sensitive-content screens without content descriptions.

Recall, we were also interested in whether the *level* of content related information on sensitive-content screens would influence the emotional cost. Therefore, we ran a second 2 (condition: brief content description, detailed content description) x 2 (time: T1, T2) mixed ANOVA on the negative affect scale of the PANAS (Table S17). As with our first ANOVA, there was a main effect of time, $F(1, 306) = 40.41$, $p < .001$, $\eta_p^2 = .17$, and a nonsignificant main effect of condition, $F(1, 201) = 0.004$, $p = .95$, $\eta_p^2 = .00002$, for negative affect; the pattern of the data was the same. The effect of time on negative affect depended on whether participants saw sensitive-content screens with brief vs detailed content descriptions; a significant interaction between condition and time for negative affect, $F(1, 201) = 4.01$, $p =$

.047, η_p^2 = .02. Specifically, participants who saw sensitive-content screens with detailed content descriptions showed larger increases in negative affect (from T1 to T2) compared with participants who saw sensitive-content screens with brief content descriptions.

Table S3.17

Study 3a: Means (and Standard Deviations) for Negative Affect by Condition and Time

	Time		
Condition	Pre-task *M* (*SD*)	Post-task *M* (*SD*)	Total *M* (*SD*)
No Content Description	13.6 (6.2)	15.2 (6.6)	14.4 (6.4)
Content Description	13.5 (6.5)	15.6 (7.5)	14.6 (6.4)
Total	13.5 (6.4)	15.5 (7.2)	
Brief Content Description	13.8 (6.6)	15.2 (7.5)	14.5 (6.6)
Detailed Content Description	13.2 (6.4)	16.0 (7.5)	14.6 (6.6)
Total	13.5 (6.5)	15.6 (7.5)	

Note. Possible scores for Negative Affect range from 10 to 50.

Study 3a: Planned Sensitivity Analysis

We allowed participants to view the negative image (once they decided to uncover it), for as long as they liked—to increase the ecological validity of our design. To test whether the time participants spent viewing the negative image affected our results, we ran a follow up hierarchical multiple regression controlling for viewing time, while testing the effect of content description condition on participants' distress. In the same analysis we also controlled for Instagram use and previous exposure to sensitive-content screens—as indicated in our preliminary analyses and per our pre-registration. We entered viewing time (b = -.52, p = .336), Instagram use (b = -4.29, p = .031), and previous exposure to sensitive-content screens

(b = -.42, p = .933) in Step 1; together, viewing time, Instagram use, and previous exposure to sensitive-content screens did not explain any (2.8%) variance in participants' distress, R^2 = .028, $F(3, 197) = 1.92$, p = .128. In Step 2, we entered content description condition (b = .23, p = .958); content description condition also did not explain any variance in participants' distress, $R^2change$ = 0.00001, $Fchange(1, 196) = 0.003$, p = .958. Thus, neither the combined effect of the time participants spent viewing the negative image, Instagram use, and previous exposure to sensitive-content screens, *or* content description condition contributed to participants' distress.

Table S3.18

Study 3b: General Task Compliance Questions

Variable	Yes: % (*n*)	No: % (*n*)
Left image viewing task	0.0% (0)	100% (192)
Technical issues	1.6% (5)	98.4% (303)

6 Investigating Whether Adding Cognitive Emotion Regulation Instructions to Sensitive-Content Screens Reduces Distress

Chapter 6 is submitted for publication:

Simister, E. T., Moeck, E. K., Bridgland, V. M. E., & Takarangi, M. K. T. (2024). Including cognitive emotion regulation instructions on sensitive-content screens reduces distress.

Authors Contributions: I developed the study design with the guidance of MKTT, EKM, and VMEB. I collected the data, performed the data analysis and interpretation (with assistance from EKM who developed the R Script and created the visualisation), and drafted the manuscript. MKTT, EKM, and VMEB contributed equally by making critical revisions to the manuscript. All authors approved the final version of the manuscript for submission.

Abstract

Sensitive-content screens do not reduce people's negative reactions to distressing social media content, perhaps because these screens do not help people emotionally prepare. Across two studies, we examined whether adding cognitive emotion regulation instructions to sensitive-content screens improves their efficacy. In Study 4a, we trained participants to use distraction and reappraisal then showed them negative images, each preceded by a sensitive-content screen with reappraisal, distraction, or no instructions (within-subjects). After each image, participants rated distress: participants reported lower distress when they received reappraisal or distraction instructions, compared to no instructions. In Study 4b, we varied the method by randomly allocating participants to a distraction *or* no instruction condition: participants who received distraction instructions reported lower distress than participants who received no instructions. Therefore, sensitive-content screens in their current format do not help people spontaneously engage in emotion regulation, but cognitive emotion regulation instructions can make these screens more effective.

Introduction

Instagram—and social media platforms alike (e.g., Facebook and TikTok)—use sensitive-content screens, a form of *trigger warning* (Bridgland et al., 2023), to deter people from engaging with potentially distressing content (e.g., images that depict self-harm) and to mitigate negative emotional reactions (e.g., state anxiety) to content once it is viewed. However, people tend to engage with such content despite the presence of these screens. This behaviour fits with research showing people are unlikely to avoid content following traditional trigger warnings (see Bridgland et al., 2023), likely because they find restricted content attractive ("forbidden fruit effect"; Weaver, 2011) and seek to resolve curiosity even when they expect negative consequences (e.g., electric shocks; "pandora effect"; Hsee & Ruan, 2016). Sensitive-content screens also do not mitigate negative emotional reactions to content once it is viewed (Takarangi et al., 2023)—perhaps because these screens fail to help people emotionally prepare (Bridgland, Barnard et al., 2022). Therefore, sensitive-content screens in their current form may be an inadequate harm-minimisation tool. Here, we investigate whether adding cognitive emotion regulation instructions to sensitive-content screens improves their efficacy. Specifically, we examine whether providing cognitive emotion regulation instructions—specifically, for distraction and reappraisal—on sensitive-content screens reduces people's distress following exposure to negative images, relative to no instruction. This research has implications for sensitive-content screens in their current format and provides a potential solution to improve the screens' efficacy as a harm-minimisation tool.

Advocates claim that trigger warnings, including sensitive-content screens, are beneficial because—among other things, such as increasing avoidance of content—they help people "emotionally prepare" for upcoming content and mitigate negative reactions (e.g., Lockhart, 2016). However, in Takarangi et al., (2023; Experiment 3) participants who saw

sensitive-content screens prior to negative images experienced similar changes in state anxiety and negative affect compared to participants who saw negative images without preceding screens. Furthermore, trigger warnings do not increase the time people spend "preparing" for distressing imagery (Bridgland & Takarangi, 2022), nor do they prompt emotion regulation strategies (e.g., to focus on non-emotional content) to come to people's minds (Bridgland, Barnard et al., 2022). These results are not surprising when we consider that sensitive-content screens were designed assuming people will *spontaneously* implement strategies to help manage their emotions after seeing such screens. Yet, some people may not recognise the need to use such emotion regulation strategies or, may have limited strategies to draw on (Gross, 2015). Furthermore, even *if* people know some emotion regulation strategies, they may not implement them, potentially because they doubt their capacity to do so (Gross, 2015). Therefore, sensitive-content screens may not help people "emotionally prepare" because they seemingly fail to address the challenges people have in selecting and implementing emotion regulation strategies.

To help people "emotionally prepare" for upcoming content then, sensitive-content screens could explain *how* to do so. Research suggests *cognitive* emotion regulation strategies—such as distraction, which involves directing attention away from emotionally salient aspects of situations or away from situations altogether, and reappraisal, which involves reinterpreting the meaning of situations (e.g., "my racing heart is not anxiety; it is me preparing to perform"; Gross, 2015)—can reduce negative emotions, alleviate psychological symptoms, and improve well-being (Kraiss et al., 2020; Webb et al., 2012). For example, participants given cognitive emotion regulation instructions—including distraction and reappraisal—reported lower negative emotions when viewing negative images (e.g., Ray et al., 2010; Thiruchselvam et al., 2011) and films (e.g., Wolgast et al., 2011) compared to when they were not given instructions/were asked to respond naturally.

Therefore, we might expect a similar pattern of results to emerge for negative images in a social media context.

Although both distraction and reappraisal can reduce negative emotions, research on cognitive emotion regulation suggests distraction may be more effective in a social media context. Distraction does not require people to attend or provide meaning to incoming emotional information (e.g., distressing aspects of situations), it is easy to implement, and facilitates immediate *short-term* relief from negative affect (Thiruchselvam et al., 2011; see also Sheppes & Gross, 2011). In contrast, reappraisal requires people to first attend and provide meaning to incoming emotional information *before* they can then reinterpret it (Thiruchselvam et al., 2011). Thus, reappraisal is more effortful and takes longer to implement than distraction (Sheppes & Gross, 2011), but facilitates *longer-term* emotional benefits (e.g., reductions in psychological symptoms; Kraiss et al., 2020). Consistent with the effort and cognitive resources required for each strategy, people choose to use distraction (over reappraisal), particularly for high intensity stressors (e.g., viewing distressing images; see Sheppes et al., 2011). Therefore, distraction may be more effective than reappraisal in the short-term and is the preferred strategy for high intensity stressors such as viewing graphic images.

Here, our primary aim was to examine the short-term effectiveness of adding cognitive emotion regulation instructions to sensitive-content screens. Examining this aim has implications for sensitive-content screens in their current format, and *if* we find adding instructions to screens mitigates negative emotional reactions to content once it is viewed, this research will provide a potential solution to improve the screens' efficacy as a harm-minimisation tool. Specifically, we examined whether providing distraction (Studies 4a and 4b) and reappraisal (Study 4a only) instructions on sensitive-content screens would reduce participants' distress following exposure to negative images, relative to screens as they

typically appear on Instagram (i.e., without instructions). In Study 4a, we predicted participants would report lower distress after images for which we instructed them to use reappraisal (Hypothesis 1) and distraction (Hypothesis 2) compared to images without regulation instructions. Furthermore, because we focused on short-term effectiveness, we predicted participants would report the *lowest* distress after images for which we instructed them to use distraction (Hypothesis 3).

Transparency and Openness

The Flinders University Human Research Ethics Committee approved this research, and we pre-registered the design, hypotheses, and analysis plan on the Open Science Framework (OSF; Study 4a: https://osf.io/mhtrf; Study 4b: https://osf.io/ap8y). We programmed the studies in Qualtrics. We report how we determined our sample sie, all data exclusions, all manipulations, and all measures in the study. The supplementary materials are at the end of the chapter and the data, including a codebook describing all variables, can be found at: https://osf.io/ah9qd/.

Study 4a

Method

Participants

Our desired sample was 191 participants, determined by a priori power analysis for a small within-person (Level-1) predictor effect, with an alpha of 0.05, power of .80, and effect size of $t = 2.50$ (the largest sample size we could achieve with the resources available for this study; Murayama et al. 2022). We recruited participants (in 2022) from the United States using Amazon's Mechanical Turk (MTurk) via CloudResearch (with the approved participants setting). To ensure data quality/minimise bots/server farmers, we screened out participants who failed a captcha, scored less than 8/10 on an English proficiency test (Moeck et al., 2022), selected "Konnect" (a bogus platform included to detect inattentive responses)

when asked about social media use, and/or indicated they do not use Instagram (because we only wanted to recruit Instagram users). Of 220 participants who completed the survey and received a payment of $2.00 USD, we excluded 28 per our pre-registration: 12 did not follow instructions (e.g., used both strategies simultaneously/only used one strategy); six did not demonstrate comprehension of training trials; three experienced technical issues that interfered with task completion; two failed two embedded attention checks; two reported not reading the regulation instructions (without a valid reason; e.g., because they did not pay attention); two reported leaving during the image task; and one did not pass the cultural check (we showed participants a picture of an eggplant and asked them what it is called [we expected participants from the United States to answer: "eggplant"]).

Our final sample of 192 participants, aged 19-71 years (M = 35.5, SD = 9.0) included 59.4% women (n = 114), and 37.0% men (n = 71); 2.6% of participants identified as non-binary (n = 5), and 1.0% preferred not to report gender (n = 2). Our sample was predominantly European American/White (65.1%); other participants were of African American/Black (10.4%), Hispanic/Latinx (6.8%), Asian (3.6%), or multiracial (9.9%) descent; 4.2% of participants specified nationality (e.g., American/USA) when given the option to self-describe their ethnicity. Most participants (54.2%) reported an income between $45,000-$140,000 and were predominantly (55.2%) college graduates (Supplementary Table S4.1). Most participants (49.5%) reported they had used Instagram every day over the past week, and for one hour or more on an average day in the last 30 days (66.7%; Supplementary Table S4.2) and reported they have seen sensitive-content screens on their own Instagram feed (71.9%). Most participants also reported using YouTube (91.1%), Facebook (85.4%), Reddit (75.0%), Twitter (71.4%), and TikTok (55.7%) on a regular basis (Supplementary Table S4.2).

Materials/Measures

Image Stimuli.

We selected the 30 most negative images from the Nencki Affective Picture System (NAPS; Marchewka et al., 2014; based on normative valence ratings: 1 = *very negative* to 9 = *very positive;* arousal ratings are also available: 1= *relaxed* to 9 = *aroused*). The content of images (i.e., people, animals, objects) are commonly found on Instagram and would likely meet the threshold for Instagram to screen them (e.g., the negative images include people/animals that have been injured/are deceased). All images were *high* in emotional intensity (current study: valence $M = 1.87$, $SD = 0.19$, arousal $M = 7.35$, $SD = 0.37$; comparable to high intensity images in previous studies, e.g., Sheppes et al., 2011). We placed all images in an Instagram border with non-functional like and comment buttons. Consistent with Instagram's sensitive-content screen format (as of November 2022), before each negative image was a blurred version of that image accompanied by a warning ("Sensitive Content: This photo may contain graphic or violent content"; Figure 4.1a).

Cognitive Emotion Regulation Instructions.

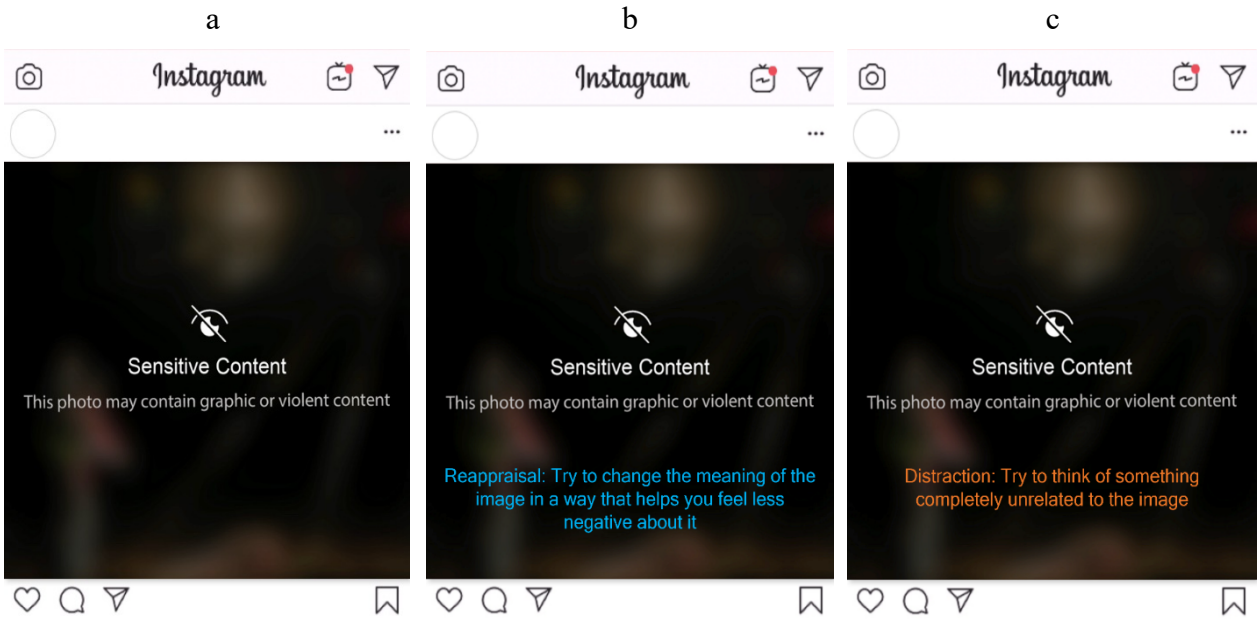
In addition to the warning, we included cognitive emotion regulation instructions on some sensitive-content screens. Thus, we manipulated instruction type within-subjects. A third of screens included a reappraisal instruction (i.e., "Reappraisal: Try to change the meaning of the image in a way that helps you feel less negative about it"), a third included a distraction instruction (i.e., "Distraction: Try to think of something completely unrelated to the image"), and the final third had no instruction. For screens without instructions, we told participants in the task preamble to respond naturally to the image (Figure 4.1a.). Screens without instructions were intended to elicit natural emotional responses (Webb et al., 2012).

To help participants quickly differentiate and switch effectively between different strategies, the reappraisal and distraction instructions had unique text color (Figure 4.1b-

4.1c). To control for possible color associations, text color was counterbalanced across participants: half the participants saw reappraisal in blue and distraction in orange, and half the participants saw reappraisal in orange and distraction in blue. We created three sets of 10 negative images matched on valence and arousal ratings and counterbalanced them across participants (set 1: valence $M = 1.84$, $SD = 0.22$, arousal $M = 7.30$, $SD = 0.46$; set 2: valence $M = 1.88$, $SD = 0.17$, arousal $M = 7.43$, $SD = 0.19$; set 3: valence $M = 1.90$, $SD = 0.16$, arousal $M = 7.33$, $SD = 0.40$; see supplementary materials for NAPS image codes).

Figure 4.1

Example NAPS Images Modified to Look Like Instagram Images with a Sensitive Content Overlay and (a) No Instruction to Regulate, (b) Instruction to Use Reappraisal, and (c) Instruction to Use Distraction



Procedure

We told participants we were collecting information about social media engagement. After providing informed consent, all participants completed Instagram use questions, and items designed to reduce suspicion about the true nature of our study: like in our previous

studies[29] (e.g., Simister, Bridgland & Takarangi et al., 2023; Studies 1a and 1b) participants rated how often they view different types of images (e.g., portraits) on Instagram. We then introduced participants to the image task (adapted from Sheppes et al., 2011). Participants read descriptions about reappraisal and distraction (see supplementary materials), before completing two training trials (one for each strategy; in a counterbalanced order). For each training trial, participants viewed a sensitive-content screen (with either a reappraisal or distraction instruction) for 5s. After 5s, the negative image (previously covered by the sensitive-content screen) appeared on the screen. The negative image remained on the screen for 5s; we instructed participants to keep their eyes on the image, and to avoid diverting their gaze, while using the specific strategy indicated on the preceding sensitive-content screen. Immediately after the negative image disappeared, participants responded to the following question: "*How distressed do you feel right now?*" (0 = *not at all distressed*, to 100 = *extremely distressed*). Participants also responded to a multiple-choice question regarding how they had used the (reappraisal or distraction) strategy while viewing the negative image. This question served as a comprehension check: if participants responded accurately (i.e., they said they "*tried to think about what was happening in the image in a new way, so that I felt less negative*" for reappraisal, and "*tried to think about something else, unrelated to what was happening in the image, so that I felt less negative*" for distraction), they moved onto the next training trial, or onto the main task once they responded accurately to both training trials. If participants responded inaccurately, they were given the strategy instruction(s) again, before repeating the trial(s). Participants who responded inaccurately on their second trial attempt (for either strategy; $n = 6$) were exited from the study at this point. Participants who demonstrated comprehension of the task proceeded to the main image task.

[29] Participants from our previous studies were not eligible for the present study.

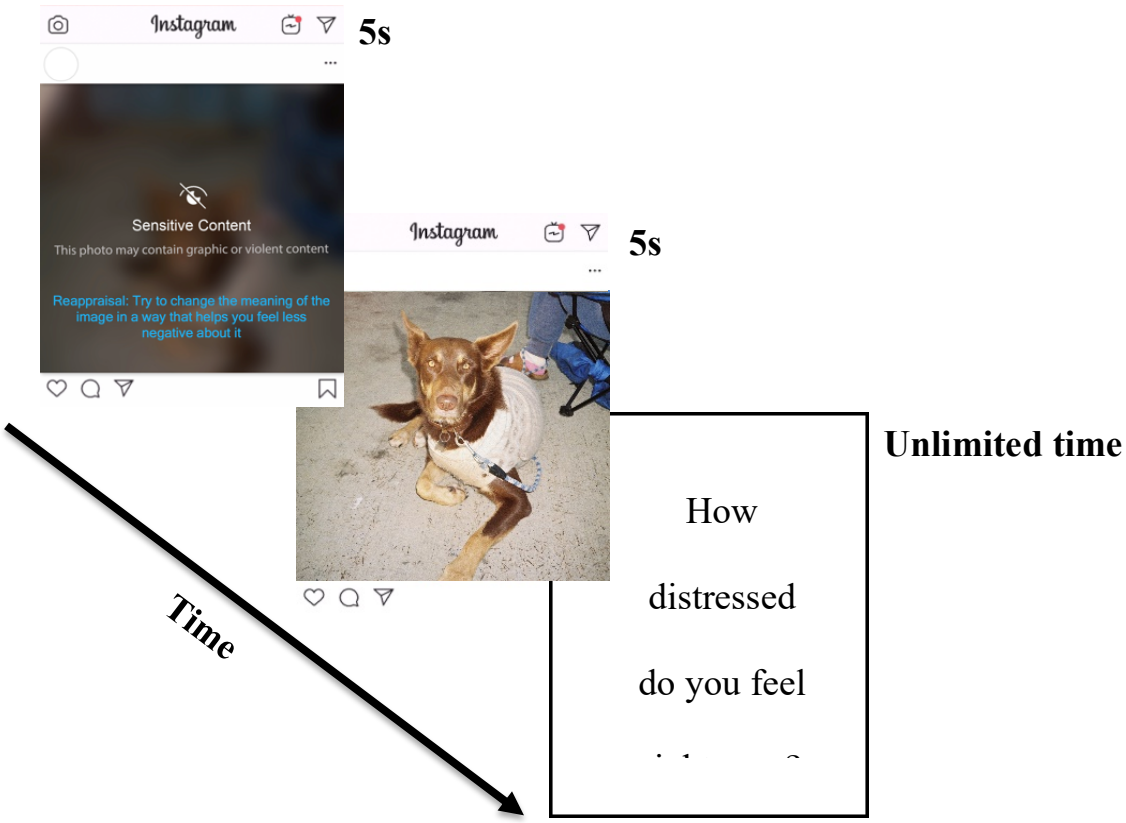
The procedure for the main image task largely matched the training trials: participants viewed sensitive-content screens (one at a time) for 5s, and then the negative images underlying each screen for 5s, while using the specific strategy indicated on the preceding sensitive-content screen (Figure 4.2). Previous research using a within-subjects design with similar trial timing has demonstrated that people *can* switch between regulatory strategies when instructed (as evidenced by different patterns of EEG data; see Thiruchselvam et al., 2011). After, participants rated their current level of distress in the same way as the training task, before proceeding to the next trial (30 trials total). Unlike training trials, there were no comprehension questions in the image task, which also included sensitive-content screens without regulation instructions. We found excellent reliability for distress ratings in all conditions (reappraisal: α = .93, distraction: α = .93, no instructions: α = .92), suggesting 10 trials per condition was sufficient.

After the task, participants rated how easy it was to use distraction and reappraisal (0 = *not at all easy*, to 100 = *extremely easy*) and the (perceived) effectiveness of each strategy (0 = *not at all effective*, to 100 = *extremely effective*). Participants also indicated which strategy they would implement if they were asked to do the task again (distraction/reappraisal/other/none) and how often they use each strategy (or other strategies) when they see negative content in their everyday lives (0 = *never* to 4 = *always*). We also asked participants if they have seen sensitive screens before (yes/no), and their preferences for sensitive content (more/standard/less) on their own Instagram feeds. Then, to detect response quality we asked participants to indicate: if they always read (yes/no) and followed (yes/no) strategy instructions, what they did while viewing images without regulation instructions, if they looked away from any negative images during the task (yes/no; and why), if they stopped the task for any extensive period (and when/for how long), or if they

experienced any technical issues. Finally, participants completed demographics. We then fully debriefed participants.

Figure 4.2

Trial Structure for the Main Image Task (an Example of a Reappraisal Trial)



Note. The example sensitive-content screen and unscreened image is a neutral photo from the authors' own collection.

Statistical Analyses

We used R (version 4.1.1) to run linear mixed effect models with the *lme4* package (Bates et al., 2015), and tested statistical significance of model parameters using *lmerTest* (Kuznetsova et al., 2017). We used two-level models, with trials (Level 1) nested within participants (Level 2). We included random intercepts and slopes for all Level-1 predictors.

In our main models, we tested whether regulation instruction condition predicted distress. Because instruction condition is a categorical variable, the intercept for each model represents the mean of the reference category, which varied depending on which hypothesis we tested (as outlined below). We then re-ran the models including person-mean centered distress from the previous trial (lagged distress) as a Level-1 covariate. This covariate allowed us to model change in distress for each condition, over and above persistence in distress across successive measurement occasions. To improve model convergence, we used the "bobyqa" optimizer and up to 250,000 iterations (Bates et al., 2015) for all analyses. We dealt with nonconverging models (1 and 2) by simplifying the random effects structure until convergence was reached (but ensured the findings did not change relative to the more complex structure; see analysis code on OSF: https://osf.io/5y49n).

Results and Discussion

Hypothesis Testing

Overall, participants' distress ratings were moderate (M = 42.04, SD_{within} = 18.32, $SD_{between}$ = 21.10) and varied similarly between as within person (ICC = .53). We first tested H1 and H2 by comparing each regulation strategy condition (reappraisal, distraction) to no instruction, by setting "no instruction" as the reference category (Table 4.1; Model 1). Consistent with H1, participants reported lower distress after images where we instructed them to use reappraisal, compared to images without regulation instructions. The model showed the same pattern for distraction, consistent with H2. However, both effects were small: relative to participants' distress rating for images without regulation instructions (intercept), there was, on average, an estimated decrease of 4-points for reappraisal and 2-points for distraction (Figure 4.3). Notably, the effects of reappraisal and distraction remained significant (and small) when we re-ran the model controlling for participants' distress on the

previous trial (Supplementary Table S4.3), meaning the reduction in participants' distress did not simply carry over from the previous trial.

Table 4.1

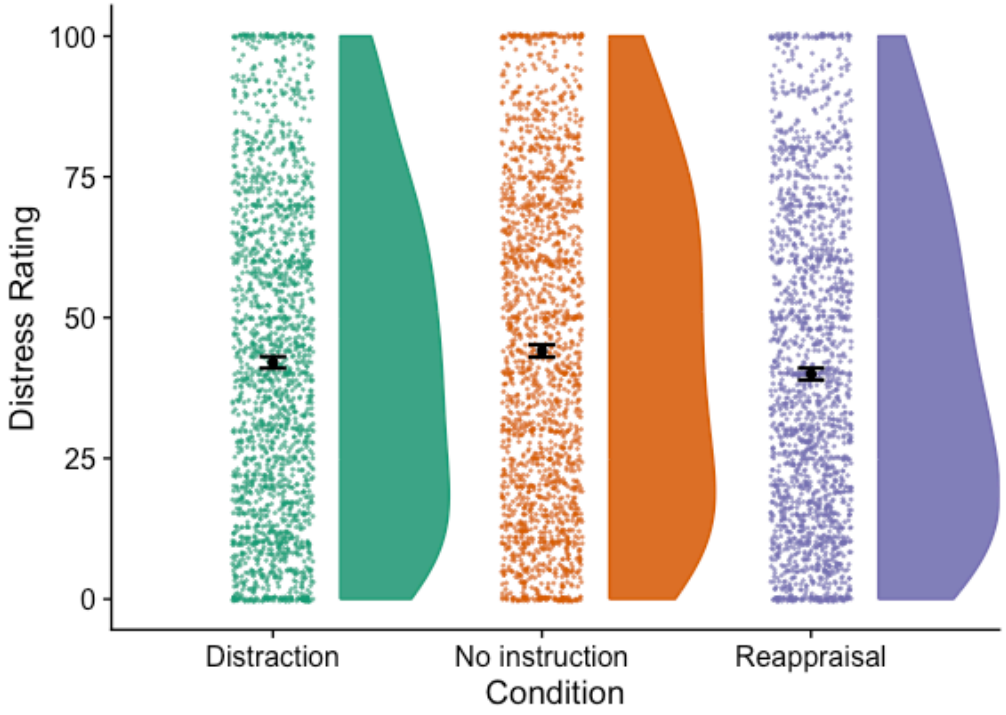
Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Condition on Distress Ratings

		Distress Rating		
	Predictor	*Estimate (SE)*	*95% CI*	*p value*
Model 1	(Intercept)	44.09 (1.59)	40.96 – 47.21	<.001
	Distraction	-2.04 (0.68)	-3.37 – -0.70	.003
	Reappraisal	-4.10 (0.78)	-5.64 – -2.57	< .001
Model 2	(Intercept)	42.05 (1.61)	38.88 – 45.22	<.001
	No Instruction	2.04 (0.68)	0.70 – 3.37	.003
	Reappraisal	-2.07 (0.74)	-3.53 – -0.61	.006

Note. The intercept for Model 1 represents the mean for the no regulation instruction condition, and the intercept for Model 2 represents the mean for the distraction condition. The condition estimates for each model indicate the difference relative to the intercept. $\alpha = .05$.

Figure 4.3

Distress Rating Estimates by Cognitive Emotion Regulation Condition



Note. The scatterplots on the left show the raw data and the density plots on the right show the distribution of the data. The black dot is the mean level of the outcome variable. The error bars represent the 95% CI around the mean. We measured distress on a 100-point distress scale (0 = *not at all distressed*, to 100 = *extremely distressed*).

We next examined whether there was a difference in participants' distress depending on *which* emotion regulation strategy they used. We re-ran the initial model with "distraction" as the reference category to compare distress between the reappraisal and distraction conditions (Table 4.1; Model 2). The model showed that participants reported a larger decrease in distress on reappraisal trials, relative to distraction trials. Therefore, contrary to H3, participants reported the *lowest* distress after images we instructed them to use reappraisal rather than distraction. However, the effect was small: there was a 2-point difference between the reappraisal and distraction conditions. Notably, the effect of

reappraisal remained significant (and small) when we re-ran the model controlling for participants' distress on the previous trial (Supplementary Table S4.4).

We also examined participants' responses to post-task questions. Participants perceived both reappraisal (M = 42.6, SD = 27.5) and distraction (M = 44.3, SD = 28.2) to be moderately effective in minimising their distress, t(191) = -0.73, p = .469, d = 0.06. Yet, more participants (50.0%; n = 96) reported they would use distraction (vs. reappraisal; 31.8%; n = 61) if they did the task again and could only use one strategy, $\chi^2(1) = 7.80$, p = .005.

Planned Sensitivity Analyses

Given the graphic nature of the images, participants may have looked away from images during the task. To test whether looking away affected our results, we re-ran our main analyses (per our pre-registration) excluding the sub-sample of participants (31.8%; n = 61) who reported looking away from *some* negative images during the task—though we cannot determine which images participants looked away from. We set "no instruction" as the reference category (Supplementary Table S4.5), then re-ran the analyses with "distraction" as the reference category (Supplementary Table S4.6). Overall, the results were consistent with our main analyses: the effects of reappraisal and distraction remained significant and small.

Perhaps the effects of distraction and reappraisal were small because participants used a regulation strategy on the trials without instructions, despite being told to respond naturally, *or* that for these participants, responding naturally involved using a regulation strategy. Although most participants (70.3%; n = 135) reported that they responded naturally (per our instructions), 8.3% (n = 16) reported using reappraisal, 18.8% (n = 36) reported using distraction, and 2.6% (n = 5) reported they did something else (e.g., "I braced myself to see something unpleasant"; Supplementary Table S4.7). Therefore, we re-ran our main analyses with the sub-sample of participants who reported they responded naturally (n = 135; note we

did not pre-register these analyses). We set "no instruction" as the reference category (Supplementary Table S4.8), then re-ran the analyses with "distraction" as the reference category (Supplementary Table S4.9). Again, the results were consistent with our main analyses.

Taken together, in Study 4a we found that participants reported lower distress after negative images where we instructed them to use distraction and reappraisal, compared to negative images without regulation instructions. Contrary to our original predictions, we also found a small difference between strategies in favour of reappraisal, despite the required effort to reappraise (Thiruchselvam et al., 2011). However, the difference is negligible when considered alongside participants' preference for using distraction rather than reappraisal in the future. This preference for distraction aligns with existing research showing that people prefer distraction for high intensity stressors (e.g., Sheppes et al., 2011), but also with evidence suggesting distraction requires less effort to implement (Sheppes & Gross, 2011). Considering (actual and perceived) effectiveness alongside participants' strategy preferences and effort requirements, distraction may be the best strategy to include on sensitive-content screens.

However, we do not know whether the differences we found between conditions (and the sizes of those effects) were due to the regulation strategies themselves, or due—in part—to three limitations of the within-subjects design. First, participants may have been *ineffective* at switching between using, and then not using, cognitive emotion regulation strategies *or*, they may have chosen to use strategies in a way other than how they were instructed (e.g., participants who liked distraction may have used distraction for most of the trials). Second, we instructed participants to respond "naturally" on trials without cognitive emotion regulation instructions, but approximately 30% of participants reported not following these instructions. Third, the training phase and presence of regulation instructions on some of the

trials may have increased participants' awareness of their emotions overall. Thus, perhaps we observed higher distress ratings for images without cognitive emotion regulation instructions because people were more aware of their emotions but did not have another "task" to do, relative to other conditions where they had a regulation "task" to engage with. We addressed these limitations by using a between-subject design in Study 4b.

Study 4b

Our primary aim was to replicate the effect of distraction (vs. no instructions) in reducing distress using a between-subjects design. Due to resource constraints, we could only replicate one experimental condition from Study 4a. We decided to focus on distraction because participants reported they preferred using distraction over reappraisal and distraction is easier to teach and implement in a short space of time (because of the relative effort and cognitive resources required for each strategy; Sheppes & Gross, 2011). Specifically, we examined whether participants who received distraction instructions on sensitive-content screens had lower distress than participants who received sensitive-content screens without instructions. In line with our findings in Study 4a, we predicted participants who received distraction instructions would report lower distress than participants who saw sensitive-content screens without emotion regulation instructions (Hypothesis 1).

Method

Participants

Our desired sample was 170 participants, determined by an a priori power analysis for a small between-person effect (Level-2 predictor), with an alpha of 0.05, power of .80, and effect size of $t = 3.00$ (calculated using the t-value from the estimate for the Level-1 difference in distress on no instruction vs. distraction trials from Study 4a; Murayama et al.

2022). We used the same recruitment and screening procedures as in Study 4a.[30] Of 180 participants who completed the survey (in 2023) and received a payment of $1.50 USD, we excluded 10 per our pre-registration: five did not demonstrate comprehension of training trials; three failed two embedded attention checks; one reported leaving during the image task; and one did not pass the cultural check.

Our final sample of 170 participants, aged 22-68 years (M = 36.5, SD = 8.65) included 68.8% women (n = 117), and 27.6% men (n = 47); 2.9% of participants identified as non-binary (n = 5), and 0.6% preferred not to report their gender (n = 1). Our sample was predominantly European American/White (60.6%); other participants were of African American/Black (10.0%), Hispanic (5.9%), Asian (4.7%), or multiracial (5.9%) descent; 10.0% of participants specified nationality (e.g., American/USA) when given the option to self-describe their ethnicity, and 1.2% preferred not to report their ethnicity. Most participants (52.4%) reported an income between $45,000-$140,000 and were predominantly (58.2%) college graduates (Supplementary Table S4.1). Most participants (61.8%) reported they had used Instagram every day over the past week, and for one hour or more on an average day in the last 30 days (78.2%; Supplementary Table S4.2) and reported they have seen sensitive-content screens on their own Instagram feed (81.2%). Most participants also reported using Facebook (90.0%), YouTube (81.8%), Reddit (63.5%), TikTok (57.6%), and Twitter (52.9%) on a regular basis (Supplementary Table S4.2).

Materials and Procedure

The cover story and the initial phase of the study were the same as in Study 4a: after providing informed consent, participants completed Instagram use questions and rated how

[30] There was one exception. We pre-registered we would recruit participants who had completed 1000+ studies with an approval rating of at least 95%, however, mid-way through data collection we allowed participants who had completed 100+ studies with an approval rating of at least 95% to complete the study because we were concerned that the pool of eligible participants was limited. We note that data quality remained high.

often they view different types of images (e.g., portraits). Next, we randomly assigned participants to either the distraction or no regulation instruction condition.

Participants in the distraction condition read about distraction and the task generally (see Appendix T), before completing a training trial. The procedure for the training trial was the same as in Study 4a, but there was only a single distraction trial (with two trial attempts). Unlike Study 4a—where unique text colour was important because participants had to quickly differentiate and switch between different strategies—distraction instructions (i.e., "Distraction: Try to think of something completely unrelated to the image") were all white to match the other text on the sensitive content screen (Figure 4.4). Participants who demonstrated comprehension of the distraction instructions (either on their first or second trial attempt) proceeded to the main image task. Participants who failed to demonstrate comprehension on their second trial attempt were exited from the survey at this point (and excluded) as in Study 4a.

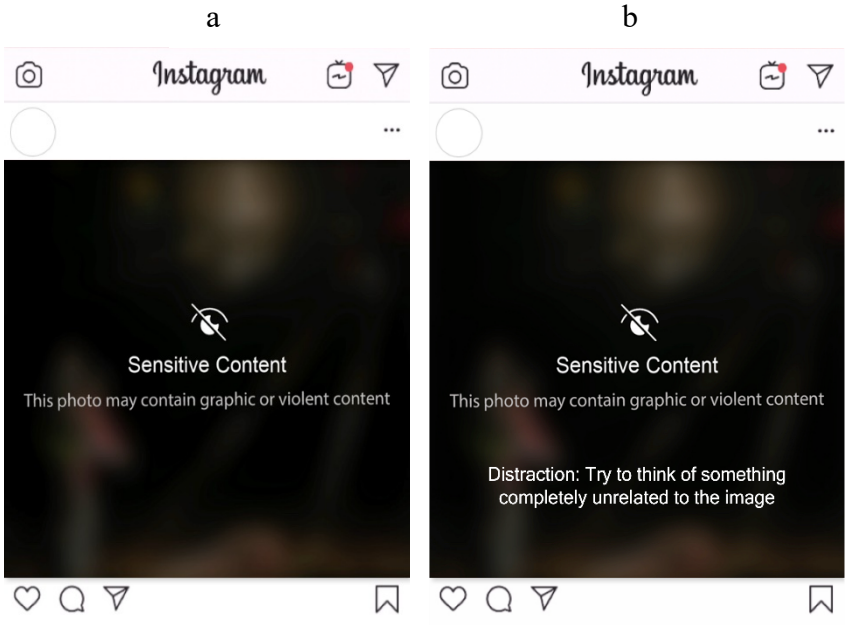
Participants in the no instruction condition read about the task generally, before completing their respective training trial—which involved viewing a sensitive-content screen for 5s, and then the negative image underlying the screen for 5s; they then rated their current level of distress, before proceeding to the main image task.

The procedure for the main image task largely matched the training trials for each respective condition: participants viewed sensitive-content screens (one at a time) for 5s, and then the negative images underlying each screen for 5s—during which participants in the distraction condition were instructed to employ distraction via the preceding sensitive-content screen. Participants rated their current level of distress before proceeding to the next trial. In total, participants viewed the 30 most negative NAPS images (as in Study 4a; Marchewka et al., 2014)—in a randomised order.

After the task, participants indicated if they have seen sensitive screens before, and their preferences for sensitive content on their own Instagram feeds. Then, participants indicated if they looked away from any negative images, if they stopped the task for any extensive period, or if they experienced any technical issues. Finally, participants completed demographics. We then fully debriefed participants.

Figure 4.4

Example NAPS Image Modified to Look Like Instagram Images with a Sensitive Content Overlay and (a) No Instruction to Regulate, and (b) Instruction to Use Distraction



Statistical Analyses

We used the same statistical approach and parameters as in Study 4a. Here, we included a random intercept for participant and random slopes for all within-person variables. In our main model, we tested whether regulation instruction condition predicted distress. The intercept for the model represents the mean of the no instruction condition.

Results and Discussion

Hypothesis Testing

Overall, participants' distress ratings were moderate (M = 45.18, SD_{within} = 17.38, $SD_{between}$ = 26.14). Distress ratings varied more between than within person (ICC = .65), likely because of our between-person manipulation. We tested H1 by comparing distraction to no instruction, setting "no instruction" as the reference category (Table 4.2; Model 1). Consistent with H1, participants who received distraction instructions reported lower distress, compared to participants who saw sensitive-content screens without instructions.[31] Unlike Study 4a, the effect was large: relative to participants who saw sensitive-content screens without emotion regulation instructions (intercept), there was, on average, an estimated decrease of 18-points for participants who received distraction instructions (Figure 4.5). Notably, the effect of distraction remained significant (and large) when we re-ran the model controlling for participants' distress on the previous trial (Supplementary Table S4.10).[32]

Planned Sensitivity Analyses

As in Study 4a, we re-ran our main analyses (per our pre-registration) excluding the sub-sample of participants (33.5%; n = 57) who reported looking away from *some* negative images during the task. Notably, the proportion of participants who reported looking away from some of the negative images during the task was similar for the distraction (34.1%; n = 29) and no instruction (32.9%; n = 28) conditions, $\chi^2(1) = 0.03$, $p = .871$. We set "no instruction" as the reference category (Supplementary Table S4.11).[33] Overall, the results were consistent with our main analyses: the effect of distraction remained significant and

[31] Distress ratings in the control condition (without instructions) compared to levels of distress reported immediately following traditional trigger warnings in the college context (also in a between-subjects design; M = 5.6; on a scale of 0-10, with higher scores indicating higher distress; Kimble et al., 2022).
[32] We dealt with nonconvergence here by rescaling the continuous variables in the model (per our pre-registration). However, because the results of the scaled and unscaled models were similar, we report the unscaled model for interpretability.
[33] We recommend the results be interpreted with caution given the reduced sample size (n = 113).

large. However, relative to the original model, the estimated distress means for participants who saw sensitive-content screens without instructions (intercept) and with distraction instructions were 5- and 2-points lower, respectively. Indeed, participants who reported looking away from negative images were, on average, more distressed (M = 53.36; SD = 25.12), than participants who did not (M = 41.04, SD = 25.88), a significant mean difference of 12.32-points, 95% CI [4.10, 20.55], $t(168) = 2.960$, $p = .002$. Such a pattern could suggest that looking away from negative images exacerbates participants' distress, or that people with an avoidant style of coping are generally more distressed. Indeed, we know that avoidance, and specifically, experiential avoidance—which refers to an unwillingness to remain in contact with distressing internal experiences as well as the attempts to control or avoid such internal experiences—can exacerbate distress (Hayes-Skelton & Eustis, 2020). However, in this case, perhaps a more parsimonious explanation is that participants looked away from negative images *because* they were distressed by the images. Indeed, when we asked these participants to report *why* they looked away from negative images, the majority explicitly referred to the disturbing (n = 11), gruesome (n = 14), or distressing (n = 6) nature of the images; others simply reported that the images were "hard to look at" (n = 13) or made them feel physically sick (n = 6; Supplementary Table S4.12).

Taken together, we found participants who received distraction instructions reported substantially lower distress than participants who saw screens without emotion regulation instructions. The effect of distraction (relative to no instruction) was larger than in Study 4a, suggesting getting people to switch between two regulation strategies and/or including no regulation trials in Study 4a may have dampened the benefit of distraction. Alternatively, perhaps assigning participants to *either* the distraction or no regulation instruction condition potentiated the effect of distraction in Study 4b. For example, one possibility is that participants in the distraction condition had no concrete sense of how distressed they would

have felt when they did not use distraction, potentially influencing their distress ratings relative to participants in Study 4a, who could compare between conditions. The same may have been true for participants in the no regulation instruction condition; having no basis for comparison may have influenced their distress rating. Our Study 4b findings showed that distress was both higher in the no instruction condition *and* lower in the distraction condition compared to Study 4a, suggesting participants may have overestimated their distress in the no instruction condition and/or underestimated their distress in the distraction condition. Although we cannot definitively say whether it was features of the within- or the between-subjects design that drove the differences between the two studies, we speculate that the true effect of distraction is likely somewhere between what we observed in Study 4a and Study 4b.

Table 4.2

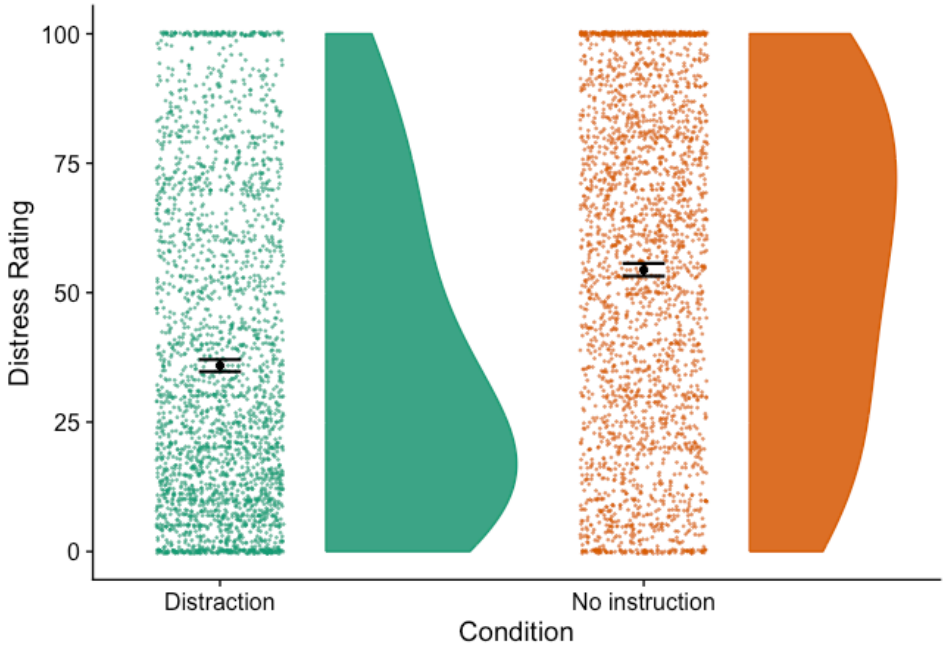
Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Conditions on Distress Ratings

		Distress Rating		
	Predictor	*Estimate (SE)*	*95% CI*	*p value*
Model 1	(Intercept)	54.43 (2.66)	49.17 – 59.70	< .001
	Distraction	-18.53 (3.77)	-25.97 – -11.08	< .001

Note. The intercept for Model 1 represents the mean for the no regulation instruction condition. The distraction estimate indicates the difference relative to the intercept. $\alpha = .05$.

Figure 4.5

Distress Rating Estimates by Cognitive Emotion Regulation Condition



Note. The scatterplots on the left show the raw data and the density plots on the right show the distribution of the data. The black dot is the mean level of the outcome variable. The error bars represent the 95% CI around the mean. We measured distress on a 100-point distress scale (0 = *not at all distressed*, to 100 = *extremely distressed*).

General Discussion

Sensitive-content screens do not mitigate negative emotional reactions to sensitive content (e.g., state anxiety; Takarangi et al., 2023). This pattern may arise because sensitive-content screens in their current format—and trigger warnings alike—do not help people emotionally prepare. Here, across two studies we examined whether putting cognitive emotion regulation instructions—specifically, for distraction and reappraisal—on sensitive-content screens could reduce people's distress following exposure to negative images, relative to screens as they typically appear on Instagram (i.e., without instructions).

Key Findings

Overall, we found that participants experienced less distress following explicit instructions to use distraction and reappraisal, versus when they received no instructions. These findings replicate prior research showing that cognitive emotion regulation instructions reduce negative emotions in response to negative images (e.g., Ray et al., 2010; Thiruchselvam et al., 2011); here, we extend these findings to a social media context. Notably, our findings were robust to various supplementary analyses. The effects of distraction and reappraisal remained significant when we controlled for participants' distress on the previous trial (Studies 4a and 4b), and when we excluded participants who reported they did not respond naturally on no instruction trials (Study 4a) or looked away from some images during the task (Studies 4a and 4b). Our findings were also robust across study design, but we observed a larger effect when participants were instructed to use one strategy the whole time (Study 4b). Hypothetically, if Instagram added emotional regulation instructions to sensitive-content screens, they could (and likely would) begin by instructing people to use one strategy; therefore, we can presume that the practical effect of adding emotional regulation instructions would be large.

We have evidence from our within-subjects design (Study 4a) to suggest that reappraisal is effective in reducing distress—relative to no instructions and also distraction—but we *may* have dampened the benefit of reappraisal (as we may have with distraction) by including another regulation task (distraction) and no regulation trials. Thus, it is possible that the effect of reappraisal would be larger using a between-subjects design, and based on Study 4a, it may even be larger than the effect we found for distraction in Study 4b. Indeed, although reappraisal is more effortful and takes longer to implement than distraction (Sheppes & Gross, 2011), reappraisal can facilitate longer-term emotional benefits (e.g., reductions in psychological symptoms; Kraiss et al., 2020), that may translate to a larger

effect in a social media context. Future research should examine reappraisal using a between-subjects design to test this possibility.

Our findings also support the idea that people do not spontaneously engage in emotion regulation following sensitive-content screens *without* emotion regulation instructions. Indeed, participants experienced the *most* distress after seeing images preceded by a warning alone—i.e., sensitive-content screens in their current format (Figure 4.4a)—even when these screens were intermixed with screens that included cognitive emotion regulation instructions in Study 4a. Broadly, this finding is consistent with research on traditional trigger warnings, which finds trigger warnings may not work as an emotion-preparation tool—perhaps because they do not prompt emotion regulation strategies (e.g., to focus on non-emotional content) to come to people's minds (Bridgland, Barnard et al., 2022). Therefore, simply alerting someone to impending negative emotions that could be caused by viewing content, via a sensitive-content screen, is not sufficient to elicit emotion regulation processes. But providing explicit instructions detailing *how* to emotionally prepare—using cognitive emotion regulation strategies—can help reduce people's distress when they view the forewarned image.

Although we aimed to use evidence-based methods to improve the *existing* sensitive-content screens, we cannot comment on whether sensitive-content screens with regulation instructions are better/worse than no screens at all. To address this issue, future research could include a no sensitive-content screen condition. But due to the cultural and legal implications around sensitive content on social media platforms (e.g., Llamas, 2023), it seems unlikely that sensitive-content screens will be abolished—meaning the best solution may be to modify and improve them with harm-minimisation in mind.

Implications

Our findings have practical implications for Instagram and social media platforms alike (e.g., TikTok). These platforms need to move beyond merely warning people about upcoming content and/or possible distressing emotional reactions and explain *how* to reduce these reactions. Users would benefit from education/training on how to implement cognitive emotion regulation strategies (like our participants received), *and* instructions on sensitive-content screens explaining how to regulate their emotions. Such education/training could occur as part of creating an account on these platforms and/or as a pop-up intervention (e.g., "Would you like to learn *how* to regulate your emotions while online?") for existing users. There are now 4.76 billion social media users around the world (~60% of the total global population); Instagram specifically, is the fourth most popular platform (after Facebook, YouTube, and WhatsApp) with 2.00 billion monthly users (Kemp, 2024). Therefore, such education/training in addition to instructions on sensitive-content screens would affect a large numbers of social media users, and subsequent reductions in distress may have cascading effects on their other behaviours (e.g., decrease distress driven self-harm behaviours; see Hetrick et al., 2020). Future research could examine how effective other strategies are in the social media context, particularly those that can be implemented with minimal education/training (e.g., acceptance: experiencing emotions without judgment; Wolgast et al., 2011).

Limitations

Our study has several limitations. First, because participants received education/training in addition to cognitive emotion regulation instructions, we do not know whether the instructions *alone* have an effect. Therefore, future research should examine whether our findings replicate with less education/training. Second, we used a general population of Instagram users, meaning we do not know if people with mental health

vulnerabilities (e.g., clinical depression) would respond differently. Although future research could examine such sub-populations, previous research has found no evidence that warnings influence emotional reactions differently among sub-populations (see Bridgland et al., 2023). Relatedly, the distribution of distress ratings (from 0-100) across all conditions suggests individual differences (e.g., distress tolerance and regulation ability) may influence people's emotional responses and how they regulate their emotions (with and without regulation instructions). Future research could include a battery of individual differences measures (e.g., Distress Tolerance Scale [DTS], Perth Emotion Regulation Competency Inventory [PERCI]; Preece et al., 2018; Simons & Gaher, 2005) to explore their influence on people's emotional responses and regulation efforts. Finally, we can only speculate on the true effect of distraction because we do not know whether it was features of the within- or the between-subjects design that drove the differences between the two studies. Future research could address this limitation by employing a within-subjects block design whereby participants are randomly allocated to first see either a distraction or no instruction block, followed by the block they did not see. Researchers could evaluate the within-subjects effect (as we did in Study 4a), but also make a between-subjects comparison on the first block (as we did in Study 4b).

Constraints on Generality

We crowdsourced online participants from the United States who use Instagram. We chose this sample because prior work on sensitive-content screens and trigger warnings more generally is predominantly based on such Western, educated, industrialised, rich, and democratic (or WEIRD) populations and we wanted to compare to this research. We have no reason to believe our results depend on particular characteristics of the participants or the materials, and therefore believe our results would be reproducible with similar participants, image stimuli and warning statements, and likely with other measures of emotional impact

(e.g., anxiety) and other social media platforms (e.g., TikTok). However, there are two key constraints to generality. First, we do not know how our findings would apply to people from different cultural and socioeconomic contexts. Although we know that people from a variety of backgrounds use social media (Kemp, 2024), and affective reactions toward emotional stimuli often differ between Western and non-Western samples (e.g., Huang et al., 2015), there is no available data on reactions to sensitive-content screens (and trigger warnings more generally) or screened sensitive content *between* different cultural and socioeconomic contexts. Therefore, caution should be exercised when considering the generalisability of our results beyond Western contexts, and future research should examine the role of cognitive emotion regulation instructions on sensitive-content screens in more diverse samples.

Second, it is also possible that our results may vary outside a controlled experimental context. Although we included the Instagram logo and "like"/comment buttons to replicate the feel and experience of viewing images on Instagram, participants saw only *negative* images—all of which were preceded by a sensitive-content screen and were not likely to be personally relevant—and had no control over what images they saw and for how long. Therefore, participants may have experienced different levels of anxiety (e.g., Havranek et al. 2015) and thus distress—relative to if they were on their own Instagram accounts and saw negative images preceded by sensitive-content screens *amongst* other neutral and positive images, and/or were given a choice to avoid the negative (or sensitive) content. Although we cannot rule out this possibility, distraction and reappraisal reduced participants' distress *relative* to a control condition (where anxiety also may have been high), meaning our conclusions would likely remain unchanged. Nonetheless, future research could incorporate cognitive emotion regulation instructions within a more ecologically valid design whereby participants are shown a series of sensitive-content screens amongst other neutral and positive images and given the option to uncover screens or not—as they would on their own

Instagram accounts, allowing us to examine distress on trials where participants choose to view the content.

Conclusions

We examined whether adding cognitive emotion regulation instructions to sensitive-content screens improves their efficacy. We found they did: participants reported less distress following exposure to negative images when they were proceeded by sensitive-content screens with distraction or reappraisal instructions, compared with no instructions. Our findings suggest that sensitive-content screens in their current format (without instructions) fail to help people emotionally prepare, and suggest that providing explicit instructions detailing *how* to emotionally prepare—using cognitive emotion regulation strategies—can reduce the negative impact of exposure to sensitive content. Therefore, social media platforms should move beyond merely warning people about upcoming content and add cognitive emotion regulation instructions to sensitive-content screens to make them a more effective harm-minimisation tool.

Supplementary Materials

Table S4.1

Demographic Characteristics

Variable	Study 4a % (*n*)	Study 4b % (*n*)
Ethnicity		
European American/White	65.1% (125)	60.6% (103)
African American/Black	10.4% (20)	10.0% (17)
Asian	3.6% (7)	4.7% (8)
Middle Eastern	0.5% (1)	0.0% (0)
European	0.0% (0)	1.8% (3)
Hispanic	6.8% (13)	5.9% (10)
Mixed race	9.9% (19)	5.9% (10)
Prefer not to answer	0.0% (0)	1.2% (2)
Specified nationality (e.g., American/USA)	4.2% (8)	10.0% (17)
Household income		
<$20,000	9.4% (18)	9.4% (16)
$20,000 - $45,000	28.6% (55)	19.4% (33)
$45,000 - $140,000	54.2% (104)	52.4% (89)
$140,000 - $150,000	2.6% (5)	8.2% (14)
$150,000 - $200,000	2.6% (5)	4.1% (7)
>$200,000	2.6% (5)	6.5% (11)
Education		
Less than high school graduate	0.5% (1)	0.0% (0)
High school graduate	12.0% (23)	11.8% (20)
Some college	32.3% (62)	30.0% (51)
College Graduate	55.2% (106)	58.2% (99)

Table S4.2

Social Media Use

Variable	Study 4a % (*n*)	Study 4b % (*n*)
Social media platform		
Facebook	85.4% (164)	90.0% (153)
Instagram	100.0% (192)	100.0% (170)
Twitter	71.4% (137)	52.9% (90)
Snapchat	34.9% (67)	38.8% (66)
WhatsApp	17.2% (33)	14.7% (25)
Tumblr	16.1% (31)	9.4% (16)
YouTube	91.1% (175)	81.8% (139)
TikTok	55.7% (107)	57.6% (98)
Reddit	75.0% (144)	63.5% (108)
Pinterest	40.1% (77)	34.7% (59)
Other (open text)	1.0% (2)	4.7% (8)
In the last 7 days, how many days did you use Instagram?		
Never	0.5% (1)	0.6% (1)
1 day	4.7% (9)	1.8% (3)
2 days	8.3% (16)	4.7% (8)
3 days	10.4% (20)	10.6% (18)
4 days	8.9% (17)	5.9% (10)
5 days	12.0% (23)	9.4% (16)
6 days	5.7% (11)	5.3% (9)
Everyday	49.5% (95)	61.8% (105)
In the last 30 days, on an average day how many hours did you use Instagram?		
Less than half an hour	33.3% (64)	21.8% (37)
1 hour	36.5% (70)	34.1% (58)
2-3 hours	14.6% (28)	30.6% (52)
4-5 hours	7.8% (15)	8.2% (14)
More than 6 hours	7.8% (15)	5.3% (9)

Table S4.3

Study 4a: Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Conditions on Distress Ratings While Controlling for Participants' Distress on the Previous Trial

	Distress Rating		
Predictor	*Estimate (SE)*	*95% CI*	*p value*
(Intercept)	43.97 (1.58)	40.85 – 47.09	<.001
Distraction	-2.18 (0.69)	-3.54 – -0.82	.002
Reappraisal	-3.89 (0.77)	-5.41 – -2.38	< .001
Lagged distress (person mean centered)	0.15 (0.02)	0.11 – 0.18	< .001

Note. The intercept for the model represents the mean for the no regulation instruction condition, with the condition estimates indicating the difference relative to the intercept. $\alpha = .05$.

Table S4.4

Study 4a: Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Conditions (Distraction vs. Reappraisal) on Distress Ratings While Controlling for Participants' Distress on the Previous Trial

	Distress Rating		
Predictor	*Estimate (SE)*	*95% CI*	*p value*
(Intercept)	41.79 (1.62)	38.59 – 44.99	<.001
No Instruction	2.18 (0.69)	0.82 – 3.54	.002
Reappraisal	-1.71 (0.73)	-3.15 – -0.28	.019
Lagged distress (person mean centered)	0.15 (0.02)	0.11 – 0.18	< .001

Note. The intercept for the model represents the mean for the distraction condition, with the condition estimates indicating the difference relative to the intercept. $\alpha = .05$.

Table S4.5

Study 4a: Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Conditions on Distress Ratings Excluding Participants Who Looked Away from Some Negative Images, with No Instruction as Reference Category

	Distress Rating		
Predictor	*Estimate (SE)*	*95% CI*	*p value*
(Intercept)	41.05 (1.85)	37.39 – 44.71	<.001
Distraction	-2.34 (0.83)	-3.98 – -0.69	.006
Reappraisal	-4.17 (0.97)	-6.08 – -2.27	<.001

Note. The intercept for the model represents the mean for the no regulation instruction condition, with the condition estimates indicating the difference relative to the intercept. $\alpha = .05$.

Table S4.6

Study 4a: Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Conditions on Distress Ratings Excluding Participants Who Looked Away from Some Negative Images, with Distraction as Reference Category

	Distress Rating		
Predictor	*Estimate (SE)*	*95% CI*	*p value*
(Intercept)	38.72 (1.82)	35.12 – 42.31	<.001
No Instruction	2.34 (0.74)	0.88 – -3.79	.002
Reappraisal	-1.84 (0.91)	-3.65 – -0.03	.047

Note. The intercept for the model represents the mean for the distraction condition, with the condition estimates indicating the difference relative to the intercept. $\alpha = .05$.

Table S4.7

Study 4a: What Participants Did in Response to Images Without Instructions to Regulate

Response	% (*n*)
I tried to think about what was happening in the image in a new way, so that I felt less negative (Reappraisal)	8.3% (16)
I tried to think about something else, unrelated to what was happening in the image, so that I felt less negative (Distraction)	18.8% (36)
I just looked at the image and responded naturally	70.3% (135)
Other (open text)	2.6% (5)
"I braced myself to see something unpleasant…"	0.5% (1)
"I tried to continue alternating between the two"	0.5% (1)
"I used the strategy last prompted…"	1.5% (3)

Table S4.8

Study 4a: Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Conditions on Distress Ratings Excluding Participants Who Did Not Report Responding Naturally, with No Instruction as Reference Category

	Distress Rating		
Predictor	*Estimate (SE)*	*95% CI*	*p value*
(Intercept)	43.95 (1.89)	40.22 – 47.69	<.001
Distraction	-2.35 (0.86)	-4.05 – -0.65	.007
Reappraisal	-4.58 (0.95)	-6.46 – -2.69	<.001

Note. The intercept for the model represents the mean for the no regulation instruction condition, with the condition estimates indicating the difference relative to the intercept. $\alpha = .05$.

Table S4.9

Study 4a: Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Conditions on Distress Ratings Excluding Participants Who Did Not Report Responding Naturally, with Distraction as Reference Category

	Distress Rating		
Predictor	*Estimate (SE)*	*95% CI*	*p value*
(Intercept)	41.60 (1.86)	37.93 – 45.27	<.001
No Instruction	2.35 (0.85)	0.68 – 4.03	.006
Reappraisal	-2.22 (0.86)	-3.92 – -0.53	.011

Note. The intercept for the model represents the mean for the distraction condition, with the condition estimates indicating the difference relative to the intercept. α = .05.

Table S4.10

Study 4b: Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Distraction on Distress Ratings While Controlling for Participants' Distress on the Previous Trial

	Distress Rating		
Predictor	*Estimate (SE)*	*95% CI*	*p value*
(Intercept)	54.56 (2.56)	49.47 – 59.66	<.001
Distraction	-18.76 (3.78)	-26.22 – -11.30	< .001
Lagged distress (person mean centered)	0.13 (0.02)	0.09 – 0.17	< .001

Note. The intercept for the model represents the mean for the no regulation instruction condition, with the distraction estimate indicating the difference relative to the intercept. α = .05.

Table S4.11

Study 4b: Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Distraction on Distress Ratings Excluding Participants Who Looked Away from Some Negative Images, with No Instruction as Reference Category

	Distress Rating		
Predictor	*Estimate (SE)*	*95% CI*	*p value*
(Intercept)	49.11(3.27)	42.57 – 55.65	<.001
Distraction	-16.29 (4.64)	-25.48 – -7.09	0.001

Note. The intercept for the model represents the mean for the no regulation instruction condition, with the distraction estimate indicating the difference relative to the intercept. $\alpha = .05$.

Table S4.12

Study 4b: Themes for why Participants Looked Away from Some Negative Images During the Task

Themes	% of total N (*n*)
The images were hard to look at	7.6% (13)
The content was gross/gruesome	8.2% (14)
The images made me upset/distressed	3.5% (6)
I felt like it was something I wasn't supposed to see	1.2% (2)
The content was disturbing	6.5% (11)
The images made me feel sick	3.5% (6)
Other	2.9% (5)

Study 4a: Repeated Use of Emotion Regulation Strategies

As an exploratory aim, we examined the changes in effectiveness of strategies with repeated use (over the task). We ran linear mixed effects models to examine this exploratory aim. We compared each regulation strategy condition (reappraisal, distraction) to no instruction, by setting "no instruction" as the reference category. We also included a task order variable (which we created from each participants' randomisation data). See Table S4.13 for model coefficients and their associated inferential statistics. Overall, the model showed that there was a small negative effect of order, meaning that as image trials progressed participants reported slightly lower distress regardless of condition. This pattern is consistent with people habituating to the images over time (i.e., as trials progressed during the task). However, there was no interaction between either emotion regulation strategy and task order, meaning the decrease (owing to habituation) impacted both conditions similarly. Therefore, we found no evidence to suggest that the effectiveness of strategies changed over the task; but we acknowledge that we only had 10 trials per conditions, meaning we were not well positioned, methodology wise, to examine repeated use of strategies (or longer-term effectiveness).

Table S4.13

Study 4a: Coefficient Estimates for Fixed Effects from Linear Mixed Effects Models Testing the Effect of Regulation Strategy Conditions on Distress Ratings with Repeated Use, and with No Instruction as Reference Category

	Distress Rating		
Predictor	*Estimate (SE)*	*95% CI*	*p value*
(Intercept)	47.54 (1.79)	44.02 – 51.07	<.001
Distraction	-3.70 (1.27)	-6.20 – - 1.21	.004
Reappraisal	-4.28 (1.33)	-6.89 – -1.68	.001
Order	-0.22 (0.06)	-0.34 – -0.10	<.001
Distraction * Order	0.10 (0.07)	-0.04 – 0.23	0.164
Reappraisal * Order	0.02 (0.07)	-0.11 – 0.16	0.751

Note. The intercept for the model represents the mean for the no regulation instruction condition, with the condition estimates indicating the difference relative to the intercept. $\alpha = .05$.

Table S4.14

Study 4a: Participants Self-Reported Ease of Use and Effectiveness of Regulatory Strategies, and Paired-Samples T-Test Testing the Difference Between Conditions

Variable	*M* (*SD*)	Paired-samples t-test
Ease of use		
Reappraisal	47.1 (26.3)	$t(191) = -1.41$, $p = .159$
Distraction	50.6 (28.9)	
Perceived effectiveness		
Reappraisal	42.6 (27.5)	$t(191) = -0.73$, $p = .469$
Distraction	44.3 (28.2)	

Note. Ratings were made on a slider scale from 0 = not at all easy/effective, to 100 = extremely easy/effective.

Table S4.15

Study 4a: How Often Participants Use Each Strategy (or Another Strategy) When They Come Across Negative Content Online in Their Everyday Lives

Strategy	Frequency of Use				
	Never	Rarely	Sometimes	Very Often	Always
Distraction	4.7% (9)	17.7% (34)	38.0% (73)	35.4% (68)	4.2% (8)
Reappraisal	9.9% (19)	33.3% (64)	41.1% (79)	14.6% (28)	1.0% (2)
Other	11.5% (22)	25.5% (49)	45.8% (88)	14.6% (28)	2.6% (5)

Table S4.16

Study 4a: Themes for "Other" Strategies Participants Use (Sometimes, Very Often, or Always) When They Come Across Negative Content Online in Their Everyday Lives

Themes	% of total N (*n*)
Avoidance, e.g., "I remove myself from whatever situation is distressing me"	19.3% (37)
Acceptance, e.g., "If I can't change it, I have to accept it"	9.4% (18)
Detachment, e.g., "[I] disengage emotionally"	4.2% (8)
Meaning making, e.g., "I find context for the image"	3.6% (7)
Prayer	2.1% (4)
Solution focused coping, e.g., "I look for a way to fix/help"	2.1% (4)
Humour, e.g., "I try humour, just laughing at distressing things"	1.6% (3)
Mindfulness, e.g., "[I use] controlled breathing"	1.6% (3)
Suppression, e.g., "I suppress the bad emotions"	1.6% (3)
Compartmentalisation, e.g., "I put them in an area of the brain that isn't used much"	1.0% (2)
Pleasure seeking, e.g., "I play music that I enjoy"	1.0% (2)
Situation modification, e.g., "I close my eyes halfway to make the image look blurry"	1.0% (2)
Comparison, e.g., "I compare that situation to worse ones"	0.5% (1)
Gratitude, e.g., "I try to be grateful for my own fortune…"	0.5% (1)
Support seeking, e.g., "I talk to my best friends"	0.5% (1)
Reappraisal (based on description provided), e.g., "[I] change the meaning in my mind"	4.7% (9)
Distraction (based on description provided), e.g., "[I] think about something to eat"	3.6% (7)
Other (e.g., unsure/none)	4.7% (9)

Table S4.17

Study 4a: Participants Strategy Choice if They Were Asked to do The Task Again

Response	% (*n*)
Reappraisal	31.8% (61)
Distraction	50.0% (96)
I would use a different strategy to manage my emotions	13.0% (25)
What strategy? (Open Text)	
Acceptance, e.g., "Letting myself be upset helps me move on"	2.1% (4)
Avoidance, e.g., "I would look away completely"	5.7% (11)
Suppression	0.5% (1)
Detachment	0.5% (1)
Prayer	0.5% (1)
Mindfulness, e.g., "breathing techniques"	0.5% (1)
Respond naturally	0.5% (1)
Unsure	1.6% (3)
Reappraisal or distraction (based on description provided)	1.0% (2)
I would not use any strategy to manage my emotions	5.2% (10)
Why? (Open Text)	
Prefer to use acceptance, e.g., "…I feel and try not to control anything"	2.1% (4)
The strategies were hard/didn't work, e.g., "It was hard to try to suppress my emotions or use tactics"	2.1% (5)
Not needed, e.g., "most things don't upset me"	0.5% (1)

Table S4.18

Study 4a: Participants Preferences for Instagram's Sensitive Content Control Feature

Response	% (*n*)
Have used feature:	18.2% (35)
More	5.7% (11)
Standard	9.4% (18)
Less	3.1% (60)
Have not used feature, but *hypothetically would choose:*	81.8% (157)
More	11.5% (22)
Standard	52.1% (100)
Less	18.2% (35)

Table S4.19

Study 4b: Participants Preferences for Instagram's Sensitive Content Control Feature

Response	% (*n*)
Have used feature:	22.9% (39)
More	10.0% (17)
Standard	9.4% (16)
Less	3.5% (6)
Have not used feature, but *hypothetically would choose:*	77.1% (131)
More	11.8% (20)
Standard	34.7% (59)
Less	30.6% (52)

Table S4.20

Study 4a: General Task Compliance Questions

Variable	Yes: % (*n*)	No: % (*n*)
Always read instructions	97.4% (187)	2.6% (5)
Always followed instructions	98.4% (189)	1.6% (3)
Looked away from images	31.8% (61)	68.2% (131)
Left image viewing task	0.0% (0)	100% (192)

Table S4.21

Study 4a: What Participants Did in Response to Images Without Instructions to Regulate

Response	% (*n*)
I tried to think about what was happening in the image in a new way, so that I felt less negative (Reappraisal)	8.3% (16)
I tried to think about something else, unrelated to what was happening in the image, so that I felt less negative (Distraction)	18.8% (36)
I just looked at the image and responded naturally	70.3% (135)
Other (open text)	2.6% (5)
"I braced myself to see something unpleasant…"	0.5% (1)
"I tried to continue alternating between the two"	0.5% (1)
"I used the strategy last prompted…"	1.5% (3)

Table S4.22

Study 4b: General Task Compliance Questions

Variable	Yes: % (*n*)	No: % (*n*)
Looked away from images	33.5% (57)	66.5% (113)
Left image viewing task	0.0% (0)	100% (170)

7 General Discussion

My thesis, broadly speaking, aimed to investigate the empirical basis of sensitive-content screens. Specifically, it aimed to answer questions left from the first (albeit small) wave of research on sensitive-content screens by examining 1) *how* people respond to sensitive-content screens when they see more than one screen—both in terms of their behaviour and their emotional experiences—2) *why* they respond the way they do, and 3) two potential ways social media platforms could adapt sensitive-content screens to improve the screens' utility as a harm minimisation tool. This final chapter draws together the findings from my four empirical chapters in the context of previous research and theories I identified in Chapter 1. I also discuss the theoretical, methodological, practical, and clinical implications of my findings, acknowledge key limitations of my research, and suggest future directions.

7.1 Summary of Findings

Recall, advocates make two key claims about sensitive-content screens. First, they claim sensitive-content screens deter people from viewing sensitive content by giving them an opportunity to avoid it. Second, they claim that *if* people decide to uncover sensitive-content screens, such forewarning helps them emotionally prepare for the content (e.g., Cripps, 2020, Manne, 2015). To examine these claims, my thesis first examined *how* people respond to sensitive-content screens. I focused on the first claim related to deterrence, because there is more comprehensive existing evidence for the second claim related to emotional preparation—meaning I was already well positioned to investigate adaptions related to emotional preparation. Nonetheless, for completeness, I still discuss how my findings align with (and support) the existing evidence on emotional preparation.

Do Sensitive-Content Screens Deter People from Viewing Sensitive Content?

In Studies 1a and 1b (Chapter 2), I began exploring the first claim related to deterrence by examining how people respond to a *series* of sensitive-content screens during an image-viewing task. Specifically, I examined behaviour over a series of sensitive-content screens because we know that people—especially people who seek out sensitive content, and then see more of it because of the algorithm (e.g., Within Health, 2023)—are likely to see more than *one* sensitive-content screen in real life, thereby improving the ecological validity of my design.

To compare with the first investigation of sensitive-content screens—which found most people indicated a desire (80.0%; Study 1) or made a choice (84.7%; Study 2) to uncover a *single* sensitive-content screen (Bridgland, Bellet et al., 2022)—I initially examined how people responded to the *first* sensitive-content screen they saw. Consistent with the prior research, most participants opted to uncover the first sensitive-content screen they came across during the image-viewing task. This pattern occurred irrespective of whether the screens warned about content type (e.g., "This photo may contain graphic or violent content"), or the emotional responses people may experience (e.g., "This photo contains sensitive content which some people may find offensive or disturbing"). This pattern also occurred irrespective of whether the task had an image-response delay (i.e., a 3s delay between screen presentation and when response options [i.e., *See Photo* and *Next Photo*] appeared; 86.0%; Study 1b) or not (84.4%; Study 1a). I replicated this finding again in Study 3b (Chapter 4): 84.5% of participants uncovered the first sensitive-content screen they came across during the image-viewing task, following a 3s image-response delay—though in this study half of the sample saw the first sensitive-content screen with brief content-related information. Therefore, forcing people to pause for a few of seconds *before* responding to the first sensitive-content screen does not mean they will be any less likely to uncover it. This

finding suggests such behaviour is not merely a result of people inattentively uncovering, but rather a conscious decision. Merely encouraging people to pause before responding to sensitive-content screens then, would likely not reduce uncovering behaviour.

In Studies 1a and 1b, I also examined how people responded to *subsequent* sensitive-content screens during an image-viewing task—because we know that exposure to sensitive content can occur relatively frequently for some people (e.g., Fulcher et al., 2020; Wang et al., 2018), and uncovering behaviour may change with cumulative exposure (e.g., after exposure to aversive stimuli; for a review on avoidance learning see Krypotos et al., 2015). Across both studies, many people continued to uncover sensitive-content screens—despite seeing negative images underneath each screen. In fact, in Study 1a, 51.7% of participants uncovered every sensitive-content screen, and in Study 1b, 38.7% of participants uncovered over half of the sensitive-content screens (i.e., 15 of 30).

Notably, in Study 1b—when there was an image-response delay—only 17.5% of participants uncovered every sensitive-content screen (compared with 51.7% in Study 1a). As I noted in Chapter 2, there are several possible explanations for this discrepancy. First, in Study 1a, we used a pool of 70 negative images, from which participants saw a subset of varying size and content, whereas in Study 1b, all participants saw the *most negative* 30 of these 70 images. Therefore, the images participants uncovered in Study 1b were likely more negative, which may have made participants less likely to uncover them, especially images subsequent to those they uncovered initially. Indeed, when images are similarly negative people may stop uncovering them because they use information from previous images to fill information gaps. But, when the level of negativity varies between images, people may have unresolved information gaps—meaning their curiosity about the images, and thus their desire to uncover sensitive-content screens, may remain high throughout the image-viewing task. Notably, given that screened images are likely to vary in negativity on Instagram in *real-*

life—because guidelines encompass a broad range of sensitive content (e.g., from hate speech to violent and graphic content)—uncovering behaviour in Study 1a is mostly likely to generalise to real-life. Second, because participants saw more sensitive-content screens in Study 1b (30 vs. 20 in Study 1a), there was greater opportunity for people's curiosity to 'wear' off (Day, 1982). Although this explanation is similar to the first explanation, it reflects a more general declining of curiosity related to the content, rather than changes in curiosity due to the variability of image negativity. Notably, the data pattern in Study 1b supports both possibilities; more participants uncovered screens initially (51.9%-86.0% over the first five screens), then uncovering steadily decreased until it plateaued over the final 10 screens (just below 30.0%).

In Studies 1a and 1b, I also examined whether vulnerable people (e.g., people with higher depression symptoms) were more susceptible to uncovering sensitive-content screens, relative to people with less severe psychopathological symptoms. Prior research on sensitive-content screens has found evidence that suggests that vulnerable people may be more susceptible to uncovering sensitive-content screens (Study 1; Bridgland, Bellet et al., 2023), but also evidence that uncovering behaviour is not related to people's vulnerabilities (Study 2; Bridgland, Bellet et al., 2023). Consistent with the latter finding, I found no evidence suggesting vulnerable people were more susceptible to uncovering sensitive-content screens, either on the first sensitive-content screen they came across, or when they viewed a series of sensitive-content screens (Studies 1a and 1b). However, my findings also suggest that vulnerable people were no more likely to avoid such content (e.g., by actively deciding not to uncover sensitive-content screens, or by viewing images at a slower pace; Study 1a). Broadly, then, these findings are consistent with traditional trigger warning research; for example, within the educational context, Kimble et al., (2021) found most (95.6%) of students were

willing to read triggering material (even when they had the opportunity to avoid it)—including students with experience of trauma (96.9%) and probable PTSD (97.6%).

Together, these findings suggest that sensitive-content screens do not necessarily deter people from viewing sensitive content. Rather, people tend to engage with sensitive content irrespective of their vulnerabilities *and* whether they receive a forewarning. As well as adding to the existing literature on sensitive-content screens specifically, this research is the first to suggest that traditional trigger warnings are ineffective at promoting deterrence *within* a social media context. These findings also fit, more broadly, with some of the existing theory and related literature I discussed in Chapter 1 including the information-gap hypothesis (Loewenstein, 1994), the "Pandora effect" (Hsee & Ruan, 2016; Yagi et al., 2023), morbid curiosity (Oosterwijk, 2017). Collectively, these findings suggest people are motivated to fill information gaps, willing to risk negative (or aversive) consequences, and have a genuine interest in highly negative information. Indeed, labelling sensitive content, like sensitive-content screens do, may well elicit "forbidden fruit" (Weaver, 2011), and "boomerang" (Brehm, 1966) effects, whereby people view restricted content as *more* attractive and intentionally engage with it.

Do Sensitive-Content Screens Emotionally Prepare People to View Sensitive Content?

Recall, existing evidence demonstrates that sensitive-content screens create a noxious anticipatory period that does not translate to an emotional benefit when people view the forewarned content (Takarangi et al., 2023). Therefore, I did not directly investigate the claim relating to emotional preparation. However, for completeness, here I discuss how my findings from Studies 3a, 3b (Chapter 5), 4a and 4b (Chapter 6)—in which I examined possible adaptions—align with (and support) the existing evidence on emotional preparation. Specifically, I examine the control conditions—where participants viewed screens in the current format, without content descriptions or emotion regulation instructions—within these

studies; therefore, I can draw on these conditions to see whether people's emotional responses to such screens, as they currently appear on social media, align with the idea of emotional preparation. First, I examine people's emotional experiences in the anticipatory period, before turning to people's emotional reactions while viewing the forewarned content.

Anticipatory Period

In Study 3a, I examined participants' change in state anxiety when exposed to sensitive-content screens (with and without brief and detailed content descriptions) during an image-viewing task. Participants' state anxiety was higher (i.e., more negative) after they saw sensitive-content screens (compared to baseline) in every condition, including the control condition when the screens appeared in their current format (i.e., without content descriptions). Therefore, in line with existing evidence (Takarangi et al., 2023), sensitive-content screens appear to create a noxious anticipatory period.

Emotional Reactions to Forewarned Content

However, as discussed in Chapter 1, we could conceptualise the noxious anticipatory period as a form of emotional preparation *if* it mitigates the emotional impact of viewing the forewarned content. Therefore, in Study 3b I examined whether participants' distress was offset when they viewed sensitive content preceded by a sensitive-content screen (though I note, I did not compare this condition to a no screen condition). Participants who saw sensitive-content screens *and* decided to uncover them (n = 100), were mildly to moderately distressed (M = 36.6, SD = 28.8; 0 = *not at all distressed*, to 100 = *extremely distressed*) after viewing the negative image. Participants who viewed negative images preceded by sensitive-content screens (without instructions) in Studies 4a and 4b (Chapter 6) reported similar levels of distress (Study 4a: M = 44.1, SD = 22.2; Study 4b: M = 54.4, SD = 35.2)—irrespective of the fact that participants in Study 4a also saw *other* sensitive-content screens with instructions to regulate their emotions within the same image-viewing task. I note that

distress ratings were higher in Studies 4a and 4b, compared with Study 3b, possibly because participants did not have an option to avoid the sensitive content (like they did in Study 3a). Indeed, as noted in Chapter 5, people who were more susceptible to amplifications in distress following sensitive-content screens may have decided not to uncover them in Study 3a. Nonetheless, these findings demonstrate that people experience mild to moderate negative emotional reactions while viewing the forewarned content, irrespective of whether they receive a forewarning—inconsistent with the idea of emotional preparation.

Together, these findings align with prior research on sensitive-content screens (Takarangi et al., 2023), and trigger warnings more generally (see Bridgland, Jones et al., 2023), which finds sensitive-content screens do not help people emotionally prepare to view sensitive content. Rather, sensitive-content screens appear to create a noxious anticipatory period that does not translate to an emotional benefit when people view the forewarned content. This finding specifically, fits with what we know about bracing for the worst, which has negative impacts during the anticipatory period (i.e., before the outcome is known), and provides little to no benefit after the outcome is known (e.g., Golub et al., 2009; Neubauer et al., 2018; Sweeny et al., 2016). As discussed in Chapters 1 and 6, sensitive-content screens (and traditional trigger warnings) may fail to help people emotionally prepare *because* they do not equip people with strategies for emotional preparation (e.g., emotion regulation strategies; Bridgland, Barnard et al., 2022), or assist people in taking a moment to pause before proceeding to the forewarned content (Bridgland & Takarangi, 2022).

Overall, these findings suggest sensitive-content screens do not deter people from viewing sensitive content *or* help them emotionally prepare for the content—contrary to advocates claims. Put simply then, sensitive-content screens in their current format do not function as intended.

Why do People Respond to Sensitive-Content Screens the way they do?

In Studies 1a and 1b (Chapter 3), I began investigating the reasons underpinning people's uncovering behaviour—aiming to understand why sensitive-content screens do not function as intended. Despite the growing body of literature on trigger warnings (and sensitive-content screens specifically), only one study (Bridgland, Bellet et al., 2023) has examined participants behaviour *and* explicitly asked them to report reasons for their behaviour. Most participants uncovered (or said they would uncover) a single sensitive-content screen because they were curious; other people said they would decide based on the context of the image (e.g., posting account, and content descriptions) and/or their ability to cope with distressing content (Bridgland, Bellet et al., 2023). However, the reasons underpinning people's uncovering behaviour may change on subsequent screens. For example, people may initially uncover sensitive-content screens because they are curious, but then continue uncovering because they learn they are able to manage their image-related distress. Therefore, I explored participant's reasons for their behaviour over a series of sensitive-content screens, beginning first with the reasons for *uncovering* sensitive-content screens.

Information seeking behaviour was the most commonly endorsed reason for uncovering sensitive-content screens in Studies 1a and 1b. This finding is consistent with prior research (e.g., Bridgland, Bellet et al., 2023), and with literature I discussed in Chapter 1, specifically the information-gap hypothesis (Loewenstein, 1994), the "Pandora effect" (Hsee & Ruan, 2016; Yagi et al., 2023), and morbid curiosity (Oosterwijk, 2017). However, because I derived the factors in Studies 1a and 1b from self-report, and I asked participants to reflect on their motivations for behaviour *after* completing the image-viewing task, retrospective bias and/or reporting inaccuracies may have influenced the results (see Schwarz, 2007). Therefore, in Study 2 I re-assessed information seeking behaviour using an

experimental paradigm—a method less vulnerable to bias and reporting inaccuracies. Completing such a follow up also facilitated method triangulation—the use of multiple research strategies to examine the same research question—thereby strengthening our confidence in the validity of the findings (Carter et al., 2014). Specifically, in Study 2 I varied the amount of content-related information—by including content descriptions on some sensitive-content screens during a simulated Instagram task—and then examined participants' uncovering behaviour. Participants uncovered sensitive-content screens *most* often when they had the least amount of information available to them; that is, when they saw sensitive-content screens as they typically appear on Instagram—with a non-specific warning. This finding is consistent with the idea that people uncover sensitive-content screens because they want to obtain information about the image and/or alleviate uncertainty and curiosity—which arguably stems from the ambiguity of screens to begin with. Put simply, sensitive-content screens may *increase* engagement with—rather than deter people from—sensitive content because they prompt information seeking behaviour in their current format.

Participants also endorsed uncovering sensitive-content screens because of their past, current, and/or anticipated affective states, both *negative* and *positive*. This finding is consistent with the idea that responses to sensitive-content screens may reflect regulation efforts, as discussed in Chapter 1. Notably, although affect can drive uncovering behaviour, the emotion *goal* (or desired end-state; Tamir, 2016) of such behaviour is unknown. Often people are motivated to experience emotions for their *hedonic* value (i.e., their immediate phenomenology)—and therefore, in most situations seek to up-regulate immediate pleasure and/or down-regulate immediate pain (Tamir, 2016). Indeed, the idea that sensitive-content screens will elicit helpful emotional preparation stems from an assumption that people will engage in hedonically driven regulation. However, we also know that people can be motivated to experience emotions for their potential benefits in the future (i.e., their

instrumental value; Tamir, 2016). For example, some people with depression seek out sadness for self-verification motives (e.g., because this affective state aligns with their negative self-view; Millgram et al., 2015), even though such behaviour may serve to maintain their depression (Beck & Alford, 2009). Therefore, people can engage in affect driven behaviour in attempts to up- or down-regulate their affective states, and/or to maintain them (Millgram et al., 2020). This finding suggests that people may *purposely* not emotionally prepare for the upcoming content after receiving a forewarning if their emotion goals do not align with down-regulating negative affect—irrespective of whether they have strategies to draw on for emotional preparation.

In Studies 1a and 1b, I examined whether vulnerable people (e.g., people with higher depression symptoms) were more likely to endorse certain reasons for uncovering sensitive-content screens. Overall, I found large discrepancies in findings between the two studies. For example, in Study 1, there was no relationship between people's overall PTSD symptomology and their reasons for uncovering screens. However, in Study 2, the higher people's PTSD symptomology, the more likely they were to endorse being motivated by information seeking behaviour *and* negative and positive affect driven behaviour. Methodological changes between the two studies (e.g., the number/nature of images could have influenced participants' reasons for uncovering screens, alongside their actual uncovering behaviour) may explain these discrepancies. Nonetheless, there does not appear to be *one* reason underpinning vulnerable people's behaviour. In fact, it likely that existing trait vulnerabilities (e.g., depression) interact with *state* mood factors (e.g., low mood) *and* contextual motivations (e.g., being alone at night vs. with others during the day) to influence uncovering decisions—making such relationships difficult to identify with simple correlations. Future research could measure trait, state *and* contextual factors and examine

their independent and interacting influence on uncovering behaviour using more complex analyses (e.g., regression analyses).

My thesis focused on the reasons for *avoiding* sensitive content to a lesser extent—because I had significantly fewer participants to draw upon in this sub-sample—but I found that people also avoided sensitive content for different reasons. Although the reasons were not distinct enough to load onto separate factors—perhaps because I did not include enough items for each reason—the items seemed conceptually different (e.g., "I did not uncover the screened image(s) because I do not enjoy taking risks" vs. "I did not uncover the screened image(s) because I do not like viewing distressing material"). Such *avoidance behaviour* is often viewed as maladaptive because it is a hallmark feature and maintaining factor of many emotional disorders, including anxiety disorders (e.g., agoraphobia, specific phobias, and social anxiety) and PTSD (e.g., Barlow, 2021). However, we can also conceptualise such behaviours as adaptive. For example, avoidance behaviour could be a form of problem-focused disengagement coping, whereby people avoid a perceived threat (e.g., the forewarned content; Carver & Connor-Smith, 2010; Skinner et al., 2003). Indeed, it may well be adaptive for people to protect their wellbeing by consciously consuming content, and flexibly avoiding potentially distressing content. However, an overreliance on problem-focused disengagement coping may become maladaptive if it is not appropriate for the context (e.g., when managing anxiety disorders; see Hofmann & Hay, 2018).

Together, these findings demonstrate that people view (and avoid) sensitive content for different reasons—reasons that may change on subsequent screens and with varying emotion goals. Not only may people view sensitive content because sensitive-content screens prompt information seeking behaviour, but they may also intentionally seek out sensitive content because they want to regulate their affect and doing so may help them achieve their emotion goal(s). Therefore, sensitive-content screens may not function as intended because

they prompt engagement with sensitive content *and* fail to counteract the motivation some people have to regulate their affect via sensitive content.

How can Social Media Platforms Adapt Sensitive-Content Screens to Improve the Screens Utility as a Harm Minimisation Tool?

Harm minimisation tools aim to mitigate the negative impact associated with engaging in potentially harmful behaviours (e.g., Leslie, 2008). Within the context of social media then, sensitive-content screens were originally designed to reduce the harms associated with being exposed to sensitive content online. Indeed, when they introduced sensitive-content screens, Instagram explained that such screening would balance out *their need* to create a safe space for people to talk about their experiences with *their responsibility* to reduce the potential harm that such content might have on other people who see it (Mosseri, 2019). However, as my thesis demonstrates, sensitive-content screens—at least in their current format—do not achieve these harm minimisation aims. In fact, the presence of sensitive-content screens may also prevent people from implementing other evidence-based strategies (e.g., cognitive emotion regulation strategies; Gross, 2015) to improve their mental health and wellbeing—which may bring about additional harms. Therefore, one solution is for social media platforms to remove sensitive-content screens from their platforms completely. Such action may encourage users to seek out professional support if they notice themselves feeling distressed when they come across sensitive content online. However, this action is unlikely because there are potential legal implications for social media platforms (e.g., Llamas, 2023)—especially, if they do not appear to be making attempts to improve the online experience for their users. Another solution is for social media platforms to make evidence-based adaptions to sensitive-content screens to improve the screens' utility as a harm minimisation tool. Indeed, such evidence-based adaptions could serve the dual purpose of reducing uninformed engagement with sensitive content, and—if/when people decide to

uncover sensitive-content screens—mitigating the impact of exposure to such content. I now discuss two possible evidence-based adaptions in turn.

Reducing Uninformed Engagement

In Study 2 (Chapter 4), I drew upon data from Studies 1a and 1b, which suggested participants uncover sensitive-content screens *because* they want to obtain information about the image. I wondered whether reducing the desire for information could reduce uncovering behaviour. Specifically, I presented participants with content-related information, in the form of brief and detailed content descriptions, and measured frequency of uncovering behaviour in each condition. People uncovered sensitive-content screens *less* often with content-related information, both brief and detailed. Participants also reported that content descriptions helped them make an informed decision about whether they should engage with the forewarned sensitive content. For example, one participant said, "There were some images I did not want to see. I appreciated the information provided that allowed me to make a more informed decision". This finding is consistent with a recent qualitative study that found people *want* contextual information alongside warnings (while avoiding overly explicit details) to help them make informed choices about their content consumption (Gupta, 2023). Therefore, reducing the desire for information by adding content-related information to sensitive-content screens can reduce uncovering behaviour, and help people make informed uncovering decisions—which aligns with what they want.

But does this reduction in exposure owing to content-related information create an emotional cost? In Studies 3a and 3b (Chapter 5), I considered two key issues. First, I tested the possibility that content-related information causes people to experience more anxiety when they view sensitive-content screens. In Study 3a, I examined whether sensitive-content screens are more anxiety provoking if they appear with brief or detailed content descriptions, compared with when they appear as they typically do on Instagram (i.e., with no description).

Consistent with the first possibility, *detailed* content descriptions increased anticipatory anxiety relative to brief content descriptions (and sensitive-content screens without content descriptions). Indeed, detailed content-related information may exacerbate anticipatory anxiety *because* the details in and of themselves may be aversive/triggering (Gupta, 2023), and/or because people have more details to imagine—which may be *as* anxiety provoking or distressing as viewing the content itself would be (see Blackwell, 2019; 2021 for review). Therefore, although detailed content descriptions can reduce uncovering behaviour and uninformed engagement with sensitive content, they seemingly come at an emotional cost. However, brief content descriptions offer the same reduction in uncovering behaviour, but do not increase people's anticipatory anxiety, relative to sensitive-content screens without content descriptions. Thus, it seems that briefer content descriptions strike an appropriate balance between providing sufficient context and avoiding overly explicit details.

Second, I tested the possibility that content-related information exacerbates the negative reactions people have to sensitive content if/when they decide to view it. In Study 3b, I examined whether participants report content as more distressing when the preceding sensitive-content screen appears with a brief (vs. no) content description. I found no evidence for the second possibility; people reported similar levels of image-related distress irrespective of whether they saw sensitive-content screens with or without brief content descriptions. This finding is consistent with research on traditional trigger warnings that finds warnings have a trivial effect on people's emotional reactions towards forewarned content (Bridgland, Jones et al., 2023). Therefore, adding brief content-related information to sensitive-content screens does not impact people's immediate reactions to sensitive content (i.e., how negative [or distressed] a person feels if they decide to uncover the screen and view the negative image).

Together, I found that adding brief content-related information to sensitive-content screens not only shifts behaviour (by minimising engagement with negative content), but

also—and perhaps more importantly—bolsters people's ability to make informed decisions about which content they want to engage with. I also found no evidence of an emotional cost associated with *brief* content-related information. Therefore, social media platforms could add brief content-related information, in the form of brief content descriptions, to sensitive-content screens to improve the screens' harm minimisation utility.

However, data from Studies 2 and 3b suggests that some people *will* still decide to uncover sensitive-content screens to view sensitive content—irrespective of whether additional content-related information accompanies the typical warning. Indeed, we know that some people intentionally seek out sensitive content because they want to experience negative affect (i.e., they have counterhedonic emotion goals), and others may be high sensation seekers and simply want to take the risk for the sake of having a novel and potentially intense experience (Zuckerman, 2007). For these people, merely providing more information about the content is unlikely to counteract uncovering behaviour. Therefore, I also investigated another adaption with the intention of mitigating the impact of exposure to sensitive content when people decide to uncover sensitive-content screens.

Mitigating the Impact of Exposure to Sensitive Content

In Studies 4a and 4b (Chapter 6), I drew upon the existing research that suggests sensitive-content screens do not help people mentally prepare for sensitive content *because* they do not help people bring coping strategies to mind (Bridgland, Barnard et al., 2022). I wondered whether providing explicit instructions detailing *how* to emotionally prepare could assist people with mental preparation. Perhaps making coping strategies more accessible—and encouraging people to engage in hedonically driven emotion regulation to down-regulate negative emotions and up-regulate positive emotions (Larsen, 2000)—could provide an emotional benefit when people view the forewarned content.

Specifically, I examined whether providing distraction (Studies 4a and 4b) and reappraisal (Study 4a) instructions on sensitive-content screens reduce participants' distress following exposure to negative images. Using a within-subjects design, I found participants reported lower image-related distress when they received reappraisal or distraction instructions, compared to no instructions (Study 4a). *And,* using a between-subjects design, I found participants who received distraction instructions reported substantially lower image-related distress than participants who received no instructions (Study 4b). These findings align with existing research showing that emotion regulation instructions reduce negative emotions in response to negative images (e.g., Ray et al., 2010; Thiruchselvam et al., 2011), but extends them to a social media context.

Notably, I observed a larger effect of distraction when I instructed participants to use one strategy the whole time (Study 4b). One explanation for this finding is that having participants switch between two regulation strategies and/or including no regulation trials (i.e., varying regulation instructions within-subjects) dampened the effects in Study 4a. Alternatively, perhaps assigning participants to *either* the distraction or no regulation instruction condition potentiated the effect of distraction in Study 4b. Having no basis for comparison regarding their levels of distress may have prompted participants in either condition to underestimate or overestimate their distress. Our Study 4b findings showed that distress was both higher in the no instruction condition *and* lower in the distraction condition compared to Study 4a, suggesting participants may have overestimated their distress in the no instruction condition and/or underestimated their distress in the distraction condition.

Nevertheless, providing explicit emotion regulation instructions appears to address some of the challenges people—especially vulnerable people—experience when they need to implement strategies for emotional preparation. However, there may be important individual differences to consider here. Mere encouragement to engage in hedonically driven emotion

regulation (i.e., to down-regulate negative emotions and up-regulate positive emotions) may be insufficient to emotionally benefit people who intentionally seek out sensitive content to fulfil counterhedonic emotion goals. For example, people who want to experience negative affect simply may not implement such emotion regulation instructions. Therefore, adding explicit instructions to use distraction and reappraisal on sensitive-content screens appears to improve the screens' utility as a harm minimisation tool for most people, but—like most harm minimisation tools—it is not a one size fits all solution.

Together, these findings suggest that social media platforms could adapt their sensitive-content screens to improve the screens' utility as a harm minimisation tool. Although any one adaption is unlikely to benefit *every* user—given that individual differences (e.g., counterhedonic emotion goals) may interfere with their effectiveness—overall, people appear to benefit from additional information and explicit emotion regulation instructions. Specifically, brief content-related information can reduce uninformed engagement with sensitive content, and distraction and reappraisal instructions can mitigate the impact of exposure to such content.

7.2 Theoretical Implications

Behavioural and Emotional Responses to Sensitive-Content Screens are Complex

My thesis forms the beginnings of a—previously non-existent—theoretical framework for understanding *how* and *why* people respond to sensitive-content screens the way they do. We now know that people uncover sensitive-content screens, irrespective of their vulnerabilities, and do so even after they view a series of sensitive images. We now also have a more sophisticated understanding of the reasons underpinning people's uncovering behaviour. Specifically, we now know that people view (and avoid) sensitive content for different reasons—so we should not simply define the decision to uncover sensitive-content screens (or not) as adaptive or maladaptive. Indeed, emotional experiences arising from

viewing sensitive content may depend on the reasons and/or emotion goals underpinning such behaviour. For example, people who seek out sensitive content as a means of filling an information gap may be less likely to experience maladaptive consequences (e.g., a worsening emotional state) then people who seek out sensitive content as a means of experiencing negative affect. Similarly, the context surrounding avoidance behaviour may be important in determining whether such behaviour is adaptive or maladaptive. Avoiding potentially distressing content when the situation is uncontrollable may be adaptive in terms of protecting people's wellbeing, but this behaviour may become maladaptive if people use it to avoid *all* negative emotions (see Hofmann & Hay, 2018). Therefore, uncovering sensitive-content screens (or not) may not warrant concern in and of itself, but behaviour underpinned by maladaptive reasons and/or emotion goals—particularly those pursued in an inflexible and context-insensitive manner (e.g., Bonanno & Burton, 2013; Kashdan & Rottenberg, 2010)—may.

Trigger Warnings in Other Contexts may also not Function as Intended

Sensitive-content screens are a unique form of trigger warning, in that they blur the forewarned content in addition to providing a warning statement, but they operate with the same *intent*—to help people avoid sensitive content or emotionally prepare for it. Therefore, not only do my findings help us understand how and why people respond to trigger warnings in a social media context, but they also apply to traditional trigger warnings. Thus, trigger warnings in other contexts (e.g., on other forms of media, such as books and podcasts) may not function as intended—perhaps because they too prompt engagement with the forewarned content and fail to counteract the motivation some people have to regulate their affect via such content.

There is a Marginal Parameter Between the Level of Information Required to Elicit Curiosity vs. Anxiety

Research suggests that the relationship between information and curiosity follows an inverted U-shaped function (e.g., Day, 1982; Kang et al., 2009). That is, people's curiosity increases as information increases, until it reaches to an optimal level (or the "zone of curiosity"; Day, 1982), characterised by exploration, and approach-driven behaviour (e.g., uncovering sensitive-content screens). But with too much information, people can experience a reduction in curiosity and enter a "zone of anxiety" whereby they are defensive, disinterested, and avoidant (e.g., of sensitive content). The "cut offs" for each zone are arbitrary, but my thesis has established some parameters for the level of information required to elicit curiosity vs. anxiety. Although both brief and detailed content descriptions reduced uncovering behaviour in Study 2—suggesting people may have moved away from the zone of curiosity, and perhaps towards the zone of anxiety—in Study 3a we found *only* detailed content descriptions increased anxiety, relative to sensitive-content screens without content descriptions. Put differently, detailed (11-15 words, e.g., "A person receives treatment for a severe burn on their hand") but not brief (1-3 words, e.g., "Burns") content descriptions appeared to move people towards the zone of anxiety. This finding suggests there is only a marginal parameter (i.e., of 8-14 words) between the level of information required to elicit curiosity vs. anxiety.

Cognitive Emotion Regulation Strategies are Effective in a Social Media Context but Require Explicit Prompting

Previous research shows that cognitive emotion regulation strategies are effective in reducing negative emotions (e.g., when viewing negative images and films; Ray et al., 2010; Thiruchselvam et al., 2011; Wolgast et al., 2011), alleviating psychological symptoms, and improving well-being (Kraiss et al., 2020; Webb et al., 2012). My thesis demonstrates that

they are also effective in a social media context, specifically for sensitive and potentially distressing content, but that people need explicit instructions about how to use these strategies when encountering such content online. That is, simply alerting someone to an impending negative affective state that may arise from viewing negative content, via a sensitive-content screen, is not sufficient to elicit emotion regulation processes. These findings may extend to other potentially negative situations in everyday life where emotion regulation strategies may be useful. For example, doctors may inform people to "stay calm", or use reappraisal, as they await their medical tests results—especially if there could be bad news (e.g., a lump returning a positive result for cancer). However, it is unlikely people will spontaneously draw upon emotion regulation strategies, such as reappraisal, unless they receive explicit instructions describing how to apply the strategy. Instead, people may begin bracing for the worst (e.g., Sweeny & Cavanaugh, 2012), which—as discussed in Chapter 1—could have negative impacts on their psychological (e.g., negative affect; Golub et al., 2009; Sweeny et al., 2016) and physiological wellbeing (e.g., increased blood pressure; Spacapan & Cohen, 1983). Future research on cognitive emotion regulation processes, more broadly, should consider the role of explicit regulation instructions in regulatory success.

7.3 Methodological Implications

A Mock Social Media Paradigm Can Investigate the Effects Warning Systems have on Behavioural and Emotional Responses

Across each of my studies, I developed (and refined) a mock social media paradigm. Behavioural trials within the mock social media paradigm allowed me to assess behaviour (e.g., the decision to view or avoid sensitive content) and emotional responses (e.g., distress) in the moment, and on consecutive trials. Only several trigger warning studies have examined behavioural avoidance of content accompanied by a warning (e.g., choosing a video title presented with or without a trigger warning; Gainsburg & Earl, 2018); and of these, some

have used dropout as an avoidance analogue (e.g., Jones et al., 2020)—which is limited because the accuracy of such an analogue is unknown. Therefore, my thesis provides a novel contribution to the broader trigger warning literature and suggests a way forward for future research on such warning systems.

Additionally, I found subtle differences in behavioural responses between time-based designs (when I fixed the time to 5-min, and participants could view and uncover as many screens/images as they liked within the time; Study 1a) and image-based designs (when I fixed the number of screens/images within the image-viewing task, and introduced an image-response delay; Studies 1b and 2). Notably, the percentage of sensitive-content screens participants uncovered was higher when I used a time-based rather than image-based design—perhaps because participants were able to rush through the task and may have not been paying as much attention to their responses. However, in real-life people are also able to move quickly from one image to the next and may pay similar attention to those images too. Nonetheless, these methodological differences highlight the importance of matching methodology with key research aims. For example, to examine how people naturally respond to warning systems (as we did in Study 1a), it would be sensible to prioritise giving participants a choice to view and uncover sensitive-content screens in their own time but, to compare behaviour between conditions (as we did in Study 2), it would be sensible to prioritise controlling for participant's exposure to screens/images.

7.4 Practical Implications

Instagram Must Revise Their Community Standards to Account for Cumulative Impact of Viewing Sensitive Content

Currently, Instagram follows a narrow set of community standards (available at the Transparency Center: https://transparency.fb.com/en-gb/), which stipulate whether content is removed or screened based on the risk of viewing a *single* piece of content. But we now

know that some people repeatedly uncover sensitive-content screens, and as such, repeatedly engage with sensitive content. Therefore, a potentially more harmful (systemic) risk to users comes from the *cumulative* impact of repeatedly viewing sensitive content—even though such content may not violate current community standards on its own (e.g., healed scars or other non-graphic self-injury imagery in a recovery context; Meta, 2023). Indeed, content moderators—who are responsible for reviewing large volumes of potentially violating content during a shift (when the sentiment of a post is unclear, or the content is context-dependent)—experience a range of negative psychological impacts (e.g., intrusive thoughts and anxiety; Spence et al., 2023). The experience of content moderation—in terms of the nature and volume of content—is arguably not dissimilar from the user experience when they find themselves in a "dark rabbit hole" (Crawford, 2019), and repeatedly engage with sensitive content. Thus, Instagram needs to revise their policies—perhaps beginning with lowering their threshold for removing sensitive content—to account for the cumulative impact of viewing such content. Other social media sites alike (e.g., TikTok) could also benefit, in terms of providing a safer environment for their users, from making similar revisions to their policies.

Instagram Must Move Beyond Merely Warning About Upcoming Content

My findings have practical implications for Instagram's sensitive-content screens. Instagram has previously argued that sensitive-content screens help protect the mental health and wellbeing of their users, but my thesis demonstrates that screens are ineffective at achieving their purported harm minimisation aims in their current format. There is a risk that continued reliance on sensitive-content screens as a harm minimisation tool becomes a "sticker-fix" (Fagan, 2019) at the expense of Instagram making other efforts to present distressing content in a conscientious and evidence-based way. On a larger scale, failure to acknowledge that the current harm minimisation tools are ineffective may prevent Instagram

from providing other wraparound mental health support (e.g., funding and developing educational programs about online safety).

Encouragingly, there are some relatively simple evidence-based adaptions that Instagram could make to sensitive-content screens to improve screens' utility as a harm minimisation tool. To minimise engagement with negative content and bolster people's ability to make informed decisions with respect to what content they want to engage with, Instagram could provide brief content-related information on sensitive-content screens, alongside the typical warning. To mitigate the impact of exposure to sensitive content—when people decide to view it—Instagram could provide explicit instructions for people to regulate their emotions. I found distraction and reappraisal were effective in reducing image-related distress, but other strategies (e.g., acceptance) may also be effective in the social media context.

Notably, I did not examine the possibility that emotion regulation instructions, in of themselves, increase curiosity and subsequent uncovering behaviour. Indeed, people may wonder why they need to emotionally regulate, and find themselves curious about the nature of the content—especially when sensitive-content screens appear in their current format (i.e., without content descriptions). However, social media platforms could circumvent such possibility by incorporating both adaptions within a single sensitive-content screen—as a two-pronged harm minimisation tool. Specifically, they could present brief content-related information alongside sensitive-content screens to begin with, to reduce uninformed engagement, and *only when* a person decides to uncover them, emotion regulation instructions could appear on the screen—before people they then go on to view the forewarned content. Not only may such a combined (and sequential) approach get around the issue of increasing uncovering behaviour, but it may have the largest (or two-pronged) effect

in terms of improving their utility as a harm minimisation tool. Nonetheless, future research is needed to examine the effectiveness of this combined (and sequential) approach.

Other Social Media Platforms Should Adapt Their Warning Systems

Other social media platforms, such as TikTok, Facebook, Twitter, Reddit and Buzzfeed, should also consider making similar evidence-based adaptions to their warning systems. These adaptions may be particularly important for platforms that use an algorithm to recommend relatable content to user. TikTok, for example, blindly recommends content to users based on what they have previously engaged with, including sensitive content, for example related to eating disorders, self-injury, and suicide (Morrison, 2022; Sung, 2020; Within Health, 2023). Because most users know how the algorithm operates (Bhandari & Bimo, 2022), there is also a risk that users who *want* to view sensitive content can manipulate the algorithm (e.g., by intentionally engaging with sensitive content) and find themselves in a harmful feed of such content. Indeed, TikTok has received scrutiny for "trapping" users in content feeds curated for their specific vulnerabilities (e.g., Morrison, 2022; Sung, 2020). Therefore, adapting their current warning systems is one of many steps needed to make TikTok a safer online environment for users. Moreover, adopting a more universal approach to warning systems *between* platforms may assist with reducing uncertainty/curiosity—and subsequent engagement with sensitive content—that can stem from seemingly novel warning systems.

Trigger Warnings in Other Contexts Should be Adapted in an Evidence-Based Way

My findings also suggest that trigger warnings in other contexts could benefit from evidence-based adaptions. Trigger warnings provided in educational contexts (e.g., on university campuses) vary in nature, but often take the form of a statement at the beginning of a lecture (e.g., 'This lecture includes reference to themes of x, y, z, which might trigger unwelcome and distressing memories or thoughts for some students'; University of Reading,

2021). This example includes brief content-related information, in the form of identifying lecture themes, but students may also benefit from explicit instructions explaining how to regulate their emotions. Given that distraction and reappraisal may interfere with comprehension of the lecture content (Gross, 2015), it may be more appropriate in this context to encourage people to use acceptance (e.g., try experiencing your emotions without judgment; Wolgast et al., 2011). Traditional trigger warnings accompanying other forms of media, such as books and podcasts, could also benefit from explicit content descriptions and/or appropriate emotion regulation instructions (if they are not already in use). However, irrespective of the context, we need to remain thoughtful about what types of content we provide trigger warnings for. Providing inappropriate trigger warnings (i.e., for content that does not require a forewarning) may have negative emotional impacts (e.g., by eliciting anticipatory anxiety)—even *if* they use the suggested evidence-based adaptions. Indeed, the type of content that should have a forewarning is still a debated issue (e.g., Johnson et al., 2015; Lukianoff & Haidt, 2015; Filipovic, 2014)—and warrants further exploration. Given "triggers" are complex and closely connected to personal vulnerabilities (Riachi et al., 2022), a more individualised approach to forewarning content may be necessary.

Marketing Teams Should Continue Using the "Teasing Effect"

My findings also provide support for the "teasing effect" (Ruan et al., 2018), which marketing teams have increasingly used in recent years to encourage consumer engagement with their products. The premise of the teasing effect is that by first creating and then resolving uncertainty, people experience a net gain in happiness, which enhances consumers' attitudes toward, willingness to try, and choice of the advertised product (Ruan et al., 2018). Although my thesis did not directly test the effect, I consistently found evidence to suggest that people like to resolve uncertainty, and will do so, even when the outcome is relatively

unknown (Studies 1a, 1b and 2). Therefore, marketing teams should continue to use this approach if they want to attract consumer attention and encourage product engagement.

7.5 Clinical Implications

Clinicians Should Consider the Reasons why People are Seeking out Sensitive Content to Inform Intervention

Clinicians working with people who seek out sensitive and potentially distressing content (e.g., people with PTSD who engage with intentional self-triggering like behaviours; Bellet et al., 2020) should assess the reasons underpinning this behaviour. Such assessment will help determine: 1) whether intervention is needed, and 2) which intervention approach is appropriate. For example, people who seek out sensitive content as a means of filling an information gap may benefit from psychoeducation on uncertainty and curiosity. With increased awareness of the cognitive processes underlying their uncovering decisions, people may be better positioned to pause when they see sensitive-content screens and evaluate their reasons for uncovering (e.g., Do I want to uncover the sensitive-content screen because I am uncomfortable not knowing what is beneath?). However, if people are seeking out sensitive content to regulate their affect, clinicians should assess their emotion goals (i.e., what clients *want* to feel) and the functions of sought after emotions (Arens & Stangier, 2020). If people want to experience unpleasant emotions (e.g., anxiety) because it allows them to avoid more painful emotions (e.g., sadness; Mees & Schmidt, 2008), they may benefit from psychoeducation on the cycle of avoidance and distress (Barlow, 2021). Whereas if people with depression want to experience sadness because it serves self-verification motives (Millgram et al., 2015), they may benefit from psychoeducation on the cycle of depression (Barlow, 2021) and need support in deemphasising the importance of self-verification. Such tailored intervention is important because we know that maladaptive emotion regulation plays a key role in the development and maintenance of a range of psychopathology (e.g., mood

disorders; see Aldao et al., 2010; Joormann & Siemer, 2014; Kring & Sloan, 2010; Werner & Gross, 2010).

Improved Harm Minimisation Tools Could Contribute to Improving Users' Mental Health and Wellbeing

We know users still have access to a considerable amount of sensitive content on Instagram (e.g., eSafety Commissioner, 2022; Molly Rose Foundation, 2023). If viewed in large amounts or cumulatively over time, this content can have harmful consequences (e.g., increasing self-injury, hopelessness, and suicide risk; Ardent, 2019; Funder & Ozer, 2019; Susi et al., 2023)—as in the case of Molly Russell, who died from "an act of self-harm while suffering from depression and the negative effects of online content" (Naughton, 2022). Therefore, improved harm minimisation tools that contribute to safer online environments for users, may have accumulating clinical implications in terms of improving users' overall mental health and wellbeing (e.g., by reducing their symptoms of psychopathology and enhancing their quality of life). Although such improvements would come too late for Molly Russell's family, there remains hope for preventing future deaths caused by the negative effects of online content.

7.6 Limitations and Future Directions

Online Data Collection

I used online data collection for all studies to increase ecological validity; specifically, I wanted participants to complete the studies on their own mobile/tablet devices and at their own convenience to simulate the situations in which people typically scroll on their social media feeds. Although participants would have been aware that they were, in fact, completing a study rather than scrolling on their own Instagram feed, online data collection was the most suitable method for simulating a real-life social media experience. However, there are downsides to using online data collection. Unlike in a traditional laboratory setting, I had no

control over distractions (e.g., phone calls) that may have interfered with participants' attention throughout their participation. Although such distractions *are* part of most real-life social media experiences—in that people can be interrupted by phone calls while scrolling on their own Instagram feed—in this case, poor attention, especially during image-viewing tasks, could have meant participants missed key experimental manipulations.

To minimise the impact of poor attention on my data I employed several strategies across all studies—recommended specifically for online data collection using MTurk (Moeck et al., 2022). I recruited MTurk participants via CloudResearch—a third-party website that interfaces with MTurk—which allowed me to exclude MTurk participants flagged for having low-quality data (using the "Block Low Quality Participants" setting; Hauser et al., 2022). I also included multiple attention checks of varying difficulty, including items embedded in existing questionnaires (e.g., "Please select between 30 and 50; this is an attention check"), and single item questions which asked participants response in a particular way (e.g., "The technology question you're about to answer is simple. When asked for your favourite technological device, select 'phone'. This is an attention check. Based on the text you read above, what device have you been asked to choose?") and excluded participants who failed at least two or more of these checks. Additionally, I asked participants to report if they stopped/left the task for any extensive length of time (and at what point)—after reminding them their honesty was important for our research, and that they would receive full compensation irrespective of their responses. Because it was important that participants paid during the image-viewing tasks in particular, I excluded participants who reported stopping or leaving during these tasks. However, I acknowledge that experienced participants are likely aware of attention checks—given researchers commonly use them in online surveys—such that they may pass the checks despite being otherwise distracted. I also acknowledge

that I could not determine whether participants were honest in reporting their task compliance.

Encouragingly though, contrary to the possibilities I have raised, several studies have demonstrated that the quality of data from MTurk participants is reliable (e.g., Buhrmester et al., 2011; Casler et al., 2013) and sometimes superior (e.g., fail less attention checks; Hauser & Schwarz, 2016) to participants sourced from more traditional subject pools (e.g., undergraduate samples). Nevertheless, it would be useful to replicate my work—especially the experimental work (from Studies 2, 3a, 3b, 4a and 4b) where I would expect to see a larger impact of inattention—in a traditional laboratory setting. Participants could still use their own handheld device, but I could oversee their task completion and limit distractions (e.g., by disabling incoming phone calls).

Finally, I crowdsourced online participants specifically from the United States. I chose this sample because I wanted to compare to prior work on sensitive-content screens (and trigger warnings more generally), which is predominantly based on such Western, educated, industrialised, rich, and democratic (or WEIRD) populations. Consequently though, I do not know how my findings would apply to non-WEIRD samples, for example, to people from different cultural and socioeconomic contexts. Therefore, readers should exercise caution when considering the generalisability of my results beyond Western contexts, and future research on sensitive-content screens should consider using more diverse samples.

Clinical Populations and Individual Differences

Despite their now widespread use across social media platforms (e.g., Facebook, YouTube, TikTok), the target audience for sensitive-content screens fundamentally remains the same—people who may be particularly vulnerable or "triggered" by viewing sensitive content. Therefore, another limitation of my work is that I did not recruit participants with a formal diagnosis (e.g., of PTSD/MDD) *or* use semi-structured diagnostic interviews to

formally diagnose participants (e.g., Clinician Administered PTSD Scale for DSM-5: CAPS-5; Weathers et al., 2013; Structured Clinical Interview for DSM-5: SCID-5; First, 2016). Therefore, I was not able to objectively examine the presence *or* severity of mental health disorder symptoms for each participant. Resource constraints largely contributed to this decision: it can take 45 to 90 minutes to complete such an interview, and I had limited funding available to recruit the desired number of participants *and* compensate them fairly. Instead, I used questionnaires (e.g., the PCL-5) to assess the probability that participants would qualify for a formal diagnosis (e.g., of PTSD), or were experiencing psychological distress associated with mental health disorders. Although research has consistently demonstrated that the PCL-5 has good convergent validity with more robust clinical tools, such as the CAPS-5 (Bovin et al., 2016), these self-report symptom questionnaires may be biased in some way. Therefore, my results may have differed had I specifically recruited specific clinical populations or used more robust clinical tools. Bearing this limitation in mind, research suggests that MTurk is an excellent platform for studying clinical and subclinical populations: the prevalence of mental health disorders in MTurk populations matches or exceeds that of the general population, and clinical measures taken from MTurk participants demonstrate high reliability and validity (Shapiro et al., 2013).

Relatedly, although I explored behavioural and emotional responses for some vulnerable populations (e.g., people with probable PTSD, depression, a history of self-triggering; Studies 1a and 1b) there are other noteworthy populations. For example, people in non-suicidal self-injury communities—who share visual content related to their experiences of self-injury, and more generally, mental health issues—may seek out graphic and non-graphic self-injury related content as a means of connecting with likeminded users, but also to experience negative affect (which may serve self-verification motives; Fulcher et al., 2020; Moreno et al., 2016). Additionally, content related to eating disorders (e.g., images promoting

extreme weight loss) is often screened to help people avoid "triggers" of disordered eating behaviours (e.g., Cripps, 2020). But, anecdotal reports suggest that people with eating disorders may maladaptively use sensitive-content screens to find content that motivates/encourages them to lose more weight (Hack, 2017). Despite users in these populations being amongst the most vulnerable, no research has specifically investigated *how* they use and respond to sensitive-content screens and *why*. Therefore, future research should address this important research gap.

Additionally, across all studies I assessed participants' preferences for sensitive content and previous experience with sensitive content screens; in Study 1b, I also assessed particiapants' trait curiosity, ability to regulate negative and positive emotions, psychological flexibility, trait experiential avoidance, and intolerance to uncertainty (which I measured again in Study 2). But, there are other noteworthy individual differences. For example, people's emotion goals (Tamir, 2016), and their tendancy to engage in sensation seeking (Zuckerman, 2007), may influence their behavioural and emotional responses. Future research should consider broader individual differences to develop our understanding of how people with varying characteristics respond to sensitive-content screens and forewarned content.

Finally, I collected data from adults aged 18-76, with a mean age of roughly 36. Although these samples were suitable for the purpose of my thesis aims—and I did not have ethics approval to recruit adolescents in the present work—future research should examine how adolescents respond to sensitive-content screens with and without brief content descriptions and emotion regulation instructions. Indeed, adolescents (aged 10-17) regularly use social media platforms (eSafety Commissioner, 2022), and are susceptible to social influence (Ahmed et al., 2020) and risk taking (Steinberg, 2008) online. Therefore, they may be even more likely (than our sample) to seek out sensitve content, for example, as a means

of attaining peer approval. Additionally, some adolescents are still developing their emotional awareness (i.e., their ability to identify, explain, and differentiate emotional experiences; Lane & Schwartz, 1987)—which makes them even more susceptible to having underdeveloped emotion regulation strategies (Van Beveren et al., 2019). Therefore, their emotional reactions to sensitive-content screens, especially screens without explicit instructions to regulate their emotions, may also vary from our sample. Indeed, adolescents may benefit the *most* from emotion regulation instructions. Increasing our understanding of adolescents' behaviour and emotional reactions in this context is an important next step in informing evidence-based adaptions for sensitive-content screens that provide a harm minimisation benefit to adolescents too.

Narrow Definition of Deterrence and Emotional Reactions

In Studies 1a, 1b, 2, and 3a, I examined whether sensitive-content screens (with and without content-related information) deter people from viewing sensitive-content using the decision *not* to uncover sensitive-content screens as an *indicator* of deterrence. Although this decision suggests that participants were deterred from viewing sensitive content, it is arguably a narrow definition of deterrence. In this situation, deterrence may encompass a broader range of behavioural, emotional, and cognitive reactions. For example, someone could initially decide to view the forewarned content but then look away (as some participants reported doing in Studies 4a and 4b; e.g.,"Some of [the images] were too much I had to glance away for a second"), focus on down-regulating distressing emotions by looking at less distressing parts of the content (i.e., using attentional deployment; e.g., "I would look off center if the image was of a child"), or cease viewing the content altogether (by moving to another piece of content or completely closing down the platform; e.g., "Some of the images were too gruesome for me to continue viewing"). They may also suppress their thoughts and feelings about the content during and/or after exposure—a form of emotional regulation

termed response modulation (Gross, 2015). Thus, although uncovering behaviour was similar irrespective of vulnerabilities, such behavioural, emotional, and cognitive responses may differ for people with more severe psychopathology—and may be where issues with emotion regulation arise. Future research could examine whether sensitive-content screens prompt these other forms of avoidance—and how they differ for people with varying psychopathology.

Relatedly, in Studies 3a, 3b, 4a and 4b, I examined how content-related information and emotion regulation instructions influenced participant's emotional reactions. But state anxiety and distress may have been too narrow an operationalisation to determine the true emotional impact of these harm minimisation tools. Indeed, content-related information and emotion regulation instructions may differentially impact other emotional reactions; for example, perhaps emotion regulation instructions also reduce the frequency and/or intensity of intrusions people later have about the image. However, because state anxiety and distress are sensitive to subtle state changes, they are considered reliable indicators of emotional reactions and are commonly used in psychological research (and in clinical settings, e.g., Marteau & Bekker, 1992; Benjamin et al., 2010). Nonetheless, future research could explore broader emotional reactions.

The Longer-Term Impacts of Sensitive-Content Screens on Behaviour and Emotional Responses

My thesis examined how participants responded to sensitive-content screens over a series of images; specifically, *immediately* after they saw the screens and the forewarned content. But we still do not know how people's behaviour and emotional responses vary over time (e.g., with subsequent exposures to the same type of content). Sensitive-content screens may provide little immediate benefit in terms of reducing negative emotional responses, but with subsequent exposures to the same type of content, change how people feel and respond.

To address this research gap, future research could employ a longitudinal design and have people view sensitive-content screens preceding the same (or similar) type of content (e.g., images depicting injured animals) on multiple occasions and see whether their behavioural and emotional responses change with subsequent exposures. Notably, research is emerging within the traditional trigger warning literature looking at the longer-term impacts of different warning types (i.e., anticipating neutral, positive, negative emotional reactions) on PTSD symptoms and distress (at Day 1, 2 and 14; Kimble et al., 2022)—though in this case, participants were not re-exposed to any content.

We also do not know whether people's responses change depending on the time of day, and/or their current emotional state. For example, people may be particularly susceptible to uncovering sensitive-content screens during the evening, perhaps because they are more emotionally vulnerable to negative emotions (e.g., sadness, boredom, and anger) at this time of the day (English & Carstensen, 2014). Indeed, anecdotally, "doom scrolling"—the tendency to scroll through negative content online, even though that content is saddening, disheartening, or depressing (Rodrigues, 2022; Sharma et al., 2022)—is especially common during the evening. To address this research gap, future research could use an experience-sample method across an extended (e.g., 7 day) period and have participants record their daily affect, encounters with sensitive-content screens, and their behavioural and emotional responses. Alternatively, for more experimental control, researchers could manipulate mood using a mood induction, and then examine how differences in mood change how people respond to sensitive-content screens.

Relatedly, we do not know the longer term or cumulative effects of the suggested adaptions. In Study 2, I found evidence to suggest that content-related information can reduce uncovering behaviour, but does the effect of content-related information persist beyond a single social media sitting? Or does uncertainty/curiosity get the better of people once they

have stopped uncovering sensitive-content screens for a while? Moreover, the cognitive resources required for regulating emotions via emotion regulation strategies can decrease with practice (e.g., Scheibe & Blanchard-Fields, 2009), so do the strategies become increasingly effective with time (because less cognitive input in required)? Or is there a ceiling effect? If the latter is true, do the strategies have a cumulative effect, whereby the emotional benefit is maintained from one social media sitting to the next, such that people experience progressively lower distress? Future research should address these remaining questions.

Contextual Information

Qualitative data from Studies 1a and 1b revealed that contextual elements (like the posting account name, captions, and comments) may be important factors in uncovering behaviour. Since I did not include these elements, I cannot determine what influence they may have on behaviour or generalise my results to situations where they are present. However, at present there is no standardised approach to captioning content on Instagram—irrespective of whether the content is screened or not. Indeed, content with sensitive-content screens is often posted with ambiguous or unclear captions. Nonetheless, future research could replicate our key findings using "The Misinformation Game" (Butler et al., 2023)—a new, easily adaptable, open-source online testing platform that simulates key characteristics of social media. Researchers can customise posts (e.g., images, videos), source information (e.g., posting account), and engagement information (e.g., number of likes and comments)—to determine what influence they have on behaviour. The platform also allows participants to respond to content (e.g., by liking, sharing, commenting), which could provide a means for researchers to explore broader responses to sensitive-content screens.

Content Type and Format

The images I used for all studies were negative in nature and—according to their guidelines, available at the Transparency Center: https://transparency.fb.com/en-gb/—Instagram would likely screen them. But I did not include non-graphic self-injury related content (e.g., depicting older instances of self-injury such as healed cuts) and eating disorder related content (e.g., depicting ribs, collar bones and thigh gaps)—which is commonly screened on Instagram (Meta, 2023). Relatedly, I did not match the content type to people's specific vulnerabilities. Some people report avoiding content specifically relevant to their trauma that could be triggering; for example, in one study, a participant with suicidal tendencies reported avoiding content related to suicide as part of their ongoing recovery efforts, but viewing sensitive content that they did not personally relate to (Gupta, 2023). Finally, here I examined sensitive-content screens for images, but social media platforms—and especially video-based platforms such as TikTok—also screen negative and potentially distressing *videos*. Although it is reasonable to infer that people would behave and respond similarly to sensitive-content screens proceeding videos, there may be something unique about videos that changes how people respond. For example, the dynamic and interactive nature of videos, compared to static images, may make them more emotionally engaging and prompt stronger emotional responses. Future research is needed to better understand how people respond to sensitive-content screens for different types of content, specifically content that is personally relevant to them, as well as different forms of media.

7.7 Conclusion

My thesis aimed to fill existing research gaps by examining behavioural and emotional responses to sensitive-content screens. Overall, sensitive-content screens do not deter people from viewing sensitive content; rather, they may *increase* engagement with sensitive content because they prompt information seeking behaviour. I also found sensitive-

content screens do not help people emotionally prepare for sensitive content. In fact, sensitive-content screens appear to create a noxious anticipatory period that does not translate to an emotional benefit when participants view the forewarned content—perhaps because people may not have strategies for emotional preparation and/or their emotion goals do not align with down-regulating negative affect. However, my findings also suggest that adapting sensitive-content screens can improve the screens' utility as a harm minimisation tool. Specifically, adding brief content-related information and emotion regulation instructions to sensitive-content screens can reduce uninformed engagement with sensitive content and mitigate the impact of viewing such content. Although many remaining questions warrant further investigation, it is evident that social media platforms should not rely upon sensitive-content screens in their current format to provide harm minimisation benefits. In fact, doing so would be a failure of social media platforms to meet their responsibility to protect users' mental health and wellbeing, and may ultimately come at the cost of more lives.

References

Abramowitz, J. S., Deacon, B. J., & Whiteside, S. P. (2019). Exposure therapy for anxiety: Principles and practice. Guilford Publications.

Ahmed, S., Foulkes, L., Leung, J. T., Griffin, C., Sakhardande, A., Bennett, M., Dunning, D. L., Griffiths, K., Parker, J., Kuyken, W., Williams, J. M. G., Dalgleish, T., & Blakemore, S. J. (2020). Susceptibility to prosocial and antisocial influence in adolescence. *Journal of Adolescence*, *84*, 56-68. https://doi.org/https://doi.org/10.1016/j.adolescence.2020.07.012

Aiken, L. S., & West, S. G. (1993). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Aldao, A., Nolen-Hoeksema, S., & Schweizer, S. (2010). Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical Psychology Review*, *30*, 217-237. https://doi.org/10.1016/j.cpr.2009.11.004

Anvari, F, Kievit, R., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. (2022). Not all effects are indispensable: Psychological Science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*. *18*, 503-507. https://doi.org/10.1177/17456916221091565

Arendt, F., Scherr, S., & Romer, D. (2019). Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. *New Media & Society*, *21*, 2422-2442. https://doi.org/10.1177/1461444819850106

Arens, E. A., & Stangier, U. (2020). Sad as a Matter of Evidence: The Desire for Self-Verification Motivates the Pursuit of Sadness in Clinical Depression. *Frontiers in Psychology*, *11*, 238-238. https://doi.org/10.3389/fpsyg.2020.00238

Badour, C. L., Blonigen, D. M., Boden, M. T., Feldner, M. T., & Bonn-Miller, M. O. (2012). A longitudinal test of the bi-directional relations between avoidance coping and PTSD

severity during and after PTSD treatment. *Behaviour Research and Therapy*, 50, 610-616. https://doi.org/10.1016/j.brat.2012.06.006

Barlow, D. H. (Ed.). (2021). *Clinical handbook of psychological disorders: A step-by-step treatment manual*. Guilford Publications.

Barsky, A. J., Saintfort, R., Rogers, M. P., & Borus, J. F. (2002). Nonspecific medication side effects and the nocebo phenomenon. *JAMA*, *287*, 622-627. https://doi.org/10.1001/jama.287.5.622

Bartels, D. J. P., Van Laarhoven, A. I. M., Haverkamp, E. A., Wilder-Smith, O. H., Donders, A. R. T., Van Middendorp, H., Van De Kerkhof, P. C. M., & Evers, A. W. M. (2014). Role of conditioning and verbal suggestion in placebo and nocebo effects on itch. *PLOS ONE*, *9*, e91727. https://doi.org/10.1371/journal.pone.0091727

Bartsch, A., & Mares, M.-L. (2014). Making sense of violence: Perceived meaningfulness as a predictor of audience interest in violent media content. *Journal of Communication*, *64*, 956-976. https://doi.org/10.1111/jcom.12112

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48. https://doi.org/10.18637/jss.v067.i01

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*, 323-370. https://doi.org/10.1037/10892680.5.4.323

Bazeley, K., & Jackson, P. (2013). *Qualitative data analysis with NVivo*. Sage.

Beck, A. T., & Alford, B. A. (2009). *Depression: Causes and treatment*. University of Pennsylvania Press.

Bellet, B. W., Jones, P. J., Meyersburg, C. A., Brenneman, M. M., Morehead, K. E., & McNally, R. J. (2020). Trigger warnings and resilience in college students: A

preregistered replication and extension. *Journal of Experimental Psychology: Applied*, *26*, 717–723. https://doi.org/10.1037/xap0000270

Bellet, B., Jones, P., & McNally, R. (2020). Self-Triggering? An exploration of individuals who seek reminders of trauma. *Clinical Psychological Science, 8*, 739-755. https://doi.org/10.1177/2167702620917459

Benjamin, C. L., O'Neil, K. A., Crawley, S. A., Beidas, R. S., Coles, M., & Kendall, P. C. (2010). Patterns and predictors of subjective units of distress in anxious youth. *Behavioural and Cognitive Psychotherapy*, *38*, 497-504. https://doi.org/10.1017/s1352465810000287

Berlyne, D. E. (1954). A theory of human behaviour. *The British Journal of Psychology, 43*, 180-191. https://doi.org/10.1111/j.2044-8295.1954.tb01243.x

Berntsen, D., & Jacobsen, A. S. (2008). Involuntary (spontaneous) mental time travel into the past and future. *Consciousness and Cognition*, *17*, 1093-1104. https://doi.org/https://doi.org/10.1016/j.concog.2008.03.001

Berntsen, D., & Rubin, D. C. (2006). The centrality of event scale: A measure of integrating a trauma into one's identity and its relation to post-traumatic stress disorder symptoms. *Behaviour Research & Therapy, 44*, 219–231. https://doi.org/10.1016/j.brat.2005.01.009

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, *78*, 161-188. https://doi.org/https://doi.org/10.1111/j.1751-5823.2010.00112.x

Bhandari, A., & Bimo, S. (2022). Why's everyone on TikTok now? The algorithmised self and the future of self-making on social media. *Social Media + Society*, *8*, 20563051221086241. https://doi.org/10.1177/20563051221086241

Bijvank, M. N., Konijn, E. A., Bushman, B. J., & Roelofsma, P. H. (2009). Age and violent-content labels make video games forbidden fruits for youth. *Pediatrics*, *123*, 870-876. https://doi.org/10.1542/peds.2008-0601

Blackwell, S. E. (2019). Mental imagery: From basic research to clinical practice. *Journal of Psychotherapy Integration*, *29*, 235. https://doi.org/10.1037/int0000108

Blackwell, S. E. (2021). Mental imagery in the science and practice of cognitive behaviour therapy: Past, present, and future perspectives. *International Journal of Cognitive Therapy*, *14*, 160-181. https://doi.org/10.1007/s41811-021-00102-0

Bonanno, G. A., & Burton, C. L. (2013). Regulatory flexibility: An individual differences perspective on coping and emotion regulation. *Perspectives on Psychological Science*, *8*, 591-612. https://doi.org/10.1177/1745691613504116

Bond, F., Hayes, S., Baer, R., Carpenter, K., Guenole, N., Orcutt, H., Waltz, T., & Zettle, R. (2011). Preliminary psychometric properties of the acceptance and action questionnaire - ii: A revised measure of psychological flexibility and acceptance. *Behaviour Therapy*, *42*. https://doi.org/10.1037/t11921-000

Bonn, S. (2016). Why We Are Drawn to True Crime Shows. *Time*. https://time.com/4172673/true-crime-allure/

Bovin, M. J., Marx, B. P., Weathers, F. W., Gallagher, M. W., Rodriguez, P., Schnurr, P. P., & Keane, T. M. (2015). Psychometric properties of the PTSD Checklist for Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (PCL-5) in Veterans. *Psychological Assessment, 28*, 1379-1391. https://doi.org/10.1037/pas0000254

Boysen, G. A., Isaacs, R. A., Tretter, L., & Markowski, S. (2021). Trigger warning efficacy: The impact of warnings on affect, attitudes, and learning. *Scholarship of Teaching & Learning in Psychology, 7*, 39-52. https://doi.org/10.1037/stl0000150

Bradley, M. M., & Lang, P. J. (2007). Emotion and motivation. In J. T. Cacioppo, L. G. Tassinary & G. Berntson G. (Eds.), *Handbook of psychophysiology*, 3rd ed. (pp. 581-607). Cambridge University Press. https://doi.org/10.1017/CBO9780511546396.025

Brehm, J. W. (1966). *A theory of psychological reactance*. Academic Press.

Bridgland, V. M. E., & Takarangi, M. K. T. (2022). Something distressing this way comes: The effects of trigger warnings on avoidance behaviours in an analogue trauma task. *Behaviour Therapy*, *53*, 414-427. https://doi.org/10.1016/j.beth.2021.10.005

Bridgland, V. M. E., Barnard, J. F., & Takarangi, M. K. T. (2022). Unprepared: Thinking of a trigger warning does not prompt preparation for trauma-related content. *Journal of Behaviour Therapy and Experimental Psychiatry*, *75*. 101708. https://doi.org/10.1016/j.jbtep.2021.101708

Bridgland, V. M. E., Bellet, B. W., & Takarangi, M. K. T. (2022). Curiosity disturbed the cat: Instagram's sensitive-content screens do not deter vulnerable users from viewing distressing content. *Clinical Psychological Science*. *11*, 290-307. https://doi.org/10.1177/216770262210976

Bridgland, V. M. E., Green, D. M., Oulton, J. M., & Takarangi, M. K. T. (2019). Expecting the worst: Investigating the effects of trigger warnings on reactions to ambiguously themed photos. *Journal of Experimental Psychology: Applied, 25*, 602-617. https://doi.org/10.1037/xap0000215

Bridgland, V. M. E., Jones, P. J., & Bellet, B. W. (2023). A meta-analysis of the efficacy of trigger warnings, content warnings, and content notes. *Clinical Psychological Science, 0.* https://doi.org/10.1177/21677026231186625

Bruce, M., & Roberts. D. (2020). Trigger warnings for abuse impact reading comprehension in students with histories of abuse. *College Student Journal*, *54*, 157-168.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*. https://doi.org/10.5334/joc.72

Buhr, K., & Dugas, M. J. (2002). The intolerance of uncertainty scale: Psychometric properties of the English version. *Behaviour Research and Therapy*, 40, 931-945. https://doi.org/10.1016/S0005-7967(01)00092-4

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3-5. https://doi.org/10.1177/1745691610393980

Bushman, B. J. (2006). Effects of warning and information labels on attraction to television violence in viewers of different ages. *Journal of Applied Social Psychology*, *36*, 2073-2078. https://doi.org/10.1111/j.0021-9029.2006.00094.x

Bushman, B., & Cantor, J. (2003). Media ratings for violence and sex: Implications for policymakers and parents. *The American Psychologist*, *58*, 130-141. https://doi.org/10.1037/0003-066X.58.2.130

Butler, L. H., Lamont, P., Wan, D. L. Y., Prike, T., Nasim, M., Walker, B., Fay, N., & Ecker, U. K. H. (2023). The (Mis)Information Game: A social media simulator. *Behaviour Research Methods*. https://doi.org/10.3758/s13428-023-02153-x

Campion, N., Martins, D., & Wilhelm, A. (2009). Contradictions and predictions: Two sources of uncertainty that raise the cognitive interest of readers. *Discourse Processes*, 46, 341-368. https://doi.org/10.1080/01638530802629125

Carleton, R. N., Fetzner, M. G., Hackl, J. L., & McEvoy, P. (2013). Intolerance of uncertainty as a contributor to fear and avoidance symptoms of panic attacks. *Cognitive Behaviour Therapy*, 42, 328-341. https://doi.org/10.1080/16506073.2013.792100

Carleton, R. N., Mulvogue, M. K., Thibodeau, M. A., McCabe, R. E., Antony, M. M., & Asmundson, G. J. (2012). Increasingly certain about uncertainty: Intolerance of uncertainty across anxiety and depression. *Journal of Anxiety Disorders*, *26*, 468-479. https://doi.org/10.1016/j.janxdis.2012.01.011

Carleton, R. N., Norton, M. P. J., & Asmundson, G. J. (2007). Fearing the unknown: A short version of the intolerance of uncertainty scale. *Journal of Anxiety Disorders*, 21, 105-117. https://doi.org/10.1016/j.janxdis.2006.03.014

Carleton, R. N., Weeks, J. W., Howell, A. N., Asmundson, G. J., Antony, M. M., & McCabe, R. E. (2012). Assessing the latent structure of the intolerance of uncertainty construct: An initial taxometric analysis. *Journal of Anxiety Disorders*, *26*, 150-157. https://doi.org/10.1016/j.janxdis.2011.10.006

Carlson, E. B., Smith, S. R., Palmieri, P. A., Dalenberg, C., Ruzek, J. I., Kimerling, R., Burling, T. A., & Spain, D. A. (2011). Development and validation of a brief self-report measure of trauma exposure: the Trauma History Screen. *Psychological Assessment, 23*, 463-477. https://doi.org/10.1037/a0022294

Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. J. (2014). The use of triangulation in qualitative research. *Oncology Nursing Forum*, *41*, 545-547. https://doi.org/10.1188/14.Onf.545-547

Carver, C. S., & Connor-Smith, J. (2010). Personality and coping. *Annual Review of Psychology*, *61*, 679-704. https://doi.org/10.1146/annurev.psych.093008.100352

Casler, K., Bickel, L., & Hackett, E. (2013, 2013/11/01/). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioural testing. *Computers in Human Behaviour*, *29*, 2156-2160. https://doi.org/https://doi.org/10.1016/j.chb.2013.05.009

Crawford, A. (2019). *Instagram 'helped kill my daughter'*. https://www.bbc.com/news/av/uk46966009/instagram-helped-kill-my-daughter

Cripps, C. (2020). The Crown was right to warn viewers of Princess Diana's bulimia – it gave me the chance to prepare myself. *Independent.* https://www.independent.co.uk/arts-entertainment/netflix/the-crown/princess-diana-bulimia-the-crown-b1721340.html

Crouch, T. A., Lewis, J. A., Erickson, T. M., & Newman, M. G. (2017). Prospective investigation of the contrast avoidance model of generalized anxiety and worry. *Behaviour Therapy, 48*, 544-556. https://doi.org/10.1016/j.beth.2016.10.001

Day, H. I. (1982). Curiosity and the interested explorer. *Performance & Instruction*, 21, 19–22. https://doi.org/10.1002/pfi.4170210410

Degroote, C., Schwaninger, A., Heimgartner, N., Hedinger, P., Ehlert, U., & Wirtz, P. H. (2020). Acute stress improves concentration performance: Opposite effects of anxiety and cortisol. *Experimental Psychology*, *67*, 88-98. https://doi.org/10.1027/1618-3169/a000481

English, T., & Carstensen, L. L. (2014). Emotional experience in the mornings and the evenings: consideration of age differences in specific emotions by time of day. *Frontiers in Psychology*, *5*, 185. https://doi.org/10.3389/fpsyg.2014.00185

Engs, R., & Hanson, D. J. (1989). Reactance theory: A test with collegiate drinking. *Psychological Reports*, *64*, 1083-1086. https://doi.org/10.2466/pr0.1989.64.3c.1083

eSafety Commissioner. (2022). *Mind the Gap: Parental awareness of children's exposure to risks online*. Aussie Kids Online, Melbourne: eSafety Commissioner.

Fagan, A. (2019). Do Trigger Warnings Actually Work? *Psychology Today*. https://www.psychologytoday.com/au/blog/brainstorm/201904/do-trigger-warnings-actually-work

Faul, F., Erdfelder, E., Lang, AG., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioural, and biomedical sciences. *Behaviour Research Methods*, 39, 175-191. https://doi.org/10.3758/bf03193146

Feldhege, J., Moessner, M., & Bauer, S. (2021). Detrimental effects of online pro–eating disorder communities on weight loss and desired weight: Longitudinal observational study. *Journal of Medical Internet Research*, *23*, e27153. https://doi.org/10.2196/27153

Ferguson, M., Bargh, J., & Nayak, D. (2005). After-affects: How automatic evaluations influence the interpretation of subsequent, unrelated stimuli. *Journal of Experimental Social Psychology*, *41*, 182–191. http://dx.doi.org/10.1016/j.jesp.2004.05.008

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics.* Sage Publications.

Filipovic, J. (2014). We've gone too far with 'trigger warnings'. *The Guardian*. https://www.theguardian.com/commentisfree/2014/mar/05/trigger-warnings-can-be-counterproductive

First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2016). User's guide for the SCID-5-CV Structured Clinical Interview for DSM-5® disorders: Clinical version. American Psychiatric Publishing, Inc.

Fulcher, J. A., Dunbar, S., Orlando, E., Woodruff, S. J., & Santarossa, S. (2020). #selfharn on Instagram: understanding online communities surrounding non-suicidal self-injury through conversations and common properties among authors. *DIGITAL HEALTH*, *6*, 1-13. https://doi.org/10.1177/2055207620922389

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods & Practices in Psychological Science, 2*, 156–168. https://doi.org/10.1177/2515245919847202

Gainsburg, I., & Earl, A. (2018). Trigger warnings as an interpersonal emotion-regulation tool: Avoidance, attention, and affect depend on beliefs. *Journal of Experimental Social Psychology*, *79*, 252-263. https://doi.org/https://doi.org/10.1016/j.jesp.2018.08.006

Geers, A. L., & Lassiter, G. D. (2005). Affective assimilation and contrast: Effects of expectations and prior stimulus exposure. *Basic and Applied Social Psychology*, *27*, 143–154. https://doi.org/10.1207/s15324834basp2702_5

Golman, R., Loewenstein, G., Molnar, A., & Saccardo, S. (2021). The demand for, and avoidance of, information. *Management Science*. https://doi.org/10.1287/mnsc.2021.4244

Golub, S. A., Gilbert, D. T., & Wilson, T. D. (2009). Anticipating one's troubles: The costs and benefits of negative expectations. *Emotion*, *9*, 277-281. https://doi.org/10.1037/a0014716

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2, 271-299. https://doi.org/10.1037/1089-2680.2.3.271

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry*, *26*, 1-26. https://doi.org/10.1080/1047840X.2014.940781

Gross, J. J. (2015). The extended process model of emotion regulation: Elaborations, applications, and future directions. *Psychological Inquiry*, *26*, 130-137. https://doi.org/10.1080/1047840x.2015.989751

Gunnarsson, N. V. (2020). The self-perpetuating cycle of shame and self-injury. *Humanity & Society*, *45*, 313-333. https://doi.org/10.1177/0160597620904475

Gupta, M. (2023). understanding social media users' perceptions of trigger and content warnings. faculty of the Virginia Polytechnic Institute and State University. Blacksburg, Virginia.

Hack. (2017). How trigger warnings could harm some people living with an eating disorder. *Hack*. https://www.abc.net.au/triplej/programs/hack/how-trigger-warnings-could-harm-people-living-with-an-eating-di/8728784

Hackett, S. (2023). The complete timeline of Instagram updates that have changed the way we gram. https://blog.kicksta.co/the-complete-timeline-of-instagram-updates/

Hahn, R. A. (1997). The nocebo phenomenon: Concept, evidence, and implications for public health. *Preventive Medicine*, *26*, 607-611. https://doi.org/https://doi.org/10.1006/pmed.1996.0124

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behaviour Research Methods*, *48,* 400-407. https://doi.org/10.3758/s13428-015-0578-z

Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2022). Evaluating Cloud Research's Approved Group as a solution for problematic data quality on MTurk. *Behaviour Research Methods*, *55*, 3953-3964. https://doi.org/10.3758/s13428-022-01999-x

Havranek, M. M., Bolliger, B., Roos, S., Pryce, C. R., Quednow, B. B., & Seifritz, E. (2015). Uncontrollable and unpredictable stress interacts with subclinical depression and anxiety scores in determining anxiety response. *The International Journal on the Biology of Stress, 19*, 53-62. https://doi.org/10.3109/10253890.2015.1117449

Hay, M. (2019). Do Trigger Warnings Actually Work? *Vice.* https://www.vice.com/en/article/wj9ba4/do-trigger-warnings-actually-work

Hayes-Skelton, S. A., & Eustis, E. H. (2020). Experiential avoidance. In *Clinical handbook of fear and anxiety: Maintenance processes and treatment mechanisms*. (pp. 115-131). American Psychological Association. https://doi.org/10.1037/0000150-007

Hetrick, S. E., Subasinghe, A., Anglin, K., Hart, L., Morgan, A., & Robinson, J. (2020). Understanding the needs of young people who engage in self-harm: A qualitative investigation. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.02916

Hoffner, C. A., Fujioka, Y., Ye, J., & Ibrahim, A. G. S. (2009). Why we watch: Factors affecting exposure to tragic television news. *Mass Communication and Society*, *12*, 193-216. https://doi.org/10.1080/15205430802095042

Hofmann, S. G., & Hay, A. C. (2018). Rethinking avoidance: Toward a balanced approach to avoidance in treating anxiety disorders. *Journal of Anxiety Disorders*, *55*, 14-21. https://doi.org/10.1016/j.janxdis.2018.03.004

Hsee, C. K., & Ruan, B. (2016). The pandora effect: The power and peril of curiosity. *Psychological Science*, 27, 659-666. https://doi.org/10.1177/0956797616631733

Hsee, C. K., & Ruan, B. (2016). The Pandora Effect: The power and peril of curiosity. *Psychological Science*, *27*, 659-666. https://doi.org/10.1177/0956797616631733

Hsee, C. K., & Ruan, B. (2016). The pandora effect: The power and peril of curiosity. *Psychological Science*, 27, 659-666. https://doi.org/10.1177/0956797616631733

Huang, J., Xu, D., Peterson, B. S., Hu, J., Cao, L., Wei, N., Zhang, Y., Xu, W., Xu, Y., & Hu, S. (2015). Affective reactions differ between Chinese and American healthy young adults: A cross-cultural study using the international affective picture system. *BMC Psychiatry*, *15*. https://doi.org/10.1186/s12888-015-0442-9

Hyland, M., & Birrell, J. (1979). Government health warnings and the "boomerang" effect. *Psychological Reports*, *44*, 643-647. https://doi.org/10.2466/pr0.1979.44.2.643

Instagram. (2021). Introducing Sensitive-Content Control. *Instagram.* https://about.instagram.com/blog/announcements/introducing-sensitive-content-control

Instagram. (2024). New protections to give teens more age-appropriate experiences on our apps. *Instagram*. https://about.instagram.com/blog/announcements/giving-teens-age-appropriate-experiences

Jepma, M., Verdonschot, R., van Steenbergen, H., Rombouts, S., & Nieuwenhuis, S. (2012). Neural mechanisms underlying the induction and relief of perceptual curiosity. *Frontiers in Behavioural Neuroscience*, 6. https://doi.org/10.3389/fnbeh.2012.00005

Johnson, K. L., Tanika; Monroe, Elizabeth; Wang, Tracey. (2015). Our identities matter in core classrooms. https://www.columbiaspectator.com/opinion/2015/04/30/our-identities-matter-core-classrooms/?rate=DHDTp8vVndr1wdh8dH4QM-CaruHV1cJ3wWefi15C4o8

Jones, P. J., Bellet, B. W., & McNally, R. J. (2020). Helping or harming? The effect of trigger warnings on individuals with trauma histories. *Clinical Psychological Science*, *8*, 905-917. https://doi.org/10.1177/2167702620921341

Joormann, J., & Siemer, M. (2014). *Emotion regulation in mood disorders*. In Handbook of emotion regulation, 2nd ed. (pp. 413-427). The Guilford Press.

Joormann, J., Siemer, M., & Gotlib, I. H. (2007). Mood regulation in depression: Differential effects of distraction and recall of happy memories on sad mood. *Journal of Abnormal Psychology*, *116*, 484-490. https://doi.org/10.1037/0021-843X.116.3.484

Juarascio, A. S., Shoaib, A., & Timko, C. A. (2010). Pro-eating disorder communities on social networking sites: a content analysis. *Eating Disorders*, *18*, 393-407. https://doi.org/10.1080/10640266.2010.511918

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31-36. https://doi.org/10.1007/BF02291575

Kalokerinos, E. K., Moeck, E. K., Rummens, K., Meers, K., & Mestdagh, M. (2023). Ready for the worst? Negative affect in anticipation of a stressor does not protect against affective reactivity. *Journal of Personality*, 91, 1123–1139. https://doi.org/10.1111/jopy.12787

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, *20*, 963-973. https://doi.org/10.1111/j.1467-9280.2009.02402.x

Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review*, *30*, 865-878. https://doi.org/10.1016/j.cpr.2010.03.001

Kashdan, T. B., Barrios, V., Forsyth, J. P., & Steger, M. F. (2006). Experiential avoidance as a generalized psychological vulnerability: Comparisons with coping and emotion regulation strategies. *Behaviour Research & Therapy, 44*, 1301-1320. https://doi.org/10.1016/j.brat.2005.10.003

Kashdan, T. B., Disabato, D. J., Goodman, F. R., & McKnight, P. E. (2020). The five-dimensional curiosity scale revised (5DCR: Briefer subscales while separating overt and covert social curiosity. *Personality and Individual Differences*, *157*, 109836. https://doi.org/10.1016/j.paid.2020.109836

Keitel, A., Wojtecki, L., Hirschmann, J., Hartmann, C. J., Ferrea, S., Südmeyer, M., & Schnitzler, A. (2013). Motor and cognitive placebo-/nocebo-responses in Parkinson's disease patients with deep brain stimulation. *Behavioural Brain Research*, *250*, 199-205. https://doi.org/https://doi.org/10.1016/j.bbr.2013.04.051

Kemp, S (2024). Digital 2024: Global overview report. *Datareportal.* https://datareportal.com/reports/digital-2024-global-overview-report

Keyes, C. L. M., Dhingra, S. S., & Simoes, E. J. (2010). Change in level of positive mental health as a predictor of future risk of mental illness. *American Journal of Public Health*, *100*, 2366-2371. https://doi.org/10.2105/ajph.2010.192245

Kimble, M., Flack, W., Koide, J., Bennion, K., Brenneman, M., & Meyersburg, C. (2021). Student reactions to traumatic material in literature: Implications for trigger warnings. *PLOS ONE, 16*, e0247579. https://doi.org/10.1371/journal.pone.0247579

Kimble, M., Koide, J., & Flack, W. F. (2022) Students responses to differing trigger warnings: A replication and extension. *Journal of American College Health*, 1-4. https://doi.org/10.1080/07448481.2022.2098038

Kirsch, I. (1997). Response expectancy theory and application: A decennial review. *Applied and Preventive Psychology*, *6*, 69-79. https://doi.org/https://doi.org/10.1016/S09621849(05)80012-5

Koval, P., Kuppens, P., Allen, N. B., & Sheeber, L. (2012). Getting stuck in depression: the roles of rumination and emotional inertia. *Cognition & Emotion*, *26*, 1412-1427. https://doi.org/10.1080/02699931.2012.667392

Kraiss, J. T., Ten Klooster, P. M., Moskowitz, J. T., & Bohlmeijer, E. T. (2020). The relationship between emotion regulation and well-being in patients with mental disorders: A meta-analysis. *Comprehensive Psychiatry*, *102*, 152189. https://doi.org/10.1016/j.comppsych.2020.152189

Kring, A. M., & Sloan, D. M. (Eds.). (2010). *Emotion regulation and psychopathology: A transdiagnostic approach to etiology and treatment.* The Guilford Press.

Krypotos, A.-M., Effting, M., Kindt, M., & Beckers, T. (2015). Avoidance learning: a review of theoretical models and recent developments. *Frontiers in Behavioural Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00189

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). emotional inertia and psychological maladjustment. *Psychological Science*, *21*, 984-991. https://doi.org/10.1177/0956797610372634

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1-26. https://doi.org/10.18637/jss.v082.i13

Lane, R. D., & Schwartz, G. E. (1987). Levels of emotional awareness: a cognitive-developmental theory and its application to psychopathology. *The American Journal of Psychiatry*, *144*, 133-143. https://doi.org/10.1176/ajp.144.2.133

Larsen, R. J. (2000). Towards a science of mood regulation. *Psychological Inquiry*, *11*, 129-141. https://doi.org/10.2307/1449791

Lavis, A., & Winter, R. (2020). #Online harms or benefits? An ethnographic analysis of the positives and negatives of peer-support around self-harm on social media. *Journal of Child Psychology and Psychiatry, 61*, 842-854. https://doi.org/10.1111/jcpp.13245

Lench, H. C., Levine, L. J., Dang, V., Kaiser, K. A., Carpenter, Z. K., Carlson, S. J., Flynn, E., Perez, K. A., & Winckler, B. (2021). Optimistic expectations have benefits for effort and emotion with little cost. *Emotion*, *21*, 1213-1223. https://doi.org/10.1037/emo0000957

Leslie, K. M. (2008). Harm reduction: An approach to reducing risky health behaviours in adolescents. *Paediatrics & Child Health*, 13, 53-60. https://doi.org/10.1093/pch/13.1.53

Lin, H., Gao, H., Ye, Z. e., Wang, P., Tao, L., Ke, X., Zhou, H., & Jin, H. (2012) Expectation enhances event-related responses to affective stimuli. *Neuroscience Letters*, *522*, 123-127. https://doi.org/https://doi.org/10.1016/j.neulet.2012.06.022

Litman, J. (2005). Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition and Emotion*, 19, 793-814. https://doi.org/10.1080/02699930541000101

Llamas, M. (2023). Facebook, Instagram, and other social media lawsuits. *Consumer Notice*. https://www.consumernotice.org/legal/social-media-harm-lawsuit/#:~:text=A%20federal%20lawsuit%20filed%20by,the%20platform%20at%20age%2011.

Lockhart, E. A. (2016). Why trigger warnings are beneficial, perhaps even necessary. *First Amendment Studies,* 50, 59-69. https://doi.org/10.1080/21689725.2016.1232623

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin, 116*, 75-98. https://doi.org/10.1037/0033-2909.116.1.75

Longo, Y., Coyne, I., & Joseph, S. (2018). Development of the short version of the Scales of General Well-Being: The 14-item SGWB. *Personality & Individual Differences, 124*, 31-34. https://doi.org/10.1016/j.paid.2017.11.042

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the depression anxiety stress scales*. Psychology Foundation.

Lukianoff, G. H., Jonathan. (2015). The coddling of the American mind. *The Atlantic.* https://www.theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356/

Manne, K. (2015). Opinion | Why I Use Trigger Warnings. *New York Times*. https://www.nytimes.com/2015/09/20/opinion/sunday/why-i-use-trigger-warnings.html

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-

quality, realistic picture database. *Behaviour Research Methods, 46*, 596-610. https://doi.org/10.3758/s13428-013-0379-1

Marks, D. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology*, *64*, 17-24. https://doi.org/10.1111/j.2044-8295.1973.tb01322.x

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, *31*, 301-306. https://doi.org/10.1111/j.20448260.1992.tb00997.x

Massey, A., & Hill, A. J. (2012). Dieting and food craving. A descriptive, quasi-prospective study. *Appetite*, *58*, 781-785. https://doi.org/10.1016/j.appet.2012.01.020

McGhie, S. F., Bellet, B. W., Mellen, E. J., & McNally, R. J. (2022). Self-triggering: Does function determine pathogenesis? *Psychological trauma: theory, research, practice, and policy*, 15, 951–960. https://doi.org/10.1037/tra0001195.

McKenzie, K. C., & Gross, J. J. (2014). Non-suicidal self-injury: An emotion regulation perspective. *Psychopathology*, *47*, 207-219. https://doi.org/http://dx.doi.org/10.1159/000358097

Mees, U., & Schmitt, A. (2008). Goals of action and emotional reasons for action. A modern version of the theory of ultimate psychological hedonism. *Journal for the Theory of Social Behaviour*, *38*, 157-178. https://doi.org/10.1111/j.1468-5914.2008.00364.x

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*, 423-429. https://doi.org/10.1111/j.1467-9280.1997.tb00455.x

Meta. (2023). Transparency Center. *Meta*. https://transparency.fb.com/en-gb/

Millgram, Y., Huppert, J. D., & Tamir, M. (2020). Emotion goals in psychopathology: A new perspective on dysfunctional emotion regulation. *Current Directions in Psychological Science*, *29*, 242-247. https://doi.org/10.1177/0963721420917713

Millgram, Y., Joormann, J., Huppert, J. D., & Tamir, M. (2015). Sad as a matter of choice? Emotion-regulation goals in depression. *Psychological Science*, *26*, 1216-1228. https://doi.org/10.1177/0956797615583295

Moeck, E. K. (2023). Why it might not help – and could hurt – to brace for the worst. *Psyche.* https://psyche.co/ideas/why-it-might-not-help-and-could-hurt-to-brace-for-the-worst

Moeck, E. K., Bridgland, V. M. E., & Takarangi, M. K. T. (2022). Food for thought: Commentary on Burnette et al. (2021) "Concerns and recommendations for using Amazon MTurk for eating disorder research". *International Journal of Eating Disorders*, *55*, 282-284. https://doi.org/10.1002/eat.23671

Molly Russell Foundation. (2023). Preventable yet pervasive: The prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest. *Molly Russell Foundation*. https://mollyrosefoundation.org/resources/online-safety/

Moreno, M. A., Ton, A., Selkie, E., & Evans, Y. (2016). Secret Society 123: Understanding the language of self-harm on Instagram. *Journal of Adolescent Health*, *58*, 78-84. https://doi.org/10.1016/j.jadohealth.2015.09.015

Morrison, S. (2022). TikTok won't stop serving me horror and death. *Vox.* https://www.vox.com/recode/2022/10/26/23423257/tiktok-for-you-page-algorithm

Mosseri, A. (2019a). Changes we're making to do more to support and protect the most vulnerable people who use instagram. *Instagram.* https://about.instagram.com/blog/announcements/supporting-and-protecting-vulnerable-people-on-instagram

Mosseri, A. (2019b). Taking more steps to keep the people who use Instagram safe. *Instagram.* https://about.instagram.com/blog/announcements/more-steps-to-keep-instagram-users-safe

Murayama, K., Usami, S., & Sakaki, M. (2022). Summary-statistics-based power analysis: A new and practical method to determine sample size for mixed-effects modeling. *Psychological Methods*. 27, 1014-1038. https://doi.org/10.1037/met0000330

Nahleen, S., Nixon, R. D. V., & Takarangi, M. K. T. (2021). The role of belief in memory amplification for trauma events. *Journal of Behaviour Therapy & Experimental Psychiatry, 72*, 101652. https://doi.org/10.1016/j.jbtep.2021.101652

Naughton, J. (2022). Molly Russell was trapped by the cruel algorithms of Pinterest and Instagram. *The Guardian.* https://www.theguardian.com/commentisfree/2022/oct/01/molly-russell-was-trapped-by-the-cruel-algorithms-of-pinterest-and-instagram

Neubauer, A. B., Smyth, J. M., & Sliwinski, M. J. (2018). When you see it coming: Stressor anticipation modulates stress effects on negative affect. *Emotion*, *18*, 342-354. https://doi.org/10.1037/emo0000381

Newberry, C. (2023). Instagram Hashtag Guide 2023. *Hootsuite.* https://blog.hootsuite.com/instagram-hashtags/#:~:text=They%20are%20used%20to%20categorize,posts%20tagged%20with%20that%20hashtag

Oosterwijk, S. (2017). Choosing the negative: A behavioural demonstration of morbid curiosity. *PLOS ONE*, *12*, e0178399. https://doi.org/10.1371/journal.pone.0178399

Park, E., Kim, W.-H., & Kim, S.-B. (2022). What topics do members of the eating disorder online community discuss and empathize with? An application of big data analytics. *Healthcare*, *10*, 928. https://doi.org/10.3390/healthcare10050928

Pearson, J., Rademaker, R. L., & Tong, F. (2011). Evaluating the mind's eye: the metacognition of visual imagery. *Psychological Science*, *22*, 1535-1542. https://doi.org/10.1177/0956797611417134

Preece, D. A., Becerra, R., Robinson, K., Dandy, J., & Allan, A. (2018). Measuring emotion regulation ability across negative and positive emotions: The Perth Emotion Regulation Competency Inventory (PERCI). *Personality and Individual Differences*, *135*, 229-241. https://doi.org/10.1016/j.paid.2018.07.025

Princing, M. (2021). Why do we love true crime? *Right as Rain.* https://rightasrain.uwmedicine.org/life/leisure/true-crime

Ray, R. D., McRae, K., Ochsner, K. N., & Gross, J. J. (2010). Cognitive reappraisal of negative affect: Converging evidence from EMG and self-report. *Emotion, 10*, 587-592. https://doi.org/10.1037/a0019015

Redmond, S., Jones, N. M., Holman, E. A., & Silver, R. C. (2019). Who watches an ISIS beheading-And why. *American Psychologist*, *74*, 555-568. https://doi.org/10.1037/amp0000438

Reinecke, L. (2016). Mood management theory. *The International Encyclopedia of Media Effects*, 1-13. https://doi.org/10.1002/9781118783764.wbieme0085

Riachi, E., Holma, J., & Laitila, A. (2022). Psychotherapists' views on triggering factors for psychological disorders. *Discover Psychology*, *2*, 44. https://doi.org/10.1007/s44202022-00058-y

Ringold, D. J. (2002). Boomerang effects in response to public health interventions: Some unintended consequences in the alcoholic beverage market. *Journal of Consumer Policy*, *25*, 27–63. https://doi.org/10.1023/a:1014588126336

Rodrigues, E. V. (2022). Doomscrolling – threat to Mental Health and Well-being: A Review. *International Journal of Nursing Research*, *8*, 127-130. https://doi.org/10.31690/ijnr.2022.v08i04.002

Rodway, C., Tham, S. G., Richards, N., Ibrahim, S., Turnbull, P., Kapur, N., & Appleby, L. (2023). Online harms? Suicide-related online experience: a UK-wide case series study

of young people who die by suicide. *Psychological Medicine*, *53*, 4434-4445. https://doi.org/10.1017/s0033291722001258

Rooney, T., Sharpe, L., Todd, J., Richmond, B., & Colagiuri, B. (2022). The relationship between expectancy, anxiety, and the nocebo effect: a systematic review and meta-analysis with recommendations for future research. *Health Psychology Review*, 7, 550-577. https://doi.org/10.1080/17437199.2022.2125894

Rosenberg, B., & Siegel, J. (2018). A 50-year review of psychological reactance theory: Do not read this article. *Motivation Science*, *4*, 281-300. https://doi.org/10.1037/mot0000091

Ruan, B., Hsee, C. K., & Lu, Z. Y. (2018). The Teasing Effect: An underappreciated benefit of creating and resolving an uncertainty. *Journal of Marketing Research*, *55*, 556-570. https://doi.org/10.1509/jmr.15.0346

Sanson, M., Strange, D., & Garry, M. (2019). Trigger warnings are trivially helpful at reducing negative affect, intrusive thoughts, and avoidance. *Clinical Psychological Science, 7*, 778-793. https://doi.org/10.1177/2167702619827018

Scheibe, S., & Blanchard-Fields, F. (2009). Effects of regulating emotions on cognitive performance: what is costly for young adults is not so costly for older adults. *Psychology and Aging*, *24*, 217–223. https://doi.org/10.1037/a0013807

Schönbrodt, F. D., & Perugini, M. (2018). Corrigendum to "At what sample size do correlations stabilize?" *Journal of Research in Personality, 74*, 194. https://doi.org/10.1016/j.jrp.2018.02.010

Schwarz, N. (2017). Retrospective and Concurrent Self-Reports: The Rationale for Real-Time Data Capture. In A. Stone, S. Shiffman, A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 11-26). Oxford University Press. https://doi.org/10.1201/9781584888901.ch2

Shafir, R., & Sheppes, G. (2018). When knowledge is (Not) power- the influence of anticipatory information on subsequent emotion regulation: Neural and behavioural evidence. *Journal of Experimental Psychology: General*, *147*, 1225-1240. https://doi.org/10.1037/xge0000452

Shafir, R., & Sheppes, G. (2020). How anticipatory information shapes subsequent emotion regulation. *Emotion*, *20*, 68-74. https://doi.org/10.1037/emo0000673

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using mechanical turk to study clinical populations. *Clinical Psychological Science*, *1*, 213–220. https://doi.org/10.1177/2167702612469015

Sharma, B., Lee, S. S., & Johnson, B. K. (2022). The dark at the end of the tunnel: Doomscrolling on social media newsfeeds. *Technology, Mind, and Behaviour*, *3*. https://doi.org/10.1037/tmb0000059

Shepperd, J. A., Ouellette, J. A., & Fernandez, J. K. (1996). Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback. *Journal of Personality and Social Psychology*, 70, 844-855. https://doi.org/10.1037/0022-3514.70.4.844

Sheppes, G., & Gross, J. J. (2011). Is timing everything? Temporal considerations in emotion regulation. *Personality and Social Psychology Review*, 15, 319-331. https://doi.org/10.1177/1088868310395778

Sheppes, G., Scheibe, S., Suri, G., & Gross, J. J. (2011). Emotion-regulation choice, *Psychological Science*, *22*, 1391-1392. https://doi.org/10.1177/0956797611418350

Simister, E. T., Bridgland, V. M. E., & Takarangi, M. K. T. (2023). To look or not to look: Instagram's sensitive screens do not deter vulnerable people from viewing negative content. *Behaviour Therapy*. https://doi.org/10.1016/j.beth.2023.06.001

Simister, E. T., Bridgland, V. M. E., Williamson, P & Takarangi, M. K. T. (2023) Mind the information-gap: Instagram's sensitive-content screens are more likely to deter people from viewing potentially distressing content when they provide information about the content, *Media Psychology*, *26*, 660-679, https://doi.org/10.1080/15213269.2023.2211774

Simons, J.S., Gaher, R.M. (2005). The Distress Tolerance Scale: Development and validation of a self-report measure. *Motivation and Emotion,* 29, 83–102. https://doi.org/10.1007/s11031-005-7955-3

Skinner, E. A., Edge, K., Altman, J., & Sherwood, H. (2003). Searching for the structure of coping: a review and critique of category systems for classifying ways of coping. *Psychological Bulletin*, *129*, 216-269. https://doi.org/10.1037/00332909.129.2.216

Spacapan, S., & Cohen, S. (1983). Effects and aftereffects of stressor expectations. *Journal of Personality and Social Psychology*, *45*, 1243-1254. https://doi.org/10.1037/00223514.45.6.1243

Spence, R., Bifulco, A., Bradbury, P., Martellozzo, E., & Demarco, J. (2023). The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *17*. https://doi.org/10.5817/cp2023-4-8

Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. *Developmental Review*, 28, 78-106. https://doi.org/10.1016/j.dr.2007.08.002

Stirling, N. S., Bridgland, V. M., & Takarangi, M. K. (2022). Nocebo effects on informed consent within medical and psychological settings: A scoping review. *Ethics & Behaviour*, 1-26. https://doi/org/10/1080/10508422.2022.2081853

Sung, M. (2020). It's almost impossible to avoid triggering content on TikTok. *Mashable.* https://mashable.com/article/tiktok-algorithm-triggers

Susi, K., Glover-Ford, F., Stewart, A., Knowles Bevis, R., & Hawton, K. (2023). Research Review: Viewing self-harm images on the internet and social media platforms: systematic review of the impact and associated psychological mechanisms. *Journal of Child Psychology and Psychiatry*, *64*, 1115-1139. https://doi.org/https://doi.org/10.1111/jcpp.13754

Sweeny, K., & Cavanaugh, A. G. (2012). Waiting is the hardest part: a model of uncertainty navigation in the context of health news. *Health Psychology Review*, *6*, 147-164. https://doi.org/10.1080/17437199.2010.520112

Sweeny, K., & Falkenstein, A. (2015). Is waiting the hardest part? Comparing the emotional experiences of awaiting and receiving bad news. *Personality and Social Psychology Bulletin, 41*, 1551-1559. https://doi.org/10.1177/0146167215601407

Sweeny, K., & Shepperd, J. A. (2010). The costs of optimism and the benefits of pessimism. *Emotion*, *10*, 750-753. https://doi.org/10.1037/a0019016

Sweeny, K., Reynolds, C. A., Falkenstein, A., Andrews, S. E., & Dooley, M. D. (2016). Two definitions of waiting well. *Emotion*, 16, 129-143. https://doi.org/10.1037/emo0000117

Systrom, K. [@instagram]. (2017, March 24). *As part of our goal to build a safe environment, we also have some updates to announce*. Instagram. https://www.instagram.com/p/BR-8eo5BGZw/?hl=eN

Takarangi, M. K. T., Bridgland, V. M. E., & Simister, E. T. (2023). A nervous wait: Instagram's sensitive-content screens cause anticipatory anxiety but do not mitigate reactions to negative content. *Cognition and Emotion*, *37*, 1-15. https://doi.org/10.1080/02699931.2023.2258574

Tamir, M. (2009). What do people want to feel and why? *Current Directions in Psychological Science*, *18*, 101-105. https://doi.org/10.1111/j.14678721.2009.01617.x

Tamir, M. (2016). Why do people regulate their emotions? A taxonomy of motives in emotion regulation. *Personality and Social Psychology Review*, *20*, 199-222. https://doi.org/10.1177/1088868315586325

Thiruchselvam, R., Blechert, J., Sheppes, G., Rydstrom, A., & Gross, J. J. (2011). The temporal dynamics of emotion regulation: An EEG study of distraction and reappraisal. *Biological Psychology*, *87*, 84-92. https://doi.org/10.1016/j.biopsycho.2011.02.009

Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of vectors of the mind.* University of Chicago Press.

University of Reading. (2021). Guidance on content warnings on course content ('trigger' warnings). https://www.reading.ac.uk/cqsd//media/project/functions/cqsd/documents/qap/trigger-warnings.pdf

Van Beveren, M.-L., Goossens, L., Volkaert, B., Grassmann, C., Wante, L., Vandeweghe, L., Verbeken, S., & Braet, C. (2019). How do I feel right now? Emotional awareness, emotion regulation, and depressive symptoms in youth. *European Child & Adolescent Psychiatry*, *28*, 389-398. https://doi.org/10.1007/s00787-018-1203-3

Van Dijk, E., & Zeelenberg, M. (2007). When curiosity killed regret: Avoiding or seeking the unknown in decision-making under uncertainty. *Journal of Experimental Social Psychology*, 43, 656-662. https://doi.org/10.1016/j.jesp.2006.06.004

Van Dijk, W. W., Zeelenberg, M., & Van Der Pligt, J. (2003). Blessed are those who expect nothing: Lowering expectations as a way of avoiding disappointment. *Journal of Economic Psychology*, 24, 505-516. https://doi.org/10.1016/s0167-4870(02)00211-8

van Stolk-Cooke, K., Brown, A., Maheux, A., Parent, J., Forehand, R., & Price, M. (2018). Crowdsourcing trauma: Psychopathology in a trauma exposed sample recruited via Mechanical Turk. *Journal of Traumatic Stress*, *31*, 549 –557. http://dx.doi.org/10.1002/jts.22303

Varelmann, D., Pancaro, C., Cappiello, E. C., & Camann, W. R. (2011). Nocebo-induced hyperalgesia during local anesthetic injection. *Obstetric Anesthesia Digest*, *31*. https://journals.lww.com/obstetricanesthesia/Fulltext/2011/06000/Nocebo_induced_Hyperalgesia_During_Local.68.aspx

Von Stumm, S., Hell, B., & Chamorro-Premuzic, T. (2011). The hungry mind. *Perspectives on Psychological Science*, 6, 574-588. https://doi.org/10.1177/1745691611421204

Wang, T., Brede, M., Ianni, A., & Mentzakis, E. (2018). Social interactions in online eating disorder communities: A network perspective. *PLOS ONE*, *13*, e0200800. https://doi.org/10.1371/journal.pone.0200800

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063-1070. https://doi.org/10.1037/0022-3514.54.6.1063

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). *The PTSD Checklist for DSM-5 (PCL-5)*. Available at www.ptsd.va.gov.

Weaver, A. J. (2011). A meta-analytical review of selective exposure to and the enjoyment of media violence. *Journal of Broadcasting and Electronic Media*, 55, 232-250. https://doi.org/10.1080/08838151.2011.570826

Webb, T. L., Miles, E., & Sheeran, P. (2012). Dealing with feeling: A meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation. *Psychological Bulletin*, *138*, 775-808. https://doi.org/10.1037/a0027600

Werner, K., & Gross, J. J. (2010). Emotion regulation and psychopathology: A conceptual framework. In A. M. Kring & D. M. Sloan (Eds.), *Emotion regulation and psychopathology: A transdiagnostic approach to etiology and treatment*. (pp. 13-37). The Guilford Press.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291-298. https://doi.org/10.1177/1745691611406923

Wilson, T. D., Lisle, D. J., Kraft, D., & Wetzel, C. G. (1989). Preferences as expectation-driven inferences: Effects of affective expectations on affective experience. *Journal of Personality and Social Psychology*, *56*, 519- 530. https://doi.org/10.1037//00223514.56.4.519

Within Health. (2023). Bringing light to the dark side of TikTok's algorithm: An exploration of eating disorder content on TikTok and suggestions for improvements and interventions. *Within Health.* https://withinhealth.com/explore/tiktok-research

Wolgast, M., Lundh, L. G., & Viborg, G. (2011). Cognitive reappraisal and acceptance: An experimental comparison of two emotion regulation strategies. *Behaviour Research and Therapy*, *49*, 858-866. https://doi.org/10.1016/j.brat.2011.09.011

Yagi, A., Fitzgibbon, L., Murayama, K., Shinomori, K., & Sakaki, M. (2023). Uncertainty drives exploration of negative information across younger and older adults. *Cognitive, Affective & Behavioural Neuroscience*, *23*, 809–826. https://doi.org/10.3758/s13415023-01082-8

Appendices

Appendix A: Image Stimuli (Study 1a and 1b)

Note: I used images with an asterisk (*) in Study 1b.

Negative Images

NAPS ID	Category	Description
*People_238_h	People	Dead Bodies
*People_198_h	People	Mutilated Leg
*People_237_h	People	Face Skin
*Faces_367_h	Faces	Mutilated Face
*Animals_074_h	Animals	Starved Dog
*Faces_371_v	Faces	Mutilated Face
*People_038_h	People	Assault
*Faces_364_v	Faces	Mutilated Face
*People_226_h	People	Accident
*People_240_h	People	Skin Disease
*People_218_v	People	Mutilated Hand
*Faces_143_v	Faces	Mutilated Face
*People_208_h	People	Dead Body
*People_227_h	People	Burns
*Animals_056_h	Animals	Dead Cat
*People_201_v	People	Accident
*People_221_h	People	Surgery
*People_211_v	People	Mutilated Hand
*Faces_159_h	Faces	Mutilated Face
*People_246_h	People	Black Eye
*People_220_h	People	Disease
*Faces_365_v	Faces	Mutilated Face
*Animals_077_h	Animals	Sick Dog
*Objects_125_h	Objects	Toilet
*Objects_139_h	Objects	Doll Head
*People_205_v	People	Drowned Man
*Faces_366_h	Faces	Mutilated Face

NAPS ID	Category	Description
*People_022_h	People	Car Crash
*People_127_h	People	Assault
*People_222_h	People	Hooked Skin
People_128_h	People	Wounded Child
Faces_010_h	Faces	Child With Burns
Animals_078_h	Animals	Dead Mouse
People_200_h	People	Dead Bodies
People_031_v	People	Burns
Objects_149_h	Objects	Blood
Faces_145_v	Faces	Mutilated Face
Faces_149_v	Faces	Mutilated Face
Animals_071_h	Animals	Dead Moose
Objects_003_h	Objects	Crashed Car
People_204_v	People	Skin Inflammation
People_239_h	People	Skin Disease
Faces_018_h	Faces	Armed Boys
Animals_033_h	Animals	Dead Bird
People_140_h	People	Wounded People
Landscapes_139_h	Landscapes	Waste
People_020_h	People	Car Crash
Faces_284_h	Faces	Crippled Man
Animals_024_h	Animals	Dead Cat
Faces_293_h	Faces	Police Arresting Someone
People_016_h	People	Car Crash
Faces_172_h	Faces	Elderly Man
Animals_039_h	Animals	Dead Dog
Animals_063_h	Animals	Dead Deer
People_233_h	People	Dead Animal
Faces_009_h	Faces	Hurting Child
People_118_h	People	Homeless Man
Faces_283_h	Faces	Elderly Woman Crying
Landscapes_026_h	Landscapes	Waste

NAPS ID	Category	Description
Animals_062_h	Animals	Dead Cow
People_225_h	People	Dead Body
People_241_h	People	Eye Disease
Faces_362_v	Faces	Mutilated Face
People_013_v	People	Car Crash
Animals_027_h	Animals	Dead Bird
Animals_068_h	Animals	Dead Chinchillas
Animals_001_h	Animals	Dead Stork
People_143_h	People	Homeless Woman
People_202_h	People	Surgery
People_215_h	People	Disease

Neutral Images

NAPS ID	Category	Description
Landscapes_056_h	Landscapes	Devastated House
Objects_280_v	Objects	Icicle
Objects_108_v	Objects	Objects
Objects_251_v	Objects	Buttons
Landscapes_170_h	Landscapes	Plants
People_164_h	People	Foot
Faces_167_v	Faces	Man
Objects_210_h	Objects	Window Grating
Objects_179_h	Objects	Objects
Animals_133_h	Animals	Cat
Objects_059_h	Objects	Snails
Animals_081_h	Animals	Insects
Objects_224_h	Objects	Hook
Objects_246_h	Objects	Mouse
Objects_147_v	Objects	Knife
Faces_039_h	Faces	Sad Girl
Faces_216_h	Faces	Man

NAPS ID	Category	Description
Faces_305_h	Faces	Woman
Objects_314_h	Objects	Car
Animals_141_h	Animals	Monkey
*People_066_v	People	Fish Stall
*Objects_057_h	Objects	Meal
*Animals_047_h	Animals	Lion
*People_146_h	People	Garbage Collectors
*Objects_213_h	Objects	Fence
*Faces_218_h	Faces	Woman
*Objects_112_h	Objects	Bouquet
*Objects_130_h	Objects	Lock
*People_078_v	People	Bottle
*Objects_226_h	Objects	Lighter
*Landscapes_061_h	Landscapes	Mine
*Objects_071_h	Objects	Fruits
*Landscapes_067_h	Landscapes	Balcony
*Animals_011_h	Animals	Black Panther
*Animals_072_h	Animals	Snake
*Objects_197_v	Objects	Knife
*Objects_239_v	Objects	Lock
*Animals_058_h	Animals	Snake
*Objects_067_h	Objects	Sausages
*Objects_308_h	Objects	Car
*Faces_312_h	Faces	Man
*Objects_119_h	Objects	Knife
*Objects_311_h	Objects	Bus
*Faces_320_v	Faces	Elderly Woman
*Landscapes_076_h	Landscapes	Wall
*Objects_050_h	Objects	Finished Meal
*Animals_127_h	Animals	Snails
*Objects_244_h	Objects	Car Pedals
*Objects_189_h	Objects	Shoes

NAPS ID	Category	Description
*Objects_307_v	Objects	Mower
Animals_035_h	Animals	Snake
Objects_299_h	Objects	Bicycle
Faces_286_h	Faces	Judo Fighters
Landscapes_079_v	Landscapes	House
Objects_089_h	Objects	Lychee Fruit
Objects_196_h	Objects	Watch Straps
Objects_211_h	Objects	Wheel
Animals_200_v	Animals	Cat
Objects_274_h	Objects	Car
Objects_014_h	Objects	Sausages
Animals_006_v	Animals	Snake
Landscapes_044_h	Landscapes	House
Objects_208_h	Objects	Wood
Objects_204_h	Objects	Containers
Landscapes_016_h	Landscapes	Block Of Flats
Objects_046_h	Objects	Chicken
Animals_014_h	Animals	Snake
Objects_296_h	Objects	Car
People_100_h	People	Firemen
Objects_065_h	Objects	Lobsters

Positive Images

NAPS ID	Category	Description
Landscapes_137_h	Landscapes	Flower
Animals_131_h	Animals	Ducks
Faces_140_h	Faces	Grandparents And Children
People_185_h	People	Swimming Pool
People_043_h	People	Children
Faces_050_h	Faces	Children Playing
Faces_356_h	Faces	Couple Smiling

NAPS ID	Category	Description
Animals_183_h	Animals	Dog
Landscapes_134_h	Landscapes	Flowers
Faces_120_h	Faces	Boy Smiling
Faces_346_v	Faces	Couple Smiling
Landscapes_113_h	Landscapes	Sea
Objects_327_h	Objects	Boats
Faces_079_h	Faces	Mother And Child
People_116_h	People	Beach
Landscapes_142_h	Landscapes	Water
Landscapes_138_h	Landscapes	Sky
Landscapes_174_h	Landscapes	Mountains
People_055_h	People	Child
Landscapes_178_h	Landscapes	Beach
People_026_h	People	Zoo
Animals_163_h	Animals	Llama
Landscapes_122_v	Landscapes	Bridge
Objects_077_h	Objects	Vegetables
Animals_172_h	Animals	Peacock
Landscapes_157_h	Landscapes	Mountains
Objects_326_h	Objects	Sailboat
Landscapes_121_h	Landscapes	Sea
Faces_001_h	Faces	Children With a Dog
People_103_h	People	Beach
Landscapes_117_h	Landscapes	Fields
People_096_h	People	Ski Slope
Landscapes_175_h	Landscapes	River
Faces_114_h	Faces	Baby
Objects_084_v	Objects	Table
People_052_h	People	Cat And Child
Landscapes_103_v	Landscapes	Tree
People_187_h	People	Woman Jumping
Faces_122_h	Faces	Girl Playing

NAPS ID	Category	Description
Faces_089_h	Faces	Mother And Child
*Animals_186_h	Animals	Dolphins
*Animals_156_h	Animals	Fish
*Landscapes_185_v	Landscapes	Flowers
*Animals_187_h	Animals	Dolphins
*Landscapes_132_h	Landscapes	Meadow
*Landscapes_098_h	Landscapes	Sunset
*Faces_002_v	Faces	Woman With a Dog
*Landscapes_096_h	Landscapes	Flowers
*Landscapes_140_v	Landscapes	Palm Trees
*People_115_h	People	Desert
*Animals_201_h	Animals	Turtle
*People_051_h	People	Child
*Landscapes_168_h	Landscapes	Forest
*Animals_184_h	Animals	Cows
*Animals_220_h	Animals	Fish
*People_172_v	People	Man Swinging
*Landscapes_141_h	Landscapes	Sky
*People_113_h	People	Seaside
*Landscapes_116_v	Landscapes	River
*Animals_177_h	Animals	Dog
*People_110_h	People	Hill
*Landscapes_123_h	Landscapes	Sea
*Faces_109_v	Faces	Child Smiling
*People_190_h	People	Diver
*Landscapes_165_h	Landscapes	Mountains
*Animals_166_v	Animals	Cats
*Landscapes_154_h	Landscapes	Mountains
*Landscapes_183_h	Landscapes	Sea
*Landscapes_120_v	Landscapes	Meadow
*Landscapes_180_h	Landscapes	Sea

Appendix B: Image Stimuli and Corresponding Brief and Detailed Content Descriptions (Studies 2, 3a and 3b)

Set number	Brief content description	Detailed content description
Set 1		
People_238_h	Deceased people.	The aftermath of a mass shooting where numerous people lie deceased and bloodied.
Faces_367_h	A deceased person.	An elderly man lying deceased on the floor in a pool of blood.
People_038_h	Gun violence.	A man points a rifle at a child, who is lying on the ground.
People_240_h	A skin infection.	A large skin infection covers the entire torso of a child.
People_208_h	A hand injury.	A bloodied and mutilated hand sticking out from underneath a cover.
People_201_v	A deceased person.	The body of a man who has been hit and killed by a train.
Faces_159_h	Facial burns.	A person who is receiving treatment for severe burns on their face.
Faces_365_v	Facial burns.	A man who has lost his eyes and nose due to severe facial burns.
Objects_139_h	A broken toy.	A severed and dirty doll head lying upon a pile of twigs.
People_022_h	A deceased person.	An overturned, severely damaged truck and a deceased person covered by a tarp.
Set 2		
People_198_h	An infected foot.	A person receives treatment for a dry, swollen, and infected lower leg.
Animals_074_h	Animal abuse.	A sick, old, and blind dog lying on a bed in a dirty room.
Faces_364_v	A facial injury.	A child who has sustained an injury, resulting in the loss of their facial skin.
People_218_v	Burns.	A person receives treatment for a severe burn on their hand.

Set number	Brief content description	Detailed content description
People_227_h	Burns.	A child receives treatment for burns to their stomach and thighs.
People_221_h	A facial injury.	A person receives treatment for a large laceration below their right eye.
People_246_h	A facial injury.	An elderly person who has sustained an injury, resulting in a severely swollen black eye.
Animals_077_h	Animal abuse.	An unkept, injured dog who has bloodied wounds on its face.
People_205_v	A deceased person.	A deceased man being pulled out of the water by two men.
People_127_h	Physical violence.	A group of men violently beat up another person who is lying on the ground
Set 3		
People_237_h	Removed skin.	A piece of skin removed from a person's head, with attached ear, nose, and lip.
Faces_371_v	A facial injury.	A young child who has lost their right eye and is now disfigured.
People_226_h	A deceased person.	The body of an elderly woman who has been hit and killed by a tram.
Faces_143_v	Facial burns.	A woman who has sustained extensive burns, resulting in severe facial scarring and disfigurement.
Animals_056_h	A deceased animal.	A deceased black cat lying in the dirt with a bloodied ear and bulging eyeball.
People_211_v	A hand injury.	A severely injured hand with exposed tissue and bones on four fingers.
People_220_h	A skin infection.	A large skin infection covering the left chest area of a woman.
Objects_125_h	Faeces.	A filthy, unflushed toilet containing used toilet paper and brown faecal matter.
Faces_366_h	A facial injury.	A man who is receiving stitches after sustaining deep facial cuts.
People_222_h	Body injury.	A man who is hanging from hooks in the skin of his back.

Appendix C: Pre-Task Questions (for all studies)

Social Media Sites Check

What social media sites do you use on a regular basis? Select all that apply.

a. Facebook
b. Instagram
c. Twitter
d. Snapchat
e. Konnect
f. WhatsApp
g. Tumblr
h. YouTube
i. TikTok
j. Reddit
k. Pinterest
l. Other (please list any other social media sites you use on a regular basis not listed above):

Instagram Use Questionnaire

1. In the last 7 days, how many days did you use Instagram?
   The rating scale is as follows: 1 = *never,* 2 = *1 day,* 3 = *2 days,* 4 = *3 days,* 5 = *4 days,* 6 = *5 days,* 7 = *6 days,* 8 = *every day.*
2. In the last 30 days, on an average day how many hours did you use Instagram?
   The rating scale is as follows: 1 = *less than half an hour*, 2 = *1 hour*, 3 = *2–3 hours*, 4 = *4–5 hours*, 5 = *more than 6 hours*.
3. Please rate how often you view the following kinds of images on Instagram. The rating scale is as follows: 1 = *Not at all,* 2 = *sometimes*, 3 = *often*, 4 = *very often*.
   a. Travel
   b. Landscapes
   c. Abstract art
   d. Animals
   e. Portraits
   f. Food

Appendix D: Reasons for Uncovering Questionnaire (prior to PCA)

Now we would like you to think back to the Instagram experience task you just completed…

1. Why did you or did you not uncover the screened image(s)? Open text box answer.

Please respond to the following statements regarding the *Instagram experience task* you just completed. The rating scale is as follows: 0 = *not at all true of me*, 1 = *a little bit true of me*, 2 = *moderately true of me*, 3 = *quite a bit true of me*, 4 = *extremely true of me*.

1. I uncovered the screened image(s) because I was excited to see what might lie beneath the screen. (REAS_IE_1)
2. I uncovered the screened image(s) because it was thrilling/exhilarating to do so. (REAS_IE_6)
3. I uncovered the screened image(s) because I enjoy having new and varied experiences. (REAS_IE2_2)
4. I did not uncover the screened image(s) because I don't enjoy taking risks. (REAS_IE_15)
5. I uncovered the screened image(s) because I wanted to reduce uncertainty associated with the covered image. (REAS_IE_2)
6. I uncovered the screened image(s) because I was uncomfortable when I didn't know what the image was. (REAS_IE2_9)
7. I uncovered the screened image(s) because I do not enjoy ambiguity. (REAS_IE2_5)
8. I did not uncover the screened image(s) because I thought I knew what the image was. (REAS_IE2_12)
9. I uncovered the screened image(s) because I was curious. (REAS_IE2_11)
10. I uncovered the screened image(s) because I was eager to learn what the image was. (REAS_IE_3)
11. I uncovered the screened image(s) because I wanted to know why it was covered. (REAS_IE_10)
12. I did not uncover the screened image(s) because I was uninterested. (REAS_IE_8)
13. I uncovered the screened image(s) because my freedom to view the image was restricted. (REAS_IE_4)
14. I uncovered the screened image(s) because I wanted to act with my own free will. (REAS_IE2_15)

15. I uncovered the screened image(s) because I was frustrated that I couldn't see the image. (REAS_IE2_10)
16. I did not uncover the screened image(s) because I trust they were covered for my own good. (REAS_IE2_7)
17. I uncovered the screened image(s) because I was trying to make sense of my past negative experiences. (REAS_IE_5)
18. I uncovered the screened image(s) because I was trying to remind myself of past negative experiences. (REAS_IE_12)
19. I uncovered the screened image(s) because I was trying to prevent the memories of my past negative experiences fading. (REAS_IE_14)
20. I uncovered the screened image(s) because I was trying to forget my past negative experiences. (REAS_IE2_8)
21. I uncovered the screened image(s) because I was sad. (REAS_IE_7)
22. I uncovered the screened image(s) because I wanted to have an experience that matched my negative mood. (REAS_IE_11)
23. I uncovered the screened image(s) because I was feeling down and blue. (REAS_IE_16)
24. I uncovered the screened image(s) because I was unhappy. (REAS_IE2_6)
25. I uncovered the screened image(s) because I was happy. (REAS_IE_13)
26. I uncovered the screened image(s) because I wanted to have an experience that matched my positive mood. (REAS_IE2_13)
27. I uncovered the screened image(s) because I was feeling good. (REAS_IE2_16)
28. I uncovered the screened image(s) because I was content. (REAS_IE2_3)
29. I did not uncover the screened image(s) because I do not like viewing distressing or graphic material. (REAS_IE_9)
30. I did not uncover the screened image(s) because I make an effort to avoid distressing and graphic material. (REAS_IE2_1)
31. I did not uncover the screened image(s) because I thought the material underneath the screen would make me upset. (REAS_IE2_4)
32. I did not uncover the screened image(s) because I thought it would remind me of a past negative experience. (REAS_IE2_14)

Now we would like you to think about when you come across sensitive screens on your own Instagram account. Please respond to the following statements regarding your experiences on your own Instagram account. The rating scale is as follows: 0 = *not at all true of me*, 1 = *a little bit true of me*, 2 = *moderately true of me*, 3 = *quite a bit true of me*, 4 = *extremely true of me*.

I would be more likely to uncover screened images if…

1. I was in a good mood. (REAS_RL_1)
2. the caption was interesting to me. (REAS_RL_2)
3. I was alone. (REAS_RL_3)
4. the comments were interesting to me. (REAS_RL_4)
5. I thought I knew what the image was. (REAS_RL_5)
6. I was in a bad mood. (REAS_RL_6)
7. I thought I would be interested in the subject matter. (REAS_RL_7)
8. I was around others. (REAS_RL_8)
9. I knew the person or account that posted them. (REAS_RL_9)

Appendix E: The Short-form Spielberger State-Trait Anxiety Inventory (STAI-6; Spielberger, 1983)

A number of statements which people have used to describe themselves are given below. Read each statement and then select the most appropriate number to indicate how you feel right now, at this moment (1= *not at all*, 2 = *somewhat*, 3 = *moderately*, 4 = *very much*). There are no right or wrong answers. Do not spend too much time on any one statement but give the answer which seems to describe your present feelings best.

1. I feel calm.
2. I feel tense.
3. I feel upset.
4. I feel relaxed.
5. I feel contented.
6. I feel worried.

Appendix F: The Positive and Negative Affect Schedule (PANAS; Watson et al., 1988)

This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you have felt like this in the past few hours.

The rating scale is as follows: 1 = *Very slightly or not at all*, 2 = *a little*, 3 = *moderately*, 4 = *quite a bit*, 5 = *extremely*.

1. Interested
2. Distressed
3. Excited
4. Upset
5. Strong
6. Guilty
7. Scared
8. Hostile
9. Enthusiastic
10. Proud
11. Irritable
12. Alert
13. Ashamed
14. Inspired
15. Nervous
16. Determined
17. Attentive
18. Jittery
19. Active
20. Afraid

Appendix G: The Depression, Anxiety and Stress Scale (DASS-21; Lovibond & Lovibond, 1995)

Please read each statement and circle a number 0, 1, 2 or 3 which indicates how much the statement applied to you over the past week. There are no right or wrong answers. Do not spend too much time on any statement.

The rating scale is as follows: 0 = *did not apply to me at all*, 1 = *applied to me to some degree, or some of the time*, 2 = *Applied to me to a considerable degree or a good part of time*, 3 = *Applied to me very much or most of the time.*

1. I found it hard to wind down.
2. I was aware of dryness of my mouth.
3. I couldn't seem to experience any positive feeling at all.
4. I experienced breathing difficulty (e.g., excessively rapid breathing, breathlessness in the absence of physical exertion).
5. I found it difficult to work up the initiative to do things.
6. I tended to over-react to situations.
7. I experienced trembling (e.g., in the hands).
8. I felt that I was using a lot of nervous energy.
9. I was worried about situations in which I might panic and make a fool of myself.
10. I felt that I had nothing to look forward to.
11. I found myself getting agitated.
12. I found it difficult to relax.
13. I felt downhearted and blue.
14. I was intolerant of anything that kept me from getting on with what I was doing.
15. I felt I was close to panic.
16. I was unable to become enthusiastic about anything.
17. I felt I wasn't worth much as a person.
18. I felt that I was rather touchy.
19. I was aware of the action of my heart in the absence of physical exertion (e.g., sense of heart rate increase, heart missing a beat).
20. I felt scared without any good reason.
21. I felt that life was meaningless.

Appendix H: The Scale of General Well-Being short form (SGWB-14; Longo et al., 2018)

Below you'll find fourteen statements about your experiences. Please indicate how true each statement is regarding the experiences in your life overall. There are no right or wrong answers. Please choose the answer that best reflects your experience rather than what you think your experience should be. The rating scale is as follows: 1 = *Not at all true*, 2 = *a bit true*, 3 = *somewhat true*, 4 = *mostly true*, 5 = *very true.*

1. I feel happy.
2. I feel energetic.
3. I feel calm.
4. I'm optimistic.
5. In my activities, I feel absorbed by what I'm doing.
6. I'm in touch with how I really feel inside.
7. I accept most aspects of myself.
8. I feel great about myself.
9. I am highly effective at what I do.
10. I feel I am improving.
11. I have a purpose.
12. What I do in my life is worthwhile.
13. What I do is consistent with what I believe I should do.
14. I feel close and connected to the people around me.

Appendix I: Trauma History Screen (THS; Carlson et al., 2011)

The events below may or may not have happened to you. Circle "YES" if that kind of thing has happened to you or circle "NO" if that kind of thing has not happened to you. If you circle "YES" for any events: put a number in the blank next to it to show how many times something like that happened.

A. A really bad car, boat, train, or airplane accident YES/NO _____ times

B. A really bad accident at work or home YES/NO _____ times

C. A hurricane, flood, earthquake, tornado, or fire YES/NO _____ times

D. Hit or kicked hard enough to injure - as a child YES/NO _____ times

E. Hit or kicked hard enough to injure - as an adult YES/NO _____ times

F. Forced or made to have sexual contact - as a child YES/NO _____ times

G. Forced or made to have sexual contact - as an adult YES/NO _____ times

H. Attack with a gun, knife, or weapon YES/NO _____ times

I. During military service - seeing something horrible or being badly scared YES/NO _____ times

J. Sudden death of close family or friend YES/NO _____ times

K. Seeing someone die suddenly or get badly hurt or killed YES/NO _____ times

L. Some other sudden event that made you feel very scared, helpless, or horrified YES/NO _____ times

M. Sudden move or loss of home and possessions YES/NO _____ times

N. Suddenly abandoned by spouse, partner, parent, or family YES/NO _____ times

Briefly describe (in one or two sentences) the most stressful experience of your life in the box below. We are going to ask you a number of questions about this event.

Your age when this happened: _____

When this happened, did anyone get hurt or killed? NO/YES

When this happened, were you afraid that you or someone else might get hurt or killed? NO/YES

When this happened, did you feel very afraid, helpless, or horrified? NO/YES

When this happened, did you feel unreal, spaced out, disoriented, or strange? NO/YES

After this happened, how long were you bothered by it? not at all / 1 week / 2-3 weeks / a month or more

How much did it bother you emotionally? not at all / a little / somewhat / much / very much

Appendix J: The Posttraumatic Stress Disorder Checklist (PCL-5; Weathers et al., 2013)

Below is a list of problems that people sometimes have in response to a very stressful experience. Please read each problem carefully and then circle one of the numbers to the right to indicate how much you have been bothered by that problem in the past month. The rating scale is as follows: 0 = *Not at all*, 1 = *a little bit*, 2 = *moderately*, 3 = *quite a bit*, 4 = *extremely*.

1. Repeated, disturbing, and unwanted memories of the stressful experience?
2. Repeated, disturbing dreams of the stressful experience?
3. Suddenly feeling or acting as if the stressful experience were actually happening again (as if you were actually back there reliving it)?
4. Feeling very upset when something reminded you of the stressful experience?
5. Having strong physical reactions when something reminded you of the stressful experience (for example, heart pounding, trouble breathing, sweating)?
6. Avoiding memories, thoughts, or feelings related to the stressful experience?
7. Avoiding external reminders of the stressful experience (for example, people, places, conversations, activities, objects, or situations)?
8. Trouble remembering important parts of the stressful experience?
9. Having strong negative beliefs about yourself, other people, or the world (for example, having thoughts such as: I am bad, there is something seriously wrong with me, no one can be trusted, the world is completely dangerous)?
10. Blaming yourself or someone else for the stressful experience or what happened after it?
11. Having strong negative feelings such as fear, horror, anger, guilt, or shame?
12. Loss of interest in activities that you used to enjoy?
13. Feeling distant or cut off from other people?
14. Trouble experiencing positive feelings (for example, being unable to feel happiness or have loving feelings for people close to you)?
15. Irritable behaviour, angry outbursts, or acting aggressively?
16. Taking too many risks or doing things that could cause you harm?
17. Being “superalert” or watchful or on guard?
18. Feeling jumpy or easily startled?
19. Having difficulty concentrating?
20. Trouble falling or staying asleep?

Appendix K: The Centrality of Event Scale Short Form (CES; Berntsen & Rubin, 2006)

Please think back upon the most stressful or traumatic event in your life and answer the following questions in an honest and sincere way, by circling a number from 1 = *totally disagree* to 5 = *totally agree*.

1. I feel that this event has become part of my identity.
2. This event has become a reference point for the way I understand myself and the world.
3. I feel that this event has become a central part of my life story.
4. This event has coloured the way I think and feel about other experiences.
5. This event permanently changed my life.
6. I often think about the effects this event will have on my future.
7. This event was a turning point in my life.

Appendix L: The Five-Dimensional Curiosity Scale Revised (5DCR; Kashdan et al., 2020)

Below are statements people often use to describe themselves. Please use the scale below to indicate the degree to which these statements accurately describe you. There are no right or wrong answers. The rating scale is as follows: 1 = *does not describe me at all*, 2 = *barely describes me*, 3 = *somewhat describes me*, 4 = *neutral,* 5 = *generally describes me*, 6 = *mostly describes me*, 7 = *completely describes me*.

Joyous Exploration:

1. I view challenging situations as an opportunity to grow and learn.
2. I seek out situations where it is likely that I will have to think in depth about something.
3. I enjoy learning about subjects that are unfamiliar to me.
4. I find it fascinating to learn new information.

Deprivation Sensitivity:

1. Thinking about solutions to difficult conceptual problems can keep me awake at night.
2. I can spend hours on a single problem because I just can't rest without knowing the answer.
3. I feel frustrated if I can't figure out the solution to a problem, so I work even harder to solve it.
4. I work relentlessly at problems that I feel must be solved.

Stress Tolerance: (entire subscale reverse-scored)

1. The smallest doubt can stop me from seeking out new experiences.
2. I cannot handle the stress that comes from entering uncertain situations.
3. I find it hard to explore new places when I lack confidence in my abilities.
4. It is difficult to concentrate when there is a possibility that I will be taken by surprise.

Thrill-Seeking:

1. Risk-taking is exciting to me.
2. When I have free time, I want to do things that are a little scary.
3. Creating an adventure as I go is much more appealing than a planned adventure.
4. I prefer friends who are excitingly unpredictable.

Social Curiosity:

General Social Curiosity

1. I ask a lot of questions to figure out what interests' other people.
2. When talking to someone who is excited, I am curious to find out why.
3. When talking to someone, I try to discover interesting details about them.
4. I like finding out why people behave the way they do.

Covert Social Curiosity

1. When other people are having a conversation, I like to find out what it's about.
2. When around other people, I like listening to their conversations.
3. When people quarrel, I like to know what's going on.
4. I seek out information about the private lives of people in my life.

Appendix M: Acceptance and Action Questionnaire-II (AAQ-II; Bond et al., 2011)

Below you will find a list of statements. Please rate how true each statement is for you by using the rating scale as follows: 1 = *never true*, 2 = *very seldom true,* 3 = *seldom true*, 4 = *sometimes true,* 5 = *frequently true,* 6 = *almost always true*, 7 = *always true*.

1. My painful experiences and memories make it difficult for me to live a life that I would value.
2. I'm afraid of my feelings.
3. I worry about not being able to control my worries and feelings.
4. My painful memories prevent me from having a fulfilling life.
5. Emotions cause problems in my life.
6. It seems like most people are handling their lives better than I am.
7. Worries get in the way of my success.

Appendix N: Perth Emotion Regulation Competency Inventory (PERCI; Preece et al., 2018)

This questionnaire asks about how you manage and respond to your emotions. Please score the following statements according to **how much you agree or disagree that the statement is true of you**. The rating scale is as follows: 1= *strongly disagree*, 4 = *neither agree nor disagree*, 7 = *strongly agree*.

The first half of the questionnaire asks about *bad* or *unpleasant* emotions, this means emotions like sadness, anger, or fear. The second half asks about *good* or *pleasant* emotions, this means emotions like happiness, amusement, or excitement.

1. When I'm feeling *bad* (feeling an unpleasant emotion), I don't know what to do to feel better.
2. When I'm feeling bad, I do stupid things.
3. When I'm feeling bad, I believe I need to get rid of those feelings at all costs.
4. When I'm feeling bad, I'm powerless to change how I'm feeling.
5. When I'm feeling bad, I can't complete tasks that I'm meant to be doing.
6. When I'm feeling bad, my behaviour becomes out of control.
7. When I'm feeling bad, I can't allow those feelings to be there.
8. When I'm feeling bad, I don't have many strategies (e.g., activities or techniques) to help get rid of that feeling.
9. When I'm feeling bad, I can't get motivated to do important things (work, chores, school etc.).
10. When I'm feeling bad, I can't get motivated to do important things (work, chores, school etc.).
11. When I'm feeling bad, I have trouble controlling my actions.
12. When I'm feeling bad, I must try to totally eliminate those feelings.
13. When I'm feeling bad, I have no control over the strength and duration of that feeling.
14. When I'm feeling bad, I have trouble getting anything done.
15. When I'm feeling bad, I have strong urges to do risky things.
16. When I'm feeling bad, I believe those feelings are unacceptable.
17. When I'm feeling *good* (feeling a pleasant emotion), I do stupid things.
18. When I'm feeling good, I don't have many strategies (e.g., activities or techniques) to increase the strength of that feeling.

19. When I’m feeling good, I have trouble completing tasks that I’m meant to be doing.
20. When I’m feeling good, part of me hates those feelings.
21. When I’m feeling good, my behaviour becomes out of control.
22. I don’t know what to do to create pleasant feelings in myself.
23. When I’m feeling good, I end up neglecting my responsibilities (work, chores, school etc.).
24. When I’m feeling good, I can’t allow those feelings to be there.
25. When I’m feeling good, I have strong urges to do risky things.
26. When I’m feeling good, I have no control over whether that feeling stays or goes.
27. When I’m feeling good, I have difficulty staying focused during important stuff (at work or school, etc.).
28. When I’m feeling good, I believe those feelings are unacceptable.
29. When I’m feeling good, I can’t keep control over myself (in terms of my behaviours).
30. When I’m feeling good, I don’t have any useful ways to help myself keep feeling that way.
31. When I’m feeling good, I have trouble getting anything done.
32. When I’m feeling good, I must try to eliminate those feelings.

Appendix O: Intolerance of Uncertainty Scale-Short Form (IUS-12; Carleton et al., 2007)

Please circle the number that best corresponds to how much you agree with each item. The rating scale is as follows:1 = *not at all characteristic of me*, 2 = *a little characteristic of me*, 3 = *somewhat characteristic of me*, 4 = *very characteristic of me,* 5 = *entirely characteristic of me*.

1. Unforeseen events upset me greatly.
2. It frustrates me not having all the information I need.
3. Uncertainty keeps me from living a full life.
4. One should always look ahead so as to avoid surprises.
5. A small unforeseen event can spoil everything, even with the best of planning.
6. When it's time to act, uncertainty paralyses me.
7. When I am uncertain, I can't function very well.
8. I always want to know what the future has in store for me.
9. I can't stand being taken by surprise.
10. The smallest doubt can stop me from acting.
11. I should be able to organize everything in advance.
12. I must get away from all uncertain situations.

Appendix P: Task Instructions and Post-Task Questions for Studies 1a and 1b

Study 1a: Image-Viewing Task Instructions

For the next task, you will view a series of Instagram images. There is no right or wrong way to complete this task. For example, you DO NOT have to view certain images to fulfil task requirements or properly complete the task. The task is simply to behave as you would normally on Instagram. You can move through the images at your own pace, however, you will be asked some questions about what you thought about the images in general, but not about specific images, so make sure you are paying attention.

The total task duration is fixed and therefore quickly skipping through images will not shorten the task time. In fact, if you skip through the images without paying attention to any of them, you will NOT receive the completion code for the study. Please note you are unable to interact with the images (i.e., use like or comment functions).

Study 1a: Feedback Questions

Have you seen sensitive screens (like the ones in the task you've just completed) on your own Instagram before? Yes/No.

- If YES: Did you behave as you normally would on Instagram (i.e., if you uncovered all sensitive screens, would you normally uncover them all)? Yes/No.
    - If NO: Please explain (open text response) in what ways you behaved differently and why.
- If NO: If you were to come across a sensitive screen (like the ones in the task you've just completed) on your own Instagram, did you behave here as you think you would in real life (i.e., if you uncovered all sensitive screens, would you uncover them all if you were to come across them in real life)? Yes/No.
    - If NO: Please explain (open text response) in what ways you behaved differently and why.

Would you turn off the sensitivity screen feature (i.e., meaning that all images would not be screened when browsing through Instagram) if you had the option to do so? Yes/No.

Instagram has recently introduced a new feature which gives users more control over the photos and videos they see. The control settings are below:

Allow-You may see more photos and videos that could be upsetting or offensive.
Limit (Default)-You may see some photos and videos that could be upsetting or offensive.
Limit Even More-You may see fewer photos and videos that could be upsetting or offensive.

Have you used this feature? Yes/No.

- If YES: Which option did you select?
- If NO: Which option would you select? (Same options as above)

Study 1b: Image-Viewing Task Instructions

For the next task, you will view a series of Instagram images, some of which have been covered by a blur. The task is simply to behave as you would normally on Instagram. You DO NOT have to view certain images to fulfil task requirements or properly complete the task. For example, you may wish to uncover all of the images, none of the images, or only some of the images. However, you will be asked some questions about what you thought about the images in general (not about specific images) so please pay attention.

There will be a 3 sec delay before you can move onto the next image. Please note you are unable to otherwise interact with the images (e.g., use like or comment functions).
Also, please DO NOT take a break during the middle of the task.

Study 1b: Feedback Questions

It is very important for our research that we use data only from people who followed directions exactly. We ask that you answer the following questions honestly to help us analyse our data. Your answers will not affect payment. Thank you for your honesty and for participating in this study.

1. Have you seen sensitive screens covering content on your own Instagram before? Yes/No.
2. Did you read the content descriptions when they were presented with screened images? Yes/No.
3. Did the content descriptions influence your decision to uncover screened images? Yes/No.
    - If YES: Please explain how the content descriptions influenced your decision.

- If NO: Please explain why the content descriptions did not influence your decision.

Study 1b: Demand Question

Before you started the image viewing task, we told you that you were not required to uncover certain images to fulfil task requirements. This is because we are interested in understanding how people typically interact with images covered by sensitive screens. So that we can analyse our data correctly, please respond to the following statements:

I only uncovered the sensitive screens because:

- I thought I was supposed to uncover the screens (i.e., I had no choice). True/False.
- I thought the study might have hidden requirements (e.g., a test of what images I saw). True/False.
- I thought there would be a penalty for not uncovering (e.g., I would be rejected/fail the HIT). True/ False.

Appendix Q: Task Instructions and Post-Task Questions for Study 2

Image-Viewing Task Instructions

For the next task, you will view a series of Instagram images, some of which have been covered by a blur. Some of the blurred images are also accompanied by a description. The task is simply to behave as you would normally on Instagram. You DO NOT have to view certain images to fulfil task requirements or to properly complete the task. For example, you may wish to uncover all of the images, none of the images, or only some of the images. Please note you are unable to interact with the images (e.g., use like or comment functions).

Feedback Questions

It is very important for our research that we use data only from people who followed directions exactly. We ask that you answer the following questions honestly to help us analyse our data. Your answers will not affect payment. Thank you for your honesty and for participating in this study.

Have you seen sensitive screens covering content on your own Instagram before? Yes/No.

Did you read the content descriptions when they were presented with screened images? Yes/No.

Did the content descriptions influence your decision to uncover screened images? Yes/No.

- If YES: Please explain how the content descriptions influenced your decision.
- If NO: Please explain why the content descriptions did not influence your decision.

Demand Question

Before you started the image viewing task, we told you that you were not required to uncover certain images to fulfil task requirements. This is because we are interested in understanding how people typically interact with images covered by sensitive screens. So that we can analyse our data correctly, please respond to the following statements:

I only uncovered the sensitive screens because:

- I thought I was supposed to uncover the screens (i.e., I had no choice). True/False.
- I thought the study might have hidden requirements (e.g., a test of what images I saw). True/False.
- I thought there would be a penalty for not uncovering (e.g., I would be rejected/fail the HIT). True/ False

Appendix R: Task Instructions and Post-Task Questions for Studies 3a and 3b

Study 3a: Image-Viewing Task Instructions

For the next task, you will view a series of Instagram images. These images will appear for a fixed duration (several minutes). You will be asked some questions about the images, so please make sure you are paying attention. Please note you are unable to interact with the images (i.e., use the like or comment functions). You may also see screens where a negative image has been covered. You will not be able to interact with these screens/images.

Study 3a: Feedback Questions

Imagination Questions (anchors from the Vividness of Visual Imagery Questionnaire [VVIQ; Marks, 1973]):

- How vividly did you imagine the content of the screened images?
  - 1 – No image at all (only "knowing" that you are thinking of the object)
  - 2 – Vague, and dim
  - 3 – Moderately clear and vivid
  - 4 – Clear and reasonably vivid
  - 5 – Perfectly clear and as vivid as normal vision

Uncertainty and Curiosity Questions

- How uncertain did you feel about the content of the screened images?
  - 1 – Not at all uncertain
  - 2 – Slightly uncertain
  - 3 – Somewhat uncertain
  - 4 – Moderately uncertain
  - 5 – Extremely uncertain
- How curious did you feel about the content of the screened images?
  - 1 – Not at all curious
  - 2 – Slightly curious
  - 3 – Somewhat curious
  - 4 – Moderately curious
  - 5 – Extremely curious

Study 3b: Image-Viewing Task Instructions

For the next task, you will view a series of Instagram images, some of which have been covered by a blur. The task is simply to behave as you would normally on Instagram. You do not have to view certain images to fulfil task requirements or to properly complete the task. Please note you are unable to interact with the images (i.e., use the like or comment functions).

Appendix S: Task Instructions and Post-Task Questions for Studies 4a and 4b

Introduction to Training Trials

Soon you will view a series of images, including some negative images. Before you do that, we are going to explain two strategies that you can use to manage your emotional responses to these images. We will then ask you to run through some practice trials. You will be reminded what to do for each strategy during the task, but please read the descriptions of each strategy carefully because we will test your understanding of this information.

Emotion Regulation Instructions

Reappraisal

You can use reappraisal to decrease negative emotions by changing the meaning of what is happening in a certain situation. When viewing a negative image, you could try to think of something to tell yourself about the image that helps you feel less negative about it. For example, you could tell yourself something about the outcome, like that whatever is happening will be resolved soon, or that help is on the way. Alternatively, you could focus on a detail in the situation that may not be as bad as it first seemed.

When using reappraisal, it is important that you do not think of random or unrelated things that make you feel better. Rather, you need to change your interpretation of the image in a way that helps you feel less negative about it. It is also very important that you do not think that the image is fake or a scene from a movie. Rather, you need to think of it as a real situation and then change its meaning.

Distraction

You can use distraction to decrease negative emotions by thinking of something completely unrelated to a certain situation. When viewing a negative image, you could use distraction in many ways. For example, instead of thinking about the content of the image, you could picture yourself taking a walk around your neighborhood and think about the different homes and buildings you might see. Alternatively, you could imagine yourself doing everyday tasks, such as brushing your teeth or making breakfast in the morning.

You can use any way to distract yourself that you think will work best in making you feel less negative, and you do not have to distract yourself in the same way every time. Also, when

distracting, it is important that you do not focus on something that is highly emotional. We do not want you to think about anything that brings you extreme sadness or happiness.

Training Trial Instructions

In a moment, you will practice using these two strategies. First, you will see a warning screen for five seconds; this warning will indicate which strategy to use. When we want you to use reappraisal, the instruction will appear in blue, and when we want you to use distraction, the instruction will appear in orange. It will be followed by an image (which will remain on the screen for five seconds). When the image is on the screen, keep your eyes on the image (and do not avert your gaze) while using the specific strategy indicated. Please note you are unable to interact with the images (e.g., use the like or comment functions).

After viewing the image, you will be asked to rate how distressed you feel. Sometimes you may have tried hard to use a certain strategy, but it may not have succeeded in helping you feel better. Please honestly report how you feel at the moment the scale appears. Click the next button (when it appears) to practice the first strategy.

Image-Viewing Task Instructions

You have completed the training trials. You now know how to use the reappraisal and distraction strategies. The main task will follow the same format, but some screens will not have any instruction on them: for these, just look at the image and respond naturally. There will also be a few more images than during training. Click the next button (when it appears) to begin the task.

Study 4a: Feedback Questions

Task Experience Questions

Please respond to the following questions regarding the two strategies you used. Recall, reappraisal required you to think about the content in a way that helped you feel less negative about it, while distraction required you to think about something completely unrelated to the content.

- How easy was it to use the reappraisal strategy? 0 (not at all easy) to 100 (extremely easy)

- How easy was it to use the distraction strategy? 0 (not at all easy) to 100 (extremely easy)
- How effective was the reappraisal strategy in minimising your distress? 0 (not at all effective) to 100 (extremely effective)
- How effective was the distraction strategy in minimising your distress? 0 (not at all effective) to 100 (extremely effective)

Future Use of Emotion Regulation Strategy Question

- Which strategy would you choose if you were given an option to do the task again and could only use one?
    - Reappraisal
    - Distraction
    - I would use a different strategy to manage my emotions
        - What strategy would you use to manage your emotions and why?
    - I would not use any strategy to manage my emotions.
        - Please explain your response (i.e., why you would not use any strategy to manage your emotions)

Past Use of Emotion Regulation Strategy Questions

Now thinking about your real life, and if/when you come across negative and potentially distressing content online:

- How often do you use the reappraisal strategy? (i.e., think about the content in a way that helps you feel less negative about it) 0 (never) to 4 (always)
- How often do you use the distraction strategy? (i.e., think about something completely unrelated to the content) 0 (never) to 4 (always)
- How often do you use another strategy to change how you feel about the content? 0 (never) to 4 (always).
    - *If sometimes or above selected:* What other strategies do you use to change how you feel about the content and why?