PLOS ONE

RESEARCH ARTICLE

Typology of content warnings and trigger warnings: Systematic review

Ashleigh Charles[1‡*], Laurie Hare-Duke[1‡], Hannah Nudds[1], Donna Franklin[2], Joy Llewellyn-Beardsley[1], Stefan Rennick-Egglestone[1], Onni Gust[3], Fiona Ng[1], Elizabeth Evans[4], Emily Knox[5], Ellen Townsend[6], Caroline Yeo[1], Mike Slade[1]

1 School of Health Sciences, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom, **2** Narrative Experience Online Lived Experience Advisory Panel, Nottingham, United Kingdom, **3** Department of History, University of Nottingham, Nottingham, United Kingdom, **4** School of Cultures, Languages, and Area Studies, University of Nottingham, Nottingham, United Kingdom, **5** School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, Illinois, United States of America, **6** School of Psychology, University of Nottingham, Nottingham, United Kingdom

‡ AC and LHD are joint first authors on this work.
* ashleigh.charles@nottingham.ac.uk

Citation: Charles A, Hare-Duke L, Nudds H, Franklin D, Llewellyn-Beardsley J, Rennick-Egglestone S, et al. (2022) Typology of content warnings and trigger warnings: Systematic review. PLoS ONE 17(5): e0266722. https://doi.org/10.1371/journal.pone.0266722

Editor: Michelle L. Munro-Kramer, University of Michigan, UNITED STATES

Received: May 24, 2021

Accepted: March 25, 2022

Published: May 4, 2022

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: YES - This article is independent research funded by the NIHR under its Programme Grants for Applied Research Programme (Programme Grants for Applied Research, Personal experience as a recovery resource in psychosis: Narrative Experiences Online (NEON) Programme, RP-PG-0615-20016). The funders had no role in study design, data collection and analysis, decision to

Abstract

Content and trigger warnings give information about the content of material prior to receiving it. Different typologies of content warnings have emerged across multiple sectors, including health, social media, education and entertainment. Benefits arising from their use are contested, with recent empirical evidence from educational sectors suggesting they may raise anxiety and reinforce the centrality of trauma experience to identity, whilst benefits relate to increased individual agency in making informed decisions about engaging with content. Research is hampered by the absence of a shared intersectoral typology of warnings. The aims of this systematic review are to develop a typology of content warnings and to identify the contexts in which content warnings are used. The review was pre-registered (ID: CRD42020197687, URL: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020197687) and used five sources: electronic databases covering multiple sectors (n = 19); table of contents from multi-sectoral journals (n = 5), traditional and social media websites (n = 53 spanning 36 countries); forward and backward citation tracking; and expert consultation (n = 15). In total, 6,254 documents were reviewed for eligibility and 136 documents from 32 countries were included. These were synthesised to develop the Narrative Experiences Online (NEON) content warning typology, which comprises 14 domains: Violence, Sex, Stigma, Disturbing content, Language, Risky behaviours, Mental health, Death, Parental guidance, Crime, Abuse, Socio-political, Flashing lights and Objects. Ten sectors were identified: Education, Audio-visual industries, Games and Apps, Media studies, Social sciences, Comic books, Social media, Music, Mental health, and Science and Technology. Presentation formats (n = 15) comprised: education materials, film, games, websites, television, books, social media, verbally, print media, apps, radio, music, research, DVD/video and policy document. The NEON content warning typology provides a framework for consistent warning use and specification of key contextual information (sector, presentation

publish, or preparation of the manuscript. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. AC acknowledges the support of the Economic and Social Research Council (grant number ES/P000711/1). MS acknowledges the support of the Center for Mental Health and Substance Abuse, University of South-Eastern Norway and the NIHR Nottingham Biomedical Research Centre.

Competing interests: All authors declare no competing interests.

format, target audience) in future content warning research, allowing personalisation of content warnings and investigation of global sociopolitical trends over time.

Introduction

Trigger warnings and content warnings are "*a statement at the start of a piece of writing, video, etc. alerting the reader or viewer to the fact that it contains potentially distressing material–often used to introduce a description of such content*" (p.602) [1]. This dual use–both warning about, and characterising the content of, material–may account for the terms 'trigger warning' and 'content warning' being used interchangeably across the diverse academic disciplines which research their use. Some but not all studies locate trigger warnings as a particular sub-type of content warnings which are focussed specifically on the needs of people with experience of trauma or post-traumatic stress disorder (PTSD). In addition, the use of 'trigger' has been disputed because what constitutes a 'trigger' in relationship to PTSD is highly individual and unpredictable [2]. In this paper we therefore use the general term 'content warning'.

The use of content warnings is longstanding and widespread across multiple sectors, e.g. health, education, media, arts and literature. The concept of 'triggering' describes the re-experiencing of unpleasant PTSD symptoms such as intrusive thoughts being evoked by exposure to materials which spark traumatic memories. Hence, content warnings have a long presence in psychiatric literature [3]. Similarly, most countries use classification systems for film and television media, such as Australia [4] and Singapore [5]. For example, a film rating classification system has been used in the United Kingdom (UK) since 1912. In education, use of content warnings is widespread [6], e.g. a 2016 survey of professors in the United States of America (USA) found that 428 (51%) of the 841 respondents reported some use of content warnings in their classes [7]. Content warnings have also been advocated and used in other sectors, including books [8], comics [9] and museums [10]. Multinational frameworks have been established, such as the Pan European Game Information (PEGI) categories of age labels [3, 7, 12, 16, 18] and content descriptors (Bad language, Discrimination, Drugs, Fear, Gambling, Sex, Violence, In-game purchases) for computer games.

There is no inter-sectoral consensus or widely used typology of content warnings. Incompatible frameworks have been developed, driven by the perceived need for different uses and in different legal and geographical jurisdictions. One reason for differences in content warning is the un-coordinated development of frameworks across different sectors. Each sector contains different assumptions, so for example in the arts sector provocation caused by displayed materials is valued and intended [11], with content warnings primarily giving information so individuals can make their own choices about exposure. By contrast, in the health sector the biomedical ethical imperative of non-maleficence [12] places more responsibility on the health professional to actively reduce the likelihood of triggering trauma responses in people with experience of PTSD. Hence, the content warnings in arts spaces tend to be more focussed on information, whereas those in a health context contain less information to avoid the content warning itself being triggering.

The development of shared inter-sectoral practices such as consensus on a content warning typology is made more difficult because of the contested evidence base. Broadly, some researchers argue that content warnings are a form of over-protection which inadvertently hinder the development of resilience [13], whereas others argue that content warning allows both avoidance of un-wanted exposure experiences and emotional preparation to reduce

negative reaction to content [14]. More recent studies have experimentally investigated the impact of content warnings, especially in educational settings. A randomised study found participants with no trauma history (n = 133) who received warnings before reading passages with disturbing content reported more anxiety than those not receiving (n = 137) warnings, suggesting warnings can undermine emotional resilience [15]. The same authors replicated this finding with a college student sample (n = 462) [16], and also showed in a randomised study of trauma survivors (n = 451) that content warnings inadvertently reinforce the centrality of trauma experiences to identity [17]. A 2019 meta-analysis confirmed this finding that content warnings are associated with increased anxiety and negative mood [1]. Meta-analyses of a series of studies involving students and internet volunteers, with and without a trauma history, found mainly neutral or slightly negative impact of content warnings, leading the authors to conclude that such warnings are neither meaningfully helpful or harmful [18].

Given the emerging empirical evidence base, and the publication of balanced and authoritative overviews [19], why is there difficulty in reaching consensus? Despite early suggestions in the 2000s of a new interdisciplinary area emerging called warning research [20], and signs of social media conventions developing [21, 22], much of the public discussion about content warnings has been heated. The use of emotive language such as the 'trigger warning war' [23] is perhaps related to wider cultural debates and potentially clashing priorities, for example free speech and censorship approaches versus trauma-informed and rights-informed approaches. The issue is international, and passionate articles have been written by authors in Australia [24], Ireland [25], the UK [26] and the USA [27]. For example, some have argued that, whatever the experimental evidence, content warnings are a vital approach to increasing inclusivity on campuses because they can support engagement by people who would otherwise avoid material due to past experiences [28]. Others cite evidence that content warnings can be useful in specific educational contexts, such as victimology courses [29].

An inter-sectoral typology of content warnings would advance the field, in three ways. First, by allowing the comparability of findings from across different studies to be maximised. At present there is no recommended typology, and so empirical studies use different content warnings for the same content. Identifying an agreed description for a warning and locating it in a broader and coherent typology will support more fine-grained investigation about the impact of specific types of warning. In addition, the typology will be of interest to a range of fields, including arts, media studies, medicine, mental health, and psychology. Second, the limits of content warnings would be helpful to establish, to understand what is in scope. For example, are 'Contains nuts', 'Contains flashing imagery' and the perhaps ironic 'Depicts killer robots' all appropriately understood as content warnings? Finally, identifying the typology, sector, presentation format and target population will allow more specific future research into the positive and negative impacts of content warning when used in a specific sector with a specific target population.

The aims of this review are (1) to develop a typology of content warnings, and (2) to identify the contexts in which content warnings are used, comprising the sector (e.g. education, health), format (e.g. film, music) and target audience.

Method

The protocol of this systematic review was developed in accordance with PRISMA guidelines and was registered on PROSPERO (International Prospective Register of Systematic Reviews) on 9 July 2020 (reference CRD42020197687). The study was conducted as part of the Narrative Experiences Online (NEON) study (http://www.researchintorecovery.com/neon), which aims to evaluate the impact of recorded recovery narratives [30] when used as a mental health

intervention [31]. In developing the intervention [32], we needed to decide whether to use content warnings when delivering recorded recovery narratives to participants in randomised controlled trials for people with experience of psychosis (NEON Trial: ISRCTN11152837), for people with experience of non-psychosis mental health problems (NEON-O Trial: ISRCTN63197153), and for informal mental health carers (NEON-C trial: ISRCTN76355273) [33]. The typology developed in this review is therefore called the NEON content warning typology.

Eligibility criteria

Studies were included where the document (a) reported a text-based list, set or typology of content warnings, or (b) presented empirical evidence or a structured framework regarding the context (e.g. target audience, sector, and presentation format) for the use of specific content warning. We included both peer-reviewed and non-peer reviewed literature, including empirical studies of any design (e.g. surveys, experiments, qualitative interviews), commentaries, opinion pieces, media organisation web-pages, and systematic and non-systematic literature reviews. No restrictions were placed on the population of study. We excluded studies that were non-English language documents and published before the year 2000 as a high-quality review was published in 2002 [34].

Information sources

Five data sources were used:

1. Electronic bibliographic databases (n = 19) from a range of academic disciplines and sectors were searched: ACM Digital Library, Applied Social Sciences Index and Abstracts (ASSIA), CINAHL, Education Database, Education Resources and Information Center (ERIC), e-theses online service (EThOS), IEEE Xplore, International Bibliography of the Social Sciences (IBSS), JSTOR, Library & Information Science Source, MEDLINE, Project MUSE, ProQuest Dissertation and Theses Global, PsycINFO, PubMed, Sociological Abstracts, Social Science Database, Sociology Database, and OpenGrey.
2. The table of contents from journals (n = 5) spanning multiple sectors as selected by topic experts: Communication Education; Feminist Teacher; Game Studies; Participations: Journal of Audience & Reception Studies; and Suicide and Life-Threatening Behaviour.
3. Traditional and social media organisation websites (n = 53) relating to 36 countries (Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Denmark, Finland, France, Germany, Hong Kong, Iceland, India, Indonesia, Ireland, Jamaica, Japan, Malaysia, Maldives, Netherlands, New Zealand, Nigeria, Norway, Malta, Philippines, Poland, Saudi Arabia, Singapore, South Africa, South Korea, Sweden, Taiwan, Turkey, United Arab Emirates, United Kingdom and United States of America) and 12 international organisations (Amazon, Apple Store, Blackberry, Entertainment Software Rating Board (ESRB), Facebook, Google Play, International Age Rating Coalition (IARC), Netflix, Pan European Game Information (PEGI), Reddit, Twitter, YouTube). These were identified through online searching to find English-language classification guidance and are listed in S1 Appendix.
4. Forward citation tracking was performed on all included studies using Google Scholar and backward citation tracking by hand-searching reference lists of all included studies.
5. Consultation with the multidisciplinary authorship team (n = 11) and other international experts (n = 4) was used to identify additional key documents or studies.

Search strategy

Through a combination of a preliminary scoping search and expert consultation we identified five sectors which we aimed to cover in our search strategy: education (primary, secondary and tertiary), healthcare (medicine, and mental health/psychology), media (television, film, social media and gaming), human rights (feminist and gender studies, critical race studies), and digital technologies (computing and information sciences). Search terms used were 'advisory warning', 'content note', 'content notice', 'content warning', 'trauma trigger', 'trigger warning', and 'video nasty'. The search terms were modified for each database, for example the search strategy used for PsycINFO and MEDLINE was: ("content warning" OR "trigger warning" OR "content notice" OR "content note" OR "advisory warning" OR "trauma trigger" OR "video nast*").ti,ab. All database searches were conducted from the year 2000 to the search date.

Study selection

For all searches, identified citations were collated and uploaded to EndNote X9. After removing duplicates, the titles and abstracts of all identified citations were screened for relevance against the inclusion criteria by three analysts (AC, HN and LHD) with a randomly-selected subsample (5%) independently assessed by LHD. Concordance between reviewers was 100%. Full texts were screened by AC, HN and LHD with a randomly-selected subsample independently assessed by LHD (10%), showing 100% concordance.

Data abstraction

For each document, information was extracted on:

1. Document characteristics, comprising publication year, peer reviewed journal, publication type, sector, and country of affiliation of the lead/first author
2. Content/trigger warning list characteristics, comprising function of list (listing categories vs. discussing warning use in specific contexts), type of warning list ('trigger warning' vs. 'content warning' or synonyms), warning labels (the list of warnings), source reference, presentation format (e.g. 'classroom', 'film'), target audience, number of lists in the document
3. Exclusion/inclusion criteria, sample size, and study design [empirical studies only]
4. Country and setting of list use.

The data abstraction table (DAT) is shown in S2 Appendix. The DAT was piloted by AC and LHD who independently abstracted a randomly selected 10% of the included documents. Concordance was 92%. The DAT was then refined following discussion between reviewers and agreement was reached on developing instructions for further abstraction. AC, DF, HN and LHD then independently extracted data from the remaining documents.

Quality assessment

The quality of the included documents which reported an empirical study were assessed using the Mixed Methods Appraisal Tool (MMAT) [35]. The MMAT has sections for different types of study, each with its own set of methodological quality criteria: (1) qualitative; (2) quantitative randomised controlled trials; (3) quantitative non-randomized; (4) quantitative descriptive; and (5) mixed methods. For each item the answer categories were 'Yes', 'No', or 'Can't tell' followed by comments. The MMAT was chosen as it covers the range of empirical studies involved in this review and has moderate-to-excellent inter-rater reliability. No quality

assessment was made of those documents not reporting empirical research as there are no standard criteria or processes for assessing the quality of such documents.

Data synthesis

Data synthesis was conducted on included papers. The two primary analysts (AC, MS) came from different mental health professional (nursing and clinical psychology) and academic (sociology and health research) backgrounds, and the other analysts (LHD, HN, DF, JLB, SRE, OG, FN, EE, EK, ET, CY) came from varied disciplinary and sectoral backgrounds (heath service research, psychology, sociology, youth work, social science, history, media studies).

For objective 1 (developing a typology of content warnings), warning lists and items from all the included studies were synthesised using the following process: a) duplicates of content warning items were removed; b) content warnings which perpetuate structural inequality were removed (the only instance found was a warning about same-sex marriage, parenting or sexual activity found in six included documents); c) labels (codes) were developed for important features in the data; d) initial categories were created by examining the codes and identifying significant broader patterns of meaning; e) categories were reviewed through checking the candidate categories against the dataset, in order to determine whether they closely mapped onto the data. Once confirmed, the wording of the label was reviewed to ensure maximum comprehensibility, e.g. anti-disability was chosen over ableism; and f) vote counting of number of papers identifying each category was performed to establish the strength of each category. The preliminary analysis was conducted by AC and MS. The analysis was then iteratively refined in discussion with all co-authors.

Results

The search identified 6,254 documents, from which 136 were included. The flow diagram is shown in Fig 1.

The data abstraction table for all included documents is shown in S2 Appendix. The 136 included documents comprised webpages (n = 43), academic journals (n = 38), newspaper/magazines (n = 36), technical reports (n = 12), books (n = 4), conference abstracts (n = 2), and a thesis (n = 1). Of these, 38 (28%) documents were peer-reviewed studies and 98 (72%) were non-peer reviewed. The included documents came from 32 countries covering North America (n = 74), Europe (n = 22), international which covered multiple countries (n = 15), Asia (n = 14), Australasia (n = 4), Africa (n = 3), both Europe and USA (n = 1), and South America (n = 1). The majority of documents came from the USA (n = 69) (51%).

Only 18 (10%) documents reported empirical studies, comprising observational (n = 9), experimental (n = 6), qualitative (n = 2), and mixed methods (n = 1) designs. As a result of this small proportion, it was not appropriate to conduct a sensitivity analysis comparing higher to lower quality studies. Therefore, the MMAT ratings, although retained in the data abstraction table, were not used in the analysis.

Across the 136 documents, 330 warning lists were identified containing a total of 2,209 warnings (including duplicates). The documents comprised 52 (16%) defining a list of warnings and 278 (84%) outlining specific warnings for use in a specific context, e.g. classroom. Across all lists, 134 (41%) were labelled as trigger warnings and 196 (59%) as content warnings.

Objective 1: NEON content warning typology

The 2,209 warnings were synthesised into 14 categories of content warning, as shown in Table 1.

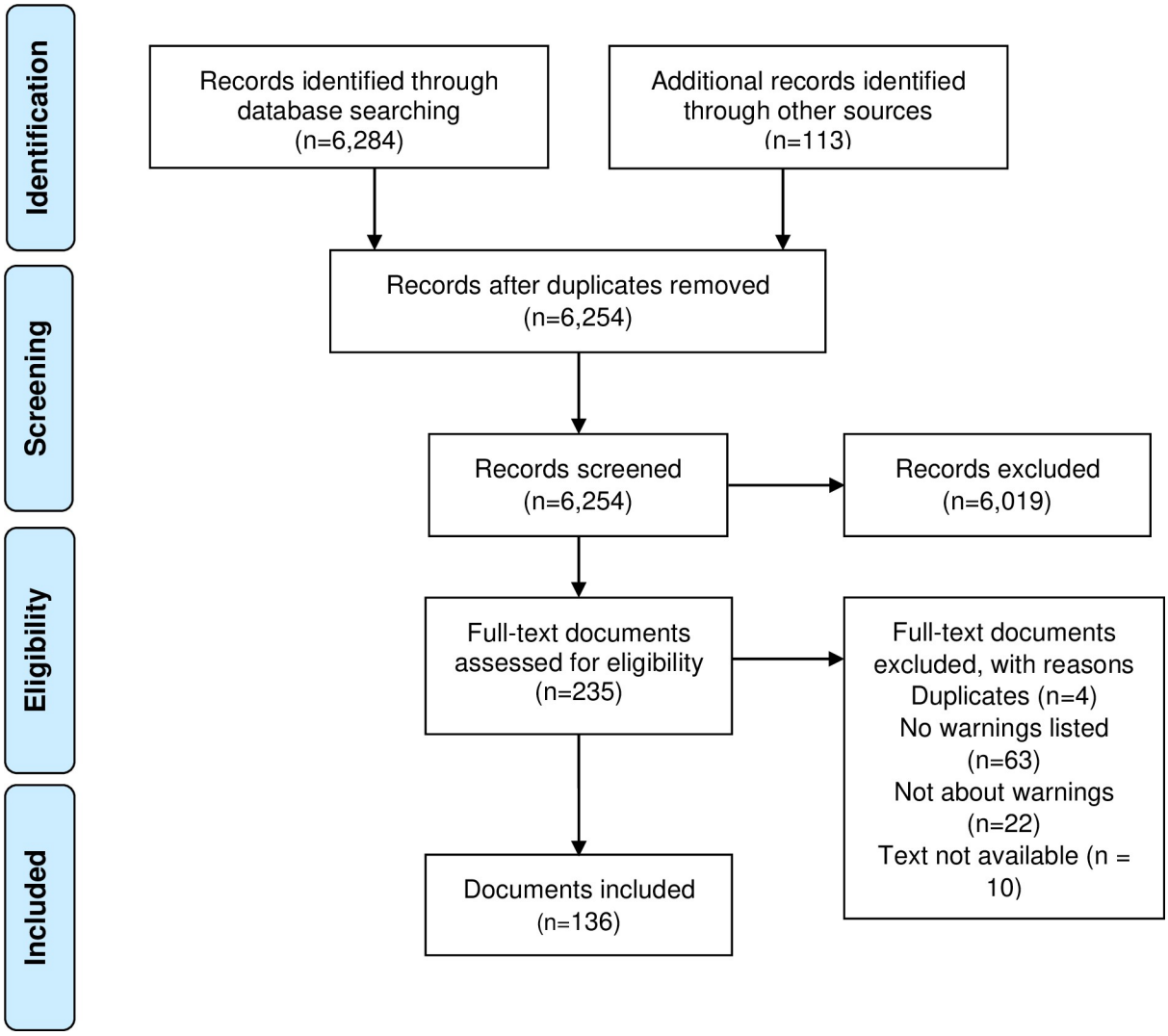

Fig 1. Flow diagram for included documents.

https://doi.org/10.1371/journal.pone.0266722.g001

The complete NEON content warning typology, including definition, number of documents using the category, and example text for all sub-categories, is shown in S3 Appendix.

Objective 2: Contexts for content warning use

Ten sectors using content warnings were identified in the 136 documents: education (n = 53), audio-visual industries (n = 34), games and apps (n = 18), media studies (n = 4), social sciences (n = 3), comic books (n = 3), social media (n = 3), music (n = 2), mental health (n = 2), science and technology (n = 1). Thirteen documents spanned multiple sectors. Compared to other sectors, Education had the most frequent mentions of warnings relating to Violence, Sex, Stigma, Risky behaviour, Mental health, Abuse, Crime, Socio-political, and Objects. Audio-visual industries were the highest users of warnings about Disturbing content, Language, Death, and Parental guidance. Games and apps were the sector most using warnings about Flashing lights.

Table 1. NEON content warning typology.

Category (n) and definition	Sub-categories
1. Violence (n = 536)	Violence; War; Weapons; Terrorism; Police brutality; Motiveless killing; Sexual violence; Animal cruelty; Torture; Genocide
Content contains violence	
2. Sex (n = 332)	Nudity; Mild sexual content; Explicit sexual content; Relationship conflict; Reproductive health
Content contains sexual themes, including nudity, sexual content and relationships	
3. Stigma (n = 328)	Racism; Anti-religious (sub-categories: Anti-Semitic; Anti-Christian; Islamophobia); Colonialism; (sub-category: Slavery); Classism; Sexism (sub-categories: Misogyny; Misandry); Transphobia; Gender-identity; Sexuality (sub-category: Homophobia); Anti-disability
Content depicts negative stereotypes about or attitudes towards a specific group, such as racism or sexism	
4. Disturbing Content (n = 236)	Disturbing content with threat; Horror and terror; Disturbing imagery; Medical content; Human bodies and functions
Content contains imagery, sounds, or effects that may frighten, disgust or scare	
5. Language (n = 235)	Sexual language; Adult humour; Swearing; Offensive language
Content contains language which is sexual, crude or offensive	
6. Risky Behaviours (n = 200)	Drug misuse; Alcohol misuse; Tobacco; Gambling
Content depicts risky lifestyle behaviours	
7. Mental Health (n = 108)	Mental health; Eating disorders; Trauma; Self-harm and suicide; Depression; OCD; Panic attacks; Anxiety (sub-categories: Spiders; Snakes; Insects; Needles; Eye contact; Irregular patterns); Hair pulling
Content relates to mental health issues	
8. Death (n = 49)	Death; Accidents; Natural disasters
Content relates to human death or dying	
9. Parental Guidance (n = 47)	Online access; Cyber-bullying; Competitive content; Imitative content; Upsetting content; Non-realistic violence
Content may not be appropriate for children	
10. Crime (n = 38)	
Content depicts or relates to criminal activity	
11. Abuse (n = 37)	Child abuse; Emotional abuse; Physical/sexual abuse; Neglect
Content depicts or relates to abuse	
12. Sociopolitical (n = 27)	Injustice; Political issues; Nazism; Class issues
Content includes social or political issues	
13. Flashing Lights (n = 27)	
Content includes strobe or flashing lighting	
14. Objects (n = 4)	
Content contains specific objects	

https://doi.org/10.1371/journal.pone.0266722.t001

A total of 15 presentation contexts for the use of content warnings were identified in the 330 lists: education materials (n = 95), film (n = 48), games (n = 43), websites (n = 32), television (n = 16), books (n = 15), social media (n = 9), verbally (n = 5), print media (n = 4), apps (n = 3), radio (n = 2), music (n = 2), research (n = 1), DVD/video (n = 1), policy documents (n = 1) and multiple contexts (e.g. classroom and online forums) (n = 59). Educational materials was the most frequently used presentation format for content warnings relating to Violence, Sex, Stigma, Mental health, Abuse, Socio-political, and Objects. Disturbing content, Language, and Risky behaviours content warning categories were most frequently used in Games. Film was the most used presentation format for Death, Parental guidance and Crime.

The target audience in the 330 warning lists comprised mainly the general public (n = 197) and students (n = 119), with other recipient groups identified including children and

adolescents (n = 8), women (n = 3), parents (n = 2), and people with post-traumatic stress disorder (PTSD) (n = 1). Generally, the included papers did not explicitly state the target audience for the content warning, so the target audience was recorded as general public. Students were a highly cited target audience across all warning categories, and were more cited than the general public for the Mental health warning. For the remaining thirteen warning categories, the most cited target audience was the general public.

Discussion

This systematic review included documents from 32 countries and developed the NEON content warning typology comprising 14 categories of warning. The categories of Violence, Sex and Stigma were the most widely used warnings, but some warnings were more sector-specific, such as Parental Guidance in relation to films. Ten distinct sectors in which content warnings are used were identified, along with 15 distinct presentation formats. The target audience for content warning usage was often not explicitly stated.

Several principles informed the typology development. First, content warnings relating to media content and not product safety were included. For example, product labelling for food (e.g. Contains nuts) and cleaning products (e.g. Keep out of reach of children) were deemed out of scope in this review. The category of Flashing Lights was a boundary condition which was included due to its use in films, even though it is also used on consumer products such as toys.

Second, the need for accessible language informed label choices where possible, e.g. Anti-disability was chosen over Ableism. However, preference was given to retaining the meaning, e.g. Misogyny and Misandry were selected since labels such as Anti-female and Anti-male did not adequately capture the full meaning of gender-based prejudice. These terms are spreading due to increased use and the increasing rights literacy among younger people [36].

Finally, the NEON content warning typology is designed to be extendable in scope and depth. This may be needed for several reasons.

a. The decision to exclude warnings that perpetuate structural inequality as part of the synthesis led to the exclusion of a Same-sex relationship warning, in order to avoid replicating heteronormative assumptions and prejudice. Alternative or additional approaches to preventing the perpetuation of structural inequality could be considered.

b. The categories in the typology are not wholly distinct. For example, Anti-Semitic was positioned in the typology as a sub-category of Anti-religious due to its use in included documents, but could also be located as a form of Racism. Similarly, there is an inter-linkage between the Stigma and Sociopolitical categories, and the challenges of locating categories in either (e.g. Racism and Sexism in Stigma and Injustice in Sociopolitical) or both (e.g. Classism in Stigma and Class issues in Sociopolitical) reflect that these complex issues can be viewed both as individual and structural problems. Similarly, warnings about depictions of violence in relationships may draw from both the Violence and Abuse categories.

c. The typology was constrained to reflect the content of included documents. This meant that some categories would benefit from future disaggregation or elaboration. For example, Political issues incorporates both pro-capitalist and anti-capitalist content, so a future iteration may develop sub-categories to differentiate these two very different concerns. The Risky Behaviours category currently does not include any online behaviours, which may be a future extension. Other candidate extensions include a Pro-religious warning in the Stigma category and a Sexualising of children warning in the Abuse category.

Globally, the translation of content warnings is an important consideration. For example, content warnings in media and printed materials may vary between differing countries and resource settings due to specific cultural and contextual factors, mediated by the power held by organisations that specify the use of content warnings, such as film classification bodies. Examples might include different thresholds in relation to depictions of sexual behaviour, drug use and clothing, which can change over time. In addition, content warnings may not be used in some settings, highlighting the prominence of content warnings in particular contexts such as higher education. The translation and use of content warnings in other settings will need to be considered. Methodologies now exist to support the proportionate translation of the content warnings into other languages [37]. In relation to use, when presenting stories in different settings, as will be done in the NEON study, specific cultural or contextual content warnings may need to be included to meet the needs of recipients from different cultural backgrounds.

Overall, the typology is designed to be extended, as social norms evolve and as research into content warnings develops. The NEON content warning typology reflects current inter-sectoral research and practice, and so refinement within its overall structure is actively encouraged and recommended.

Strengths and limitations

Several strengths can be identified. This is the first systematic review proposing an inter-sectoral typology. Previous systematic reviews have focused on individual warnings for specific products, such as cigarettes [38] or alcohol [39]. A second strength of the NEON content warning typology is that it is based on current practice, inter-sectoral research and literature, not within-sector expertise. For example, the Mental health sub-category of Irregular patterns was included which is a rare problem from a psychiatric epidemiological perspective, whereas diagnoses such as Claustrophobia are more prevalent. Starting from a mental health sectoral perspective could lead to simply reproducing sector-specific taxonomies as a content warning typology, which would limit inter-sectoral applicability. Further strengths include the wide range of sources from multiple sectors and countries, and the cross-disciplinary multi-analyst approach used in data synthesis.

Limitations include the findings that the majority (51%) of included documents came from the USA, reflecting the strong focus on content warnings in that country but raising the question of generalisability of findings. The extendibility of the NEON content warning typology is an important feature to address potential ethnocentrism. A second limitation is the inability to use the study quality metrics, due to the small proportion of scientific studies in the included documents. A third limitation relates to the inclusion criteria requiring papers to use text-based warnings, which excluded other types of warning such as graphic icon images. Future work might develop icons for each of the 14 categories in the typology.

Implications

The study has three implications for future research. First, this review can support the emergence of a coherent and aggregable evidence base allowing the impact of content warnings to be more systematically investigated. The use of content warnings relates to other safety-driven initiatives. A recent systematic review investigating removal or blurring of self-harm online imagery found parallel issues of potential harm and positive impacts [40], reinforcing that research in content warnings and cognate areas is complex. There is a need to move beyond the current somewhat simplistic focus on whether content warnings help or harm towards a contextualised understanding of the mechanisms by which content warnings impact on recipients with different characteristics in different sectors and when used for different purposes.

Future content warning studies should clearly describe the specific warnings used and the context including sector, presentation format and target audience. Using the NEON content warning typology and identified usage descriptors such as the names for the ten sectors will increase opportunities to integrate the currently diverse evidence base. A recommended citation terminology for use in studies using the NEON content warning typology would be identifying the specific warning(s) using the category number and name (e.g. 'NEON 1.3 Weapons') and specifying the sector, format and target audience involved in the research. This review found no consensus about the use of the term 'trigger warning' versus 'content warning', so to ensure relevant studies can be easily located in future reviews it is recommended that abstracts for relevant studies always include the term 'content warning'. Overall, these recommendations will support the development of a more robust and integrated scientific evidence base, which can illuminate issues such as the optimal granularity of a warning (Violence? Motiveless killing?) and the mechanisms of impact on different groups in different contexts.

A second and related future focus could be on a more fine-grained perspective about differential impacts of content warnings on specific target audiences. Most included documents did not explicitly specify the audience, which is an important omission. For example, there is emerging evidence that post-traumatic growth in psychosis is more common than expected [41, 42], which may inform the use or not of content warnings with this clinical population. Similarly, greater clarity about the relationship between positive and negative impacts of warnings on specific audiences would allow personalisation of individually delivered content warnings in online apps, games, and web-browsing. For example, web-browser extensions have been developed which automatically detect potentially sensitive content warnings for specific topics, with the goal of creating safe internet spaces for users [43]. Whilst The NEON content warning typology presents a list of content warnings, it is limited to reflect the included documents. However, the typology could inform creators of such technologies about what content warnings exist and what topics and/or words need to be included and screened for. Similarly, for audiences, future web-based innovations may allow individuals to choose what specific content warnings to screen that meets their needs, and the typology can support individuals in deciding which specific topics/words to include or exclude. Further personalisation of content may also include recording recipient characteristics as part of a personal profile on a smartphone which would allow apps to be tailored to include content warnings tailored to individuals with particular characteristics, or to include content warnings only for some recipients based on their personal profile.

A third area of research is to investigate secular or time trends in the use of different types of warning. For example, in this study Stigma was one of the most widely used categories, which may reflect an increasing global focus on issues of rights and discrimination, and the links between being a member of a stigmatised group and experiencing trauma [44]. Future research might investigate whether the changing pattern of content warning use over time can be used as a barometer to capture global socio-political trends, such as the increasing recognition of the ongoing trauma effected by historical colonialism, racism, and trans-Atlantic slavery. Changing patterns of content warning use may in this way illuminate wider societal changes, giving a new source of data for social science research seeking to characterise the evolution of societal values and priorities.

Finally, the next stage in research may involve addressing the short-term nature of content warnings. The NEON content warning typology provides a framework for alerting to potential distressing and/or sensitive content, but in other environments such as academia, content warnings are provided alongside faculty and student service support. Evaluation will need to identify if any additional resources beyond the typology alone are needed to support recipients

when content warnings are in different settings, and whether the typology is useful for recipients in deciding what content to engage with.

Overall, the NEON content warning typology is an empirically-defensible theoretical foundation for future content warnings research. It has been developed by analysts from diverse professional and academic perspectives, enhancing its intersectoral applicability. Using the typology will support investigation addressing the important and currently under-researched question of how different types of content warnings impact on different audiences in different sectors.

Supporting information

S1 Checklist. PRISMA 2020 checklist.
(DOCX)

S1 Appendix. Media.
(DOCX)

S2 Appendix. Data abstraction table.
(XLSX)

S3 Appendix. Complete coding framework.
(DOCX)

Acknowledgments

We thank Ms Emma Young (Nottinghamshire Healthcare NHS Foundation Trust) for support.

Author Contributions

Conceptualization: Ashleigh Charles, Laurie Hare-Duke, Joy Llewellyn-Beardsley, Mike Slade.

Data curation: Ashleigh Charles, Laurie Hare-Duke.

Formal analysis: Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Onni Gust, Fiona Ng, Elizabeth Evans, Emily Knox, Ellen Townsend, Caroline Yeo, Mike Slade.

Funding acquisition: Mike Slade.

Methodology: Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Mike Slade.

Writing – original draft: Ashleigh Charles, Laurie Hare-Duke, Mike Slade.

Writing – review & editing: Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Onni Gust, Fiona Ng, Elizabeth Evans, Emily Knox, Ellen Townsend, Caroline Yeo, Mike Slade.

References

1. Bridgland V, Green D, Oulton J, Takarangi M. Expecting the Worst: Investigating the Effects of Trigger Warnings on Reactions to Ambiguously Themed Photos. Journal of Experimental Psychology: Applied. 2019; 25:602–17. https://doi.org/10.1037/xap0000215 PMID: 30843709
2. Duggan L. Taking Offense: Trigger Warnings and the Neo-liberal Rhetoric of Endangerment. Bully Bloggers. 2014.

3. Haslam N. What's the difference between traumatic fear and moral anger? Trigger warnings won't tell you: The Conversation; 2017 [Available from: https://theconversation.com/whats-the-difference-between-traumatic-fear-and-moral-anger-trigger-warnings-wont-tell-you-77365.
4. ABC. Associated standard on TV programm classification Australia 2014 [Available from: https://edpols.abc.net.au/associated-standard-on-tv-program-classification/.
5. INFOCOMM Media Development Authority. Films Singapore 2020 [Available from: https://www.imda.gov.sg/regulations-and-licensing-listing/content-standards-and-classification/standards-and-classification/films.
6. Kimble M, Flack W, Koide J, Bennion K, Brenneman M, Meyersburg C. Student reactions to traumatic material in literature: Implications for trigger warnings. PLoS One. 2021; 16:e0247579. https://doi.org/10.1371/journal.pone.0247579 PMID: 33765044
7. Kamenetz K. Half Of Professors In NPR Ed Survey Have Used 'Trigger Warnings' 2016 [Available from: https://www.npr.org/sections/ed/2016/09/07/492979242/half-of-professors-in-npr-ed-survey-have-used-trigger-warnings?t=1614958676468&t=1614959066521.
8. Gold J. Content Warnings: How and What to Include? 2019 [Available from: https://jamigold.com/2019/08/content-warnings-how-and-what-to-include/.
9. Vitagliano E. Warning: explicit content 2003 [Available from: https://afajournal.org/past-issues/2003/january/warning-explicit-content/.
10. NCAC Arts Advocacy Program. Museum best practices for managing controversy 2019 [Available from: https://ncac.org/resource/museum-best-practices-for-managing-controversy.
11. Benford S, Greenhalgh C, Crabtree A, Flintham M, Walker B, Marshall J, et al. Performance-led research in the wild. ACM Transactions on Computer-Human Interaction (TOCHI). 2013; 20(3):1–22.
12. Beauchamp T, Childress J. Principles of Biomedical Ethics. Oxford: Oxford University Press; 2001.
13. Lukianoff G, Haidt J. The Coddling of the American Mind: How Good Intentions and Bad Ideas Are Setting Up a Generation for Failure. London: Penguin; 2018.
14. Lockhart E. Why trigger warnings are beneficial, perhaps even necessary. First Amendment Studies. 2016; 50:59–69.
15. Bellet B, Jones P, McNally R. Trigger warning: Empirical evidence ahead. J Behav Ther Exp Psychiatry. 2018; 61:134–41. https://doi.org/10.1016/j.jbtep.2018.07.002 PMID: 30077703
16. Bellet B, Jones P, Meyersburg C, Brenneman M, Morehead K, McNally R. Trigger Warnings and Resilience in College Students: A Preregistered Replication and Extension. Journal of Experimental Psychology: Applied. in press: https://doi.org/10.1037/xap0000270 PMID: 32281813
17. Jones P, Bellet B, McNally R. Helping or Harming? The Effect of Trigger Warnings on Individuals With Trauma Histories. Clinical Psychological Science. 2020; 8:905–17.
18. Sanson M, Strange D, Garry M. Trigger Warnings Are Trivially Helpful at Reducing Negative Affect, Intrusive Thoughts, and Avoidance. Clinical Psychological Science. 2019; 7:778–93.
19. Knox E. Trigger warnings. History, Theory, Context. Maryland: Rowman & Littlefield; 2017.
20. Rogers W, Lamson N, Rousseau G. Warning Research: An Integrative Perspective. Hum Factors. 2000; 42:102–39. https://doi.org/10.1518/001872000779656624 PMID: 10917149
21. Turner H. A guide to content and trigger warnings: The Mix; 2020 [Available from: https://www.themix.org.uk/mental-health/looking-after-yourself/a-guide-to-content-and-trigger-warnings-37946.html.
22. Many L. How to write a trigger warning?: Lookslikefilm; 2019 [Available from: https://www.lookslikefilm.com/2019/01/27/how-to-write-a-trigger-warning/.
23. Hui A. The Trigger Warning War at Harvard: Harvard Crimson; 2019 [Available from: https://www.thecrimson.com/article/2019/9/26/trigger-warning-great-article-ahead/.
24. Palmer T. Monash University trigger warning policy fires up free speech debate: ABC News; 2017 [Available from: https://www.abc.net.au/news/2017-03-28/monash-university-adopts-trigger-warningpolicy/8390264.
25. Malervy R. Castigating trigger warnings isn't only hypocritical—It's absurd: University Times; 2018 [Available from: http://www.universitytimes.ie/2018/12/castigating-trigger-warnings-isnt-only-hypocritical-its-absurd/.
26. Harper C. It's official—trigger warnings might actually be harmful 2018 [Available from: https://medium.com/@CraigHarper19/its-official-trigger-warnings-might-actually-be-harmful-3e8acaae098b.
27. Bass S, Clark M. The gravest threat to colleges comes from within. Chronicle of Higher Education. 2015; 62:A26–A7.

28. Karasek S. Trust me, trigger warnings are helpful: New York Times; 2016 [Available from: https://www.nytimes.com/roomfordebate/2016/09/13/do-trigger-warnings-work/trust-me-trigger-warnings-are-helpful.

29. Cares A, Franklin C, Fisher B, Bostaph L. "They Were There for People Who Needed Them": Student Attitudes Toward the Use of Trigger Warnings in Victimology Classrooms. Journal of Criminal Justice Education. 2019; 30:22–45.

30. Llewellyn-Beardsley J, Rennick-Egglestone S, Callard F, Crawford P, Farkas M, Hui A, et al. Characteristics of mental health recovery narratives: systematic review and narrative synthesis. PLoS One. 2019; 14:e0214678. https://doi.org/10.1371/journal.pone.0214678 PMID: 30921432

31. Rennick-Egglestone S, Ramsay A, McGranahan R, Llewellyn-Beardsley J, Hui A, Pollock K, et al. The impact of mental health recovery narratives on recipients experiencing mental health problems: qualitative analysis and change model. PLoS One. 2019; 14:e0226201. https://doi.org/10.1371/journal.pone.0226201 PMID: 31834902

32. Slade M, Rennick-Egglestone S, Llewellyn-Beardsley J, Yeo C, Roe J, Bailey S, et al. Using recorded mental health recovery narratives as a resource for others: Narrative Experiences Online (NEON) intervention development. JMIR Formative Research. in press.

33. Rennick-Egglestone S, Elliott R, Smuk M, Robinson C, Bailey S, Smith R, et al. Impact of receiving recorded mental health recovery narratives on quality of life in people experiencing psychosis, people experiencing other mental health problems and for informal carers: Narrative Experiences Online (NEON) study protocol for three randomised controlled trials. Trials. 2020; 21(1):661. https://doi.org/10.1186/s13063-020-04428-6 PMID: 32690105

34. Wogalter M, Conzola V, Smith-Jackson T. Research-based guidelines for warning design and evaluation. Appl Ergon. 2002; 33:219–30. https://doi.org/10.1016/s0003-6870(02)00009-1 PMID: 12164506

35. Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, et al. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. International journal of nursing studies. 2012; 49(1):47–53. https://doi.org/10.1016/j.ijnurstu.2011.07.002 PMID: 21835406

36. Ross A. Finding Political Identities. Young People in a Changing Europe. London: Palgrave MacMillan; 2018.

37. Charles A, Korde P, Newby C, Grayzman A, Hiltensperger R, Mahlke C, et al. Proportionate translation of study materials and measures in a multinational global health trial: methodology development and implementation. BMJ Open. 2022; 12:e058083. https://doi.org/10.1136/bmjopen-2021-058083 PMID: 35058270

38. Drovandi A, Teague P-A, Glass B, Malau-Aduli B. A systematic review of the perceptions of adolescents on graphic health warnings and plain packaging of cigarettes. Systematic Reviews. 2019; 8:25. https://doi.org/10.1186/s13643-018-0933-0 PMID: 30654833

39. Hassan L, Shiu E. A systematic review of the efficacy of alcohol warning labels: Insights from qualitative and quantitative research in the new millennium. Journal of Social Marketing. 2018; 8:333–52.

40. Marchant A, Hawton K, Burns L, Stewart A, John A. Impact of Web-Based Sharing and Viewing of Self-Harm–Related Videos and Photographs on Young People: Systematic Review. Journal of medical Internet research. 2021; 23:e18048. https://doi.org/10.2196/18048 PMID: 33739289

41. Slade M, Blackie L, Longden E. Personal growth in psychosis. World Psychiatry. 2019; 18:29–30. https://doi.org/10.1002/wps.20585 PMID: 30600621

42. Slade M, Rennick-Egglestone S, Blackie L, Llewellyn-Beardsley J, Franklin D, Hui A, et al. Post-traumatic growth in mental health recovery: qualitative study of narratives. BMJ Open. 2019; 9:e029342. https://doi.org/10.1136/bmjopen-2019-029342 PMID: 31256037

43. Stratta M, Park J, deNicola C. Automated Content Warnings for Sensitive Posts. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems 2020. p. 108.

44. Kirkinis K, Pieterse A, Martin C, Agiliga A, Brownell A. Racism, racial discrimination, and trauma: A systematic review of the social science literature. Ethn Health. 2021; 26:392–412. https://doi.org/10.1080/13557858.2018.1514453 PMID: 30165756