

What Counts as Harmful Eating Disorder Content? A Delphi Study Developing Consensus-Based Harm Identification Rules with Lived Experience and Professional Experts

Pranita Shrestha

pranita.shrestha@monash.edu

Monash University

Jue Xie

Monash University

Pari Delir Haghighi

Monash University

Scott Griffiths

The University of Melbourne

Hannah K Jarman

Deakin University

Ana Ximena Torres

Monash Health

Zoe Bradbury

Butterfly Foundation

Roisin McNaney

The University of Melbourne

Research Article

Keywords: eating disorders, social media, content moderation, Delphi method, harm identification

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10587785/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

What Counts as Harmful Eating Disorder
Content? A Delphi Study Developing
Consensus-Based Harm Identification Rules with
Lived Experience and Professional Experts

Pranita Shrestha^{1*}, Jue Xie^{1*†}, Pari Delir Haghighi^{1†},
Scott Griffiths^{2†}, Hannah K Jarman^{4,5†}, Ana Ximena Torres^{7†},
Zoe Bradbury^{6†}, Roisin McNaney^{1,3†}

^{1*}Human-Centred Computing, Monash University, Clayton, Melbourne,
VIC, Australia.

²Melbourne School of Psychological Sciences, University of Melbourne,
Parkville, Melbourne, VIC, Australia.

³School of Computing and Information Systems, University of
Melbourne, Carlton, Melbourne, VIC, Australia.

⁴School of Psychology, Deakin University, Geelong, Melbourne, VIC,
Australia.

⁵Deakin Lifespan Institute, Deakin University, Burwood, Melbourne,
VIC, Australia.

⁶Butterfly Foundation, Crows Nest, Sydney, NSW, Australia.

⁷Wellness and Recovery Centre (WRC), Monash Health, Clayton,
Melbourne, VIC, Australia.

*Corresponding author(s). E-mail(s): pranita.shrestha@monash.edu;
jue.xie@monash.edu;

Contributing authors: pari.delir.haghighi@monash.edu;
scott.griffiths@unimelb.edu.au; h.jarman@deakin.edu.au;
ana.torres@monashhealth.org; Zoe.Bradbury@butterfly.org.au;
roisin.mcnaney@unimelb.edu.au;

[†]These authors contributed equally to this work.

Abbreviations

ED - Eating disorders; PS - Pranita Shrestha; JX - Jue Xie; PDH - Pari Delir Haghighi; SG - Scott Griffiths; HJ - Hannah K Jarman; ZB - Zoe Bradbury; AXT - Ana Ximena Torres; RM - Roisin McNaney;

Declarations

Ethics approval and consent to participate

This study was approved by the Monash University Human Research Ethics Committee under approval number ID 41131. All Steering Committee members and Delphi panel members provided informed consent prior to publication.

Consent for publication

Participants provided informed consent for publication of research data and findings, with the understanding that no identifying information would be disclosed in any publications.

Availability of data and materials

Detailed results from the preliminary harm identification rules, all Steering Committee and Delphi rounds, and the final harm identification rules are provided in the Supporting Information.

Competing interests

No conflicts of interest to be declared.

Funding declaration

No funding to be declared.

Authors' Contributions

Pranita Shrestha: conceptualisation, methodology, project management, data collection, data analysis, harm identification rules review in each stage, writing - original draft, writing - reviewing and editing. **Joe Xie:** conceptualisation, harm identification rules review in each stage, supervision, writing - reviewing and editing, **Pari Delir Haghighi:** conceptualisation, harm identification rules review in each stage, supervision, writing - reviewing and editing. **Scott Griffiths:** conceptualisation, Steering Committee member, writing - reviewing and editing. **Hannah Jarman:** conceptualisation, Steering Committee member, writing - reviewing and editing. **Ana Ximena Torres:** Steering Committee member, writing - reviewing and editing. **Zoe Bradbury:** Steering Committee member, writing - reviewing and editing. **Roisin McNaney:** conceptualisation, data analysis, harm identification rules review in each stage, supervision, writing - reviewing and editing.

Acknowledgements

The authors utilised the generative artificial intelligence platform ChatGPT (GPT 5.3, OpenAI) to assist with grammar and language refinement. The authors critically reviewed and revised all AI suggestions to ensure accuracy, appropriateness, and alignment with the intended meaning.

Plain English Summary

Social media has become deeply woven into the daily lives of people at risk of, or experiencing, eating disorders, shaping how they see their bodies, connect with others,

and navigate recovery. Given this influence on risk, recovery and relapse, this study aimed to create expert-informed rules for identifying eating disorder-related content that could be harmful. The rules were built through a Delphi study involving 21 experts by profession, including clinicians, researchers and professionals working in the body image and eating disorder space, and 20 individuals with a past history of eating disorders. Using a structured process involving four Steering Committee rounds and three Delphi panel rounds (two primary and one refinement), 45 harm identification rules were created. Rules related to clearly harmful behaviour, such as promoting thinness or unsafe weight loss, reached consensus quickly. Rules about more ambiguous content, such as posts about nutrition, exercise, or recovery, took longer to reach consensus because their harm depends on context, the creator's intent, and the viewer's vulnerability. Overall, the findings show that harm depends not just on what is said in the content but also on the surrounding context, offering a foundation for better automated detection and more personalised, context-sensitive moderation.

Abstract

Objective: We aimed to develop expert-informed harm identification rules to identify eating disorder (ED)-related harmful social media content for individuals at risk of or experiencing ED. Given the increasing influence of social media on ED risk, recovery and relapse, there is a need for systematic approaches to identify harmful content. To our knowledge, this is the first study to systematically develop consensus-based harm identification rules integrating both professional expertise and lived experience perspectives.

Method: We employed a modified Delphi methodology, involving n=21 experts by profession (researchers, clinicians and professionals in the ED and body image field) and n=20 experts by lived experience (individuals with a history of ED). The

process included four Steering Committee rounds alternating with three Delphi panel rounds (two primary and one refinement) to iteratively refine the rules. We assessed agreement/consensus using predefined consensus thresholds, and analysed qualitative feedback from open-ended responses descriptively to inform refinement.

Results: We generated a total of N=67 preliminary rules, with n=45 retained. Rules capturing explicit harmful behaviours (e.g., thin-ideal promotion and unsafe weight-loss practices) achieved higher agreement, while context-dependent content (e.g., nutrition, exercise and recovery narratives) required iterative refinement. Qualitative feedback highlighted the importance of context, intent and user vulnerability in determining harm.

Conclusion: Harm in ED-related content depends on both content and context. Explicitly harmful content represents clear harm, while other forms are more nuanced and shaped by user interpretation. The resulting rules provide criteria for identifying both explicit and nuanced harm, which could support automated detection, personalised content moderation, and policy to protect vulnerable users while preserving beneficial content.

Keywords: eating disorders, social media, content moderation, Delphi method, harm identification

1 Background

Social media platforms rely heavily on algorithmic systems that organise, curate and recommend content to users based on how they interact with social media content [Rader and Gray \(2015\)](#). In the context of eating disorder (ED)-related content, these systems have been shown to amplify content associated with dieting behaviours and extreme thinness [Center for Countering Digital Hate \(2024\)](#); [Griffiths et al. \(2024\)](#), driven

7 by both active interactions (e.g., liking, commenting, sharing) and passive behaviours
8 (e.g., prolonged viewing time, delayed skipping) [Narayanan \(2023\)](#). [Griffiths et al.](#)
9 [\(2024\)](#) found that TikTok’s recommendation algorithm disproportionately exposed
10 individuals experiencing ED to harmful content, including 4343% more ED-related
11 content, 335% more dieting-related content and 142% more exercise-related content,
12 compared with individuals not experiencing ED. This amplification can lead to the
13 development of ED ‘echo chambers’, defined as spaces in which harmful ED-related
14 content is repeatedly reinforced and normalised through an algorithmic feedback loop
15 that extends beyond users’ active searches and active interactions [Griffiths et al.](#)
16 [\(2024\)](#); [Shrestha et al. \(2026\)](#). This can ultimately reinforce maladaptive beliefs and
17 behaviours.

18 In response, social media platforms have implemented moderation policies [Meta](#)
19 [\(2026\)](#); [TikTok \(2025\)](#) that combine automated systems with human review [Facebook](#)
20 [\(2024\)](#); [Instagram \(2024\)](#); [TikTok \(2024\)](#) and user-level self-moderation (e.g., unfol-
21 lowing or muting content) [Hockin-Boyers et al. \(2021\)](#). However, harmful content is
22 often highly contextual and ambiguous [Shrestha et al. \(2026\)](#). For example, “what I
23 eat in a day” content [Drivas et al. \(2024\)](#), where creators document and share the
24 food they consume over the course of a day, often falls within an ambiguous category
25 [Davis et al. \(2025\)](#). At face value, such content may appear neutral or informational.
26 However, it can easily become harmful when it promotes restrictive eating patterns,
27 showcases detailed low-calorie counts and portion sizes or implicitly encourages com-
28 parison [Davis et al. \(2025\)](#). This can be particularly harmful for individuals at risk of
29 or experiencing ED. Platform moderation policies rely on broad content-related guide-
30 lines to detect harmful or extreme content (e.g., extremely low-calorie diets; starving,
31 bingeing and purging) or age-based restrictions (e.g., not showing content related to
32 weight loss medications to those under 18) [TikTok \(2025\)](#). Yet, vulnerable popula-
33 tions, such as individuals at risk of or experiencing ED, are not adequately protected

34 by broad and generalised content moderation mechanisms alone. Platforms often re-
35 move explicitly pro-ED content (e.g., pro-anorexia/pro-ana or pro-bulimia/pro-mia
36 [Borzekowski et al. \(2010\)](#), thinspiration [Ghaznavi and Taylor \(2015\)](#)), while ambigu-
37 ous or ‘grey area’ content is frequently unaddressed or undetected. Prior research has
38 demonstrated that both ambiguous and overtly harmful content may bypass existing
39 safeguards of current moderation systems [ABC News \(2023\)](#); [Center for Countering](#)
40 [Digital Hate \(2024\)](#); [Griffiths et al. \(2024\)](#).

41 Together, these limitations point to a critical gap in current moderation ap-
42 proaches. A key challenge in addressing these limitations is the absence of clear and
43 shared definitions and interpretations of what constitutes harmful social media con-
44 tent for individuals at risk of or experiencing ED [Shrestha et al. \(2026\)](#). This gap
45 underscores the need for structured harm identification rules that reflect the nuanced,
46 context-dependent ways harm manifests across the visual, textual, and audio compo-
47 nents of social media content. To achieve this, we conducted a modified Delphi study
48 [Nasa et al. \(2021\)](#) to develop a consensus on expert-informed harm identification rules
49 for ED-related social media content, integrating expertise from experts by profession
50 (researchers, clinicians, and professionals from the ED and body image field) and ex-
51 perts by lived experience (individuals with a past history of ED). To the best of our
52 knowledge, this is the first Delphi study to establish such a consensus in this domain.

53 2 Method

54 We employed a modified Delphi methodology [Nasa et al. \(2021\)](#), using an iterative pro-
55 cess to achieve consensus on a set of predefined preliminary harm-identification rules
56 for harmful ED-related content. Unlike a traditional Delphi approach, which begins
57 with expert-generated open-ended items, the modified Delphi starts with predefined
58 items that participants rate and refine across rounds. The Delphi methodology was
59 well-suited to this study, given its capacity to support consensus-building on complex,

sensitive, and context-dependent constructs. It offered anonymity, enabled iterative expert feedback, and minimised the influence of dominant voices. It has been widely applied in health-related research [Chalmers et al. \(2014\)](#); [Ross et al. \(2012\)](#); [Young et al. \(2024\)](#) and the ED field [Austin et al. \(2023\)](#); [Bannatyne et al. \(2018\)](#); [Hanegraaf et al. \(2024\)](#); [Hart and Wade \(2020\)](#); [Lemmer et al. \(2024\)](#); [McMaster et al. \(2020\)](#) to establish definitions, guidelines, and best-practice recommendations.

We recruited n=5 members for the Steering Committee (RM, SG, HJ, ZB and AXT), who are also co-authors of this study, to guide the development and refinement of harm identification rules across different rounds. RS, SG, and HJ represented researchers working in the body image and ED fields; AXT represented a clinician; and ZB represented an individual with lived experience of ED and a professional from an ED support service (not-for-profit sector). For the Delphi panel, we recruited n=21 experts by profession and n=20 experts by lived experience. In total, we conducted n=4 Steering Committee survey rounds, n=2 primary Delphi rounds and n=1 Delphi refinement round. Figure 1 presents an overview of the study.

2.1 Ethics

We obtained ethical approval from the Monash University Human Research Ethics Committee (ID: 41131). Prior to participation, we provided an explanatory statement and consent form to all participants. Experts by lived experience were required to be over 18 years of age and to have a past history of ED, but no current diagnosis, to ensure the safety and well-being of the participants with lived experience. Additionally, we refined the recruitment materials and survey language based on feedback from professionals at Butterfly Foundation, Australia’s leading not-for-profit ED support service provider.

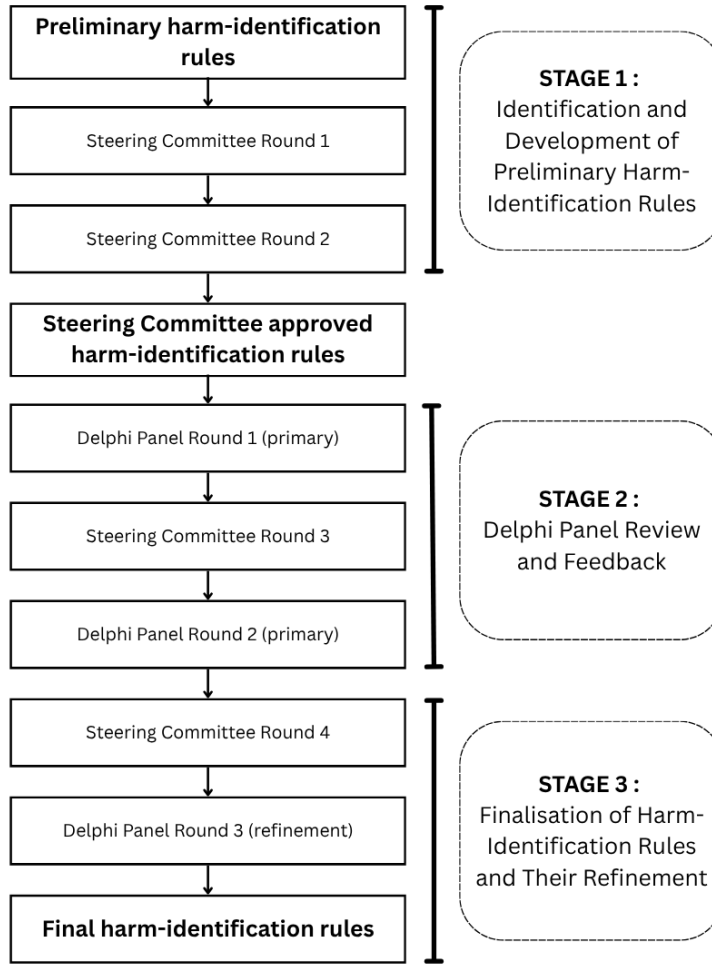


Figure 1 Overview of the study

2.2 Recruitment

We recruited experts by lived experience through the Butterfly Collective (The Butterfly Foundation's Lived Experience Network). Interested participants completed a Google Form to provide demographic and contact information, after which the researchers contacted them and invited them to complete the consent form. We reimbursed the participants from this group with AUD 80 gift cards for contributing as Delphi panel members. For the experts by profession, we purposively recruited

participants with intersecting expertise in body image, ED, and social media via our professional networks, supplemented by snowball sampling. We did not provide any reimbursement to participants in this group. During recruitment, we also asked experts by profession whether they had lived experience of ED. Regardless of their response, we classified them as experts by profession. We then provided all participants with the required information package, which included a plain-language summary and instructions for each round of the Delphi panel survey.

2.3 Stages

The entire study was divided into three stages.

2.3.1 Stage 1 - Identification and Development of Preliminary

Harm Identification Rules

We developed preliminary harm identification rules through a three-step data triangulation process. First, we reviewed the literature on ED-related harm in social media content to identify content types, themes, and mechanisms of harm already documented in existing research. Second, we conducted an initial study involving n=12 interviews with experts by profession, which were exploratory in nature and used, alongside the literature review findings, to identify an initial set of harmful content characteristics and to group them into content categories. Third, these preliminary categories were presented to n=5 focus groups with n=18 experts by lived experience, who validated and refined them from a lived-experience perspective, and surfaced additional harmful content characteristics not captured by the literature or the interviews.

Across these three sources, we identified recurring harmful content characteristics such as promotion of extreme fasting or framing body fat as something to be monitored, and through iterative discussion within the research team, converted each into

116 a discrete rule describing the content and the harm it posed. The rule structure fol-
117 lowed a consistent internal format: a content descriptor (what the content depicts),
118 a harm mechanism (how the content causes harm), and, where relevant, a boundary
119 condition (the circumstances under which the harm does not apply). This extraction
120 process was iterative: we checked and refined harmful content characteristics iden-
121 tified in the literature against the interview data, and we tested interview-derived
122 characteristics against the focus group data. This process yielded a preliminary set of
123 $n=67$ harm identification rules across eight categories: General content, Drugs, supple-
124 ments and surgeries, Beauty products, Filters and edited images, Food and nutrition,
125 Exercise, Body measurements and comparisons and Eating disorder recovery. The in-
126 terview and focus guides, along with detailed qualitative findings underpinning this
127 extraction, are provided in [Shrestha et al. \(2026\)](#). We then conducted a preliminary
128 cross-disciplinary focus group with experts in Delphi methodology ($n=1$), Artificial
129 Intelligence ($n=2$), and body image and ED ($n=2$) to build a mutual understand-
130 ing, collaboratively shape the requirements, solidify the structure and protocol of the
131 Delphi process, and standardise how rules should be structured for the Delphi study.
132 Steering Committee members and Delphi panel members rated and commented on the
133 resulting rules but did not apply or restructure them. The research team restructured
134 rules in subsequent rounds in response to panel feedback.

135 To refine and validate the preliminary rules before the Delphi launch, Steering
136 Committee members completed two rounds of anonymous surveys via REDCap. Only
137 $n=4$ Steering Committee members participated in these two rounds, as RM was
138 involved in the development of the preliminary harm identification rules. Each sur-
139 vey presented rules alongside a five-point Likert scale for including the rules in the
140 Delphi rounds: “Strongly Agree,” “Agree,” “Disagree,” “Strongly Disagree,” and “Ab-
141 stain/Unable to Answer,” as well as a free-text comment box. Members were asked
142 to provide comments where they disagreed with the inclusion of the rule or wished to

143 suggest modifications, consolidation, or adjustments, and were permitted to propose
 144 new rules. Strong consensus was defined as at least 75% of the members selecting ei-
 145 ther “Agree” or “Strongly Agree,” (i.e. 3/4). This consensus threshold corresponded
 146 to the commonly used median threshold in Delphi studies [Diamond et al. \(2014\)](#) and
 147 was appropriate for a small panel size. Rules reaching consensus with substantive
 148 comments or feedback, or those not reaching consensus in Round 1 were re-presented
 149 in the Steering Committee Round 2 alongside original and modified versions and all
 150 anonymised comments. Rules with only minor grammatical adjustments were not re-
 151 presented, as these changes did not alter the meaning or substantive content of the
 152 rules.

153 A sample of the survey is presented in Figure 2.

1. General content	
1.1. The content depicts individuals with visual ribcages, collarbones, or other bones emphasised to suggest underweight bodies.	☐ Strongly Agree ☐ Agree ☐ Disagree ☐ Strongly Disagree ☐ Abstain/ Unable to answer
1.1. Comments: [Please provide comments if you disagree with adding this rule in the Delphi Panel or if you want to suggest certain wordings, merging or adjustments. Also, please provide comments if you choose abstain/ unable to answer.]	

Figure 2 Survey for Steering Committee

154 2.3.2 Stage 2 - Delphi Panel Review and Feedback

155 Panel members evaluated the harm identification rules approved by the Steering Com-
 156 mittee via a REDCap survey (in English), which the entire research team reviewed
 157 before dispatch. The survey was divided into eight categories and used the same five-
 158 point Likert scale and free-text comment box as the Steering Committee’s. Consensus
 159 was defined as at least 80% of panel members within each expert group (experts

160 by lived experience and experts by profession) selecting either “Agree” or “Strongly
161 Agree”; if either group failed to reach this threshold, consensus was not achieved. This
162 consensus threshold was set higher than that used for the Steering Committee rounds
163 to ensure a more stringent level of agreement across the larger panel size. Follow-
164 ing Round 1, Steering Committee Round 3 was initiated, in which members reviewed
165 consensus results and anonymised comments from the Delphi panels, and proposed
166 whether rules should be retained or modified, and how. RM took part in this round
167 as a Steering Committee member.

168 Round 2 retained n=16 experts by profession and n=14 experts by lived experience,
169 as the remaining panel members were non-respondents. Panels were presented with
170 rules that either had not reached consensus or had been modified, along with the
171 original version, subsequent modifications, consensus data, and anonymised comments.
172 Panels indicated agreement using the Likert scale, provided free-text comments and,
173 where modifications had been made, selected “Prefer the original” or “Prefer the
174 modification”. Additionally, the panels were presented with the full set of rules within
175 each category. (e.g. General Content; Drugs, Supplements and Surgeries), and were
176 asked to rank their top five rules across four dimensions: priority for moderation,
177 ease of implementation, difficulty of implementation and risk of false positives. These
178 ranking data are reported in not reported in this manuscript. A sample of the survey
179 used to collect the top five rule selections is presented in Figure 3.

180 **2.3.3 Stage 3 - Finalisation of harm identification Rules and Their** 181 **Refinement**

182 In Steering Committee Round 4 (the last round), members were asked to cast a bi-
183 nary vote to confirm approved rules and exclude rejected ones, with justifications.
184 To overturn an approved rule or reinstate a rejected one, members required unani-
185 mous agreement, grounded in the Delphi panel’s feedback. Minor grammatical changes

<p>From the "List of rules for Category 7. Body measurements and comparisons", please list the top 5 rules in order of priority, from highest to lowest, using their rule numbers (e.g., 7.8, 7.1, etc.) in the comment box below. These should be the most critical rules that should not be missed during content moderation.</p> <p>Please provide the reasoning for this particular ranking.</p>	
<p>From the "List of rules for Category 7. Body measurements and comparisons" please identify the top 5 rules that are easiest to implement, listed in order from easiest to less easy, using their rule numbers (e.g., 7.8, 7.1, etc.).</p> <p>Please provide the reasoning for this particular ranking.</p>	
<p>From the "List of Rules for Category 7. Body measurements and comparisons" please list the top 5 rules that are most difficult to implement, ranked from most to least difficult, using their rule numbers (e.g., 7.8, 7.1, etc.).</p> <p>Please provide the reasoning for this particular ranking.</p>	
<p>From the "List of Rules for Category 7. Body measurements and comparisons" please identify the top 5 rules most likely to lead to incorrect takedowns or content being wrongly flagged (false positives), ranked from most to least likely to cause false positives, using their rule numbers (e.g., 7.8, 7.1, etc.).</p> <p>Please provide the reasoning for this particular ranking.</p>	

Figure 3 Top five rule selections for Delphi Panel Round 2

186 were permitted as long as they did not change the meaning. Following this round, PS
 187 compiled the final list of harm identification rules. The remaining team subsequently
 188 reviewed the final validated rules. After substantially revising one rule based on two
 189 primary rounds of the Delphi study and four rounds of Steering Committee feedback,
 190 we conducted a targeted Delphi Panel Round 3 with the same panel (n=8 experts
 191 by profession, n=8 experts by lived experience, following further attrition) to refine
 192 consensus on this single item, using the same five-point scale and comment box. The
 193 Steering Committee members and research team approved the final list of rules.

2.4 Data Analysis

Quantitative data were analysed descriptively to determine agreement levels across rounds. Consensus was predefined as $\geq 75\%$ for the Steering Committee and $\geq 80\%$ for each expert group in the Delphi panel. Rules that did not meet consensus were modified, consolidated, or removed based on the feedback. Even rules that met the consensus threshold were modified or consolidated in response to substantive panel comments. High consensus was defined as $\geq 90\%$ of the Delphi panel in each Delphi expert group (experts by lived experience and experts by profession) selecting “Strongly Agree” or “Agree”) and moderate consensus as $\geq 80\%$ and $< 90\%$ of panel in each Delphi group selecting “Strongly Agree” or “Agree”).

3 Results

The Delphi study produced a final set of $n=45$ harm identification rules across six categories, following the removal of the beauty products, and filters and edited images categories. Table 1 presents the full set of final rules and their consensus levels. Of the $n=45$ rules, $n=31$ met the high-consensus threshold, and $n=14$ met the moderate-consensus threshold. High consensus was achieved predominantly for rules addressing explicit, clearly observable harm, including thin-ideal promotion, encouragement of disordered eating behaviours and unsafe weight loss practices. Rules where harm depended on framing or context were difficult to agree on and tended to fall within the moderate consensus tier. A total of $n=12$ rules were rejected despite modifications and were not fully consolidated into other rules. Of these, $n=9$ were rejected by the Delphi panels and $n=3$ by the Steering Committee. Detailed results of the rules that did not receive consensus are provided in Supplementary Material A.

Please note that some findings were considered beyond the scope of the present manuscript and are therefore not reported. For example, in Delphi Panel Round 2, panel members were asked to identify their top five priority rules based on importance

220 for moderation, ease of implementation, difficulty of implementation and risk of false
 221 positives.

Table 1: Final harm-identification rules and consensus levels

<i>Note.</i> Green shading indicates high consensus, defined as at least 90% agreement in each expert group. Yellow shading indicates moderate consensus, defined as at least 80% but less than 90% agreement in each expert group.	
Final Rule	Consensus Level
General Content	
Rule 1. The content depicts individuals with visual ribcages, collarbones or other bones emphasised to suggest the desirability or promotion of underweight bodies.	High
Rule 2. The content uses language such as “super skinny”, “model thin”, “bare bones” and “thigh gap goals”.	High
Rule 3. The content promotes or encourages participation in viral or social media “body size-related challenges” that normalises comparison [e.g. “A4 waist challenge”, “hip bone comparison”].	High
Rule 4. The content suggests a specific body shape to aspire to and promotes unrealistic or disordered methods to achieve this (e.g. thin waist, curvy bum, six-pack abs or thigh gaps).	Moderate
Rule 5. The content uses eating disorder-specific terminology, slang or the mutation of these terms (an0rExia, bulim1a) into language that promotes harm rather than support.	High
Rule 6. The content promotes very low-calorie diets (less than 1000 calories per day) or extreme fasting (24 hours or more than 24 hours of fasting) outside of religious, cultural and medical reasons.	High

Table 1 continued

Final Rule	Consensus Level
Rule 7. The content encourages weight loss-focused habits or suggests that it is acceptable to pursue extreme weight loss through disordered eating or disordered exercise behaviours that are extreme or unsafe.	High
Rule 8. The content implies weight loss as a virtue or goal to achieve happiness, self-worth, confidence or social success.	Moderate
Rule 9. The content frames eating disorders as insignificant or unimportant issues or issues that can be easily controlled.	High
Rule 10. The content perpetuates myths that eating disorders affect only certain demographics or body types.	High
Rule 11. The content frames binge-eating episodes as fun or entertaining without context on potential health risks.	Moderate
Rule 12. The content associates larger bodies with negative traits such as laziness, lack of control and unattractiveness.	High
Rule 13. The content implies thinness is superior or essential for worthiness often paired with derogatory references to body fat.	High
Drugs, Supplements and Surgeries	
Rule 14. The content names or references regulated weight-loss drugs, supplements or pills (e.g. Ozempic, Wegovy) with claims of effectiveness.	High

Table 1 continued

Final Rule	Consensus Level
Rule 15. The content contains product-selling or advertising companies promoting unregulated weight-loss drugs, supplements or pills with claims of effectiveness.	High
Rule 16. The content promotes easy and quick weight loss through drugs or supplements.	High
Rule 17. The content mentions performance-enhancing drugs, steroids or similar substances in connection with achieving specific physical or appearance goals effectively.	High
Rule 18. The content includes potential misleading, deceptive or unfair advertising of cosmetic surgery, including the use of testimonials, discounts without conditions, and unrealistic depictions of outcomes.	Moderate
Rule 19. The content implies that a person's natural appearance is inadequate and would be improved by modifying their appearance through surgery.	Moderate
Rule 20. The content promotes weight loss methods and drugs, performance enhancement drugs or cosmetic surgeries or treatments and suggests that these will lead to improved self-worth or social success.	Moderate

Table 1 continued

Final Rule	Consensus Level
Rule 21. The content features testimonials or endorsements promoting specific weight loss pills, fad diets, performance-enhancing drugs, or cosmetic surgery procedures intended primarily for aesthetic enhancement, without any mention of consultation with licensed healthcare professionals or medically verified needs.	High
Rule 22. The content implies that drugs or surgeries alone can create drastic physical changes and dismisses the importance of exercise, balanced nutrition, or proper medical supervision, which can encourage unsafe practices.	Moderate
Rule 23. The content encourages pursuing surgery and weight loss drugs or performance-enhancing drugs without medical advice or consideration of risks.	High
Rule 24. The content minimises, denies or dismisses serious health risks of body-modifying practices, including extreme diets, intense exercise regimens, the use of weight loss and/or performance-enhancing drugs and cosmetic surgery, potentially encouraging unsafe behaviours.	High
Rule 25. The content promotes easy or quick muscle gain through illegal anabolic steroids.	High

Table 1 continued

Final Rule	Consensus Level
Food and Nutrition	
Rule 26. The content suggests that nutritional advice is suitable for all viewers without personalisation or expertise. Broad, non-prescriptive suggestions (e.g., eat more vegetables) that do not imply medical authority are acceptable.	Moderate
Rule 27. The content includes the caloric content of specific foods or meals consumed.	Moderate
Rule 28. The content includes very low portion sizes, serving counts or specific amounts that would account for less than 1000 calories in total per day, and recommends this as a daily intake.	High
Rule 29. The content encourages restraint or self-control in eating as a pathway to health or beauty.	High
Rule 30. The content sells weight loss packages or nutritional programs without proper qualifications or professional approval.	Moderate
Exercise	
Rule 31. The content emphasises exercise solely for achieving appearance-related goals such as weight loss, muscle definition or a “perfect” body shape rather than promoting health benefits like improved strength, endurance or overall wellness.	High
Rule 32. The content focuses on exercise routines that are promoted as a way to fix or improve specific body parts to meet certain beauty standards.	Moderate

Table 1 continued

Final Rule	Consensus Level
Rule 33. The content suggests high-frequency, high-intensity or long-duration workouts as ideal or necessary for maintaining a certain physique for someone who is not an athlete.	High
Rule 34. The content depicts lengthy or highly intense workouts as typical or desirable, especially framed as essential for visible results.	Moderate
Rule 35. The content highlights the number of calories burned immediately after exercising.	High
Rule 36. The content promotes a disordered concept of exercise by framing it solely as a way to “balance out” or “work off” food consumption, reinforcing harmful relationships with food and physical activity.	High
Body Measurements and Comparisons	
Rule 37. The content includes a person visually standing on a weighing scale with their body weight clearly visible, with or without imagery of extreme thinness or muscularity, or language and hashtags promoting disordered eating behaviours.	High
Rule 38. The content promotes achieving or maintaining a specific body weight as a measure of personal value, social validation, or beauty, in a way that may reinforce disordered eating behaviours or contribute to the development of eating disorders.	High

Table 1 continued

Final Rule	Consensus Level
Rule 39. The content shows tape measurements around body parts or alternative measuring methods using household items as a means of body checking.	High
Rule 40. The content shows checking or assessing body parts for fat or leanness.	High
Rule 41. The content frames body fat and/or lack of muscularity as a visible concern that should be monitored or evaluated.	High
Rule 42. The content shows visual features side-by-side (as a process of showcasing before and after progress for weight loss) and sequential progress particularly depicting body transformations (excluding gender-affirming or transgender-related transitions or pregnancy, illness or injury).	Moderate
Rule 43. The content encourages comparison or evaluation of body size, shape, or specific physical features (e.g., waist size, bone visibility), focused on achieving specific body measurements or physical traits, excluding bodybuilding or athletic performance contexts.	High
Eating Disorder Recovery	
Rule 44. The content for eating disorder recovery showcases eating or weight loss behaviours that may trigger imitation or reinforce the severity of symptoms.	High

Table 1 continued

Final Rule	Consensus Level
Rule 45. The content for eating disorder recovery mentions specific numerical values tied to weight, calorie intake, portion sizes or exercise metrics.	Moderate

3.1 Demographics

The Steering Committee comprised n=5 members (4F, 1M); four identified as White/Caucasian, and one as Latina. Both the Steering Committee and Delphi panel members were asked to self-identify their ethnicity or cultural background, in line with [Soubutts et al. \(2024\)](#), and the research team did not standardise this. Three members identified as researchers in the body image and ED field, one as a clinician with policy and strategy expertise and one as a professional in ED support services with lived experience of ED.

The Delphi panel included n=20 experts by lived experience and n=21 experts by profession in Round 1, n=14 and n=16 in Round 2, and n=8 and n=8 in Round 3, respectively, an attrition rate of 27% from Round 1 to Round 2. This level of attrition is consistent with patterns reported in other Delphi studies, where final-round retention has ranged from 19.5% to 87.1% [Hall et al. \(2018\)](#). Dropout has been attributed to time constraints [Niederberger et al. \(2025\)](#); [Schmalz et al. \(2021\)](#), illness [Niederberger et al. \(2025\)](#) and the perceived burden of participation [Hall et al. \(2018\)](#); [Niederberger et al. \(2025\)](#); [Schmalz et al. \(2021\)](#).

Experts by profession were predominantly female and primarily identified as White/Caucasian (62%), mostly aged between 30 and 44 years. The majority identified

240 as researchers, clinicians, or academics, and nearly half reported 10-20 years of experi-
241 ence in the body image and ED fields. Over half reported having a lived experience of
242 ED. Experts by lived experience were also predominantly female and primarily identi-
243 fied as White/Caucasian (70%), mostly aged 25 to 34 years. The majority reported a
244 history of anorexia nervosa. Full demographics for experts by profession and experts
245 by lived experience are presented in Supplementary Material B and Supplementary
246 Material C respectively.

247 3.2 Stage 1 - Identification and Development of Preliminary 248 Harm Identification Rules

249 The n=67 preliminary harm identification rules are provided in Supplementary Ma-
250 terial D. Full consensus in these rounds was primarily achieved for rules that were
251 specific, clearly defined, and directly tied to observable ED-related behaviours (e.g.,
252 *‘promotes very low-calorie diets (500 calories per day)’*). Rules that did not reach
253 consensus were commonly described as *“too broad”* or *“not sufficiently disordered,”*
254 particularly where constructs such as *“ideal body shape”* lacked clarity. Members noted
255 that behaviours such as exercise, dieting, or food consumption are not inherently
256 harmful but depend on intent and framing (*“this doesn’t mean that all content about*
257 *weight loss should be eliminated”*), and raised concern about over-flagging normative
258 content. Refinements that added specific thresholds or behavioural indicators (e.g.,
259 *“addition of <1000 calories a day makes this rule specific enough to merit flagging”*)
260 reliably improved agreement in the subsequent round, while feasibility of implementa-
261 tion emerged as a recurring concern for rules describing broad or subjective constructs.
262 The flow and result of Stage 1 is presented in Figure 4.

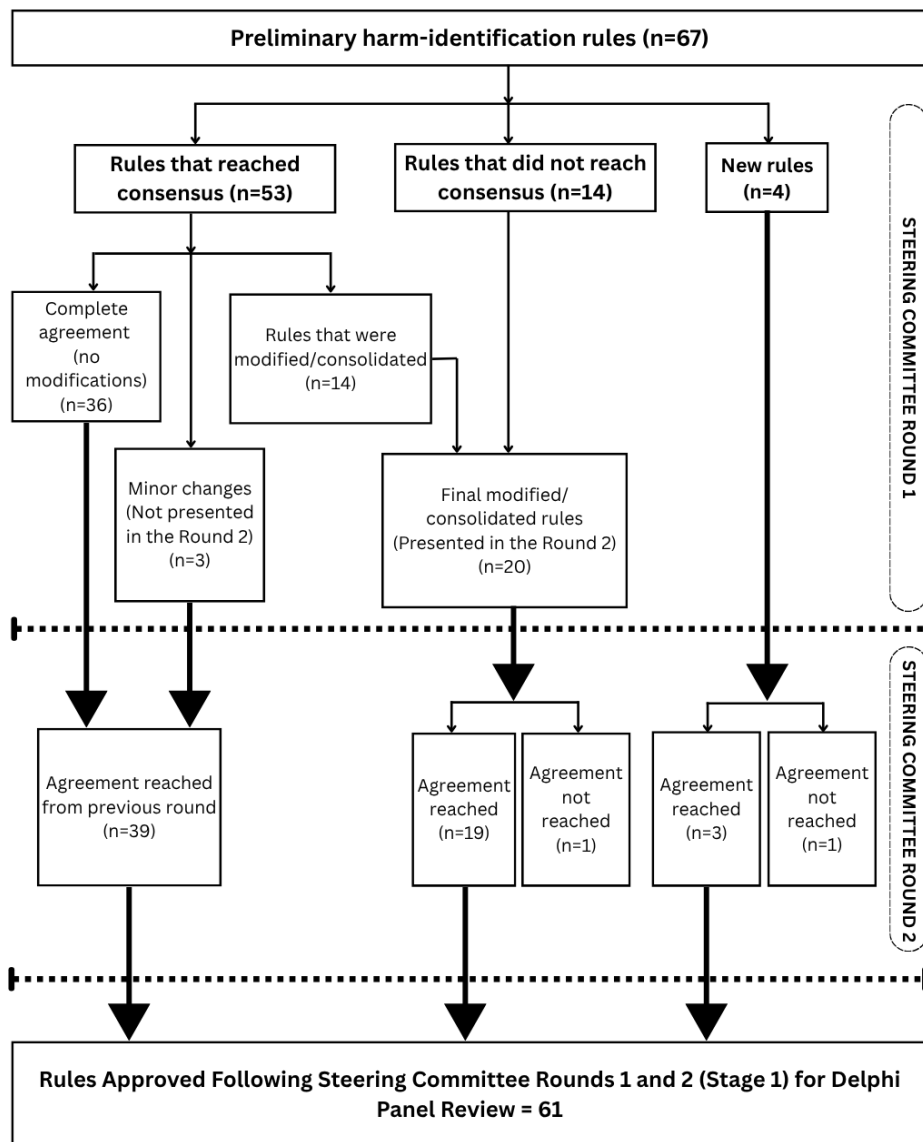


Figure 4 Result Overview of Stage 1 - Identification and Development of Preliminary Harm Identification Rules

3.2.1 Steering Committee Round 1

In Round 1, the Steering Committee members reviewed and rated n=67 preliminary harm identification rules and returned n=80 comments. Of these rules, n=53 reached

consensus. Of these, n=36 required no modifications; n=3 rules underwent minor editorial changes (e.g., grammatical revisions) and were not re-presented in Round 2, as these changes did not alter their intention or meaning, and the rules had already achieved sufficient consensus. The remaining rules were modified and/or consolidated based on feedback and re-presented in Round 2, including consolidation with related rules that had not reached consensus where conceptual overlap or duplication was identified. Overall n=24 rules (n=20 modified and/or consolidated rules and n=4 newly developed rules) were re-presented in Round 2. Full results of this round are provided in Supplementary Material E.

3.2.2 Steering Committee Round 2

In Round 2, the Steering Committee returned n=41 comments. Of the n=24 rules reviewed, n=22 received agreement for inclusion in the Delphi study and n=2 did not. Full results of this round are provided in Supplementary Material F.

3.2.3 Rules approved by the Steering Committee for inclusion in the Delphi study

This stage produced a final n=61 rules approved by the Steering Committee for inclusion in the Delphi study, distributed across the eight preliminary categories as follows: General content (n=17), Drugs, supplements and surgeries (n=12), Beauty products (n=2), Filters and edited images (n=3), Food and nutrition (n=8), Exercise (n=6), Body measurements and comparisons (n=8), and ED recovery (n=5). The final list of Steering Committee-approved harm identification rules to be presented to the Delphi Panel in Stage 2 is provided in Supplementary Material G.

3.3 Stage 2 - Delphi Panel Review and Feedback

Panels resonated with the Steering Committee's view that harm is highly context-dependent: visual indicators such as underweight bodies were seen as potentially

harmful when normalised as a beauty ideal, but not when depicting historical famine or malnourishment in an educational context. Behaviours such as dieting, exercise, or food consumption were similarly not viewed as universally disordered, given cultural, medical, religious, or everyday contexts (e.g., a relationship breakup) in which they may be appropriate. Ambiguity in terms such as “large quantities” or “casual” contributed to disagreement, and panels raised concern about unintended removal of beneficial or recovery-oriented content (e.g., *“banning or flagging collarbones could remove positive recovery content by those blogging their recovery safely”*). Revised wording that made the boundary between harmful and safe framing more explicit was associated with improved clarity and agreement in Round 2 (e.g., *“this is a good clear modification...straightforward to identify and remove this content”*). The flow and result of Stage 2 is presented in Figure 5.

3.3.1 Delphi Panel Round 1 (primary round)

In Round 1, the Delphi panel members reviewed and rated the n=61 rules to determine their inclusion in the final set of harm identification rules and returned N=292 comments (n=105 from experts by lived experience and n=187 from experts by profession). Out of n=61 rules, n=53 rules received agreement. Full results of this round are provided in Supplementary Material H.

3.3.2 Steering Committee Round 3

PS compiled the agreement levels and comments from each Delphi panel group for Steering Committee members. The Committee members provided revisions for improvement based on their expertise. PS analysed all feedback and implemented the necessary revisions, which was then reviewed by the research team and Steering Committee for final verification before commencing Delphi Panel Round 2. We received a total of n=183 comments from the Steering Committee. Overall, n=25 rules were included in Delphi Panel Round 2 for further review, which included n=20 rules that

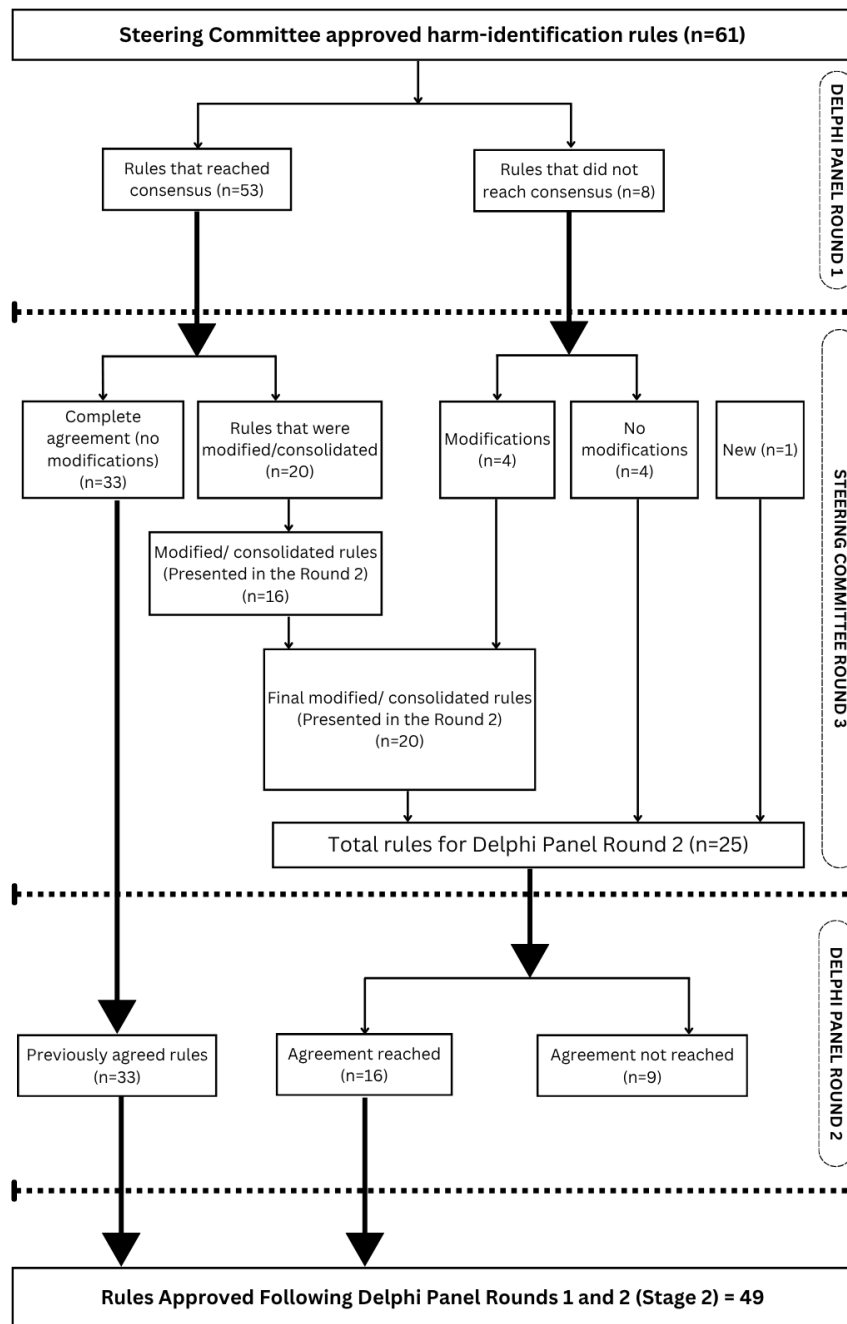


Figure 5 Result Overview of Stage 2 - Delphi Panel Review and Feedback

317 were modified and/or consolidated, n=4 rules with no modifications (that did not reach
318 agreement) and n=1 new rule developed based on the Delphi panel's feedback. The
319 detailed results of the Steering Committee Round 3 are provided in Supplementary
320 Material I.

321 **3.3.3 Delphi Panel Round 2 (primary round)**

322 In Round 2, the Delphi Panel members reviewed and rated the n=25 rules and returned
323 n=154 comments (n=69 from experts by lived experience and n=85 from experts
324 by profession). Of these, n=16 received agreement for inclusion in the final harm
325 identification rules, and n=9 rules did not. Full results of this round are provided in
326 Supplementary Material J.

327 **3.4 Stage 3 - Finalisation of harm identification Rules and** 328 **Their Refinement**

329 Steering Committee members' concerns in this round centred on distinguishing ex-
330 treme or explicit harm from common practices (e.g., intermittent fasting, HIIT
331 training), questioned the relevance of the 'Beauty products' category to ED specifically,
332 and noted mixed evidence for the benefit of 'Filters and edited images' disclaimers as
333 well as substantial detection challenges; these concerns informed the decision to re-
334 move both categories in their entirety. The flow and result of Stage 3 is presented in
335 Figure 6.

336 **3.4.1 Steering Committee Round 4**

337 Across Delphi Panel Rounds 1 and 2, there were a total of n=49 rules that were ap-
338 proved to be included in the final harm identification rules, and n=9 that were rejected.
339 PS compiled the agreement level for each Delphi panel group and then created a sur-
340 vey with all rules, along with their agreement level and agreement status (approved
341 or rejected). The Steering Committee members were required to select two options for

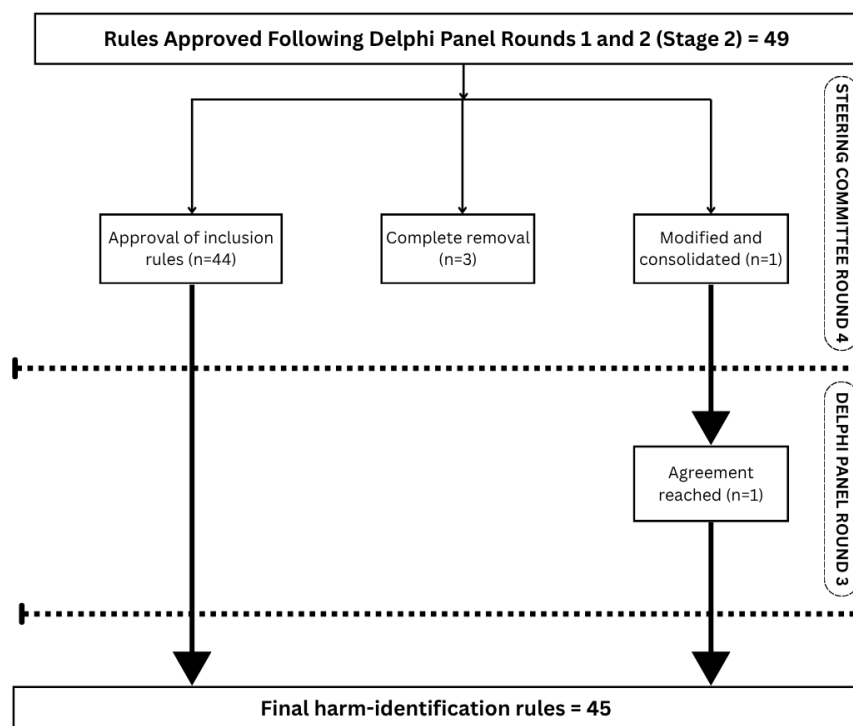


Figure 6 Result Overview of Stage 3 - Finalisation of harm identification Rules and Their Refinement

342 approved and rejected rules: “Agree with inclusion as approved/ rejected rule” and
 343 “Disagree with inclusion as approved/ rejected rule”. They were also asked to provide
 344 comments where necessary. Out of the n=49 approved rules, n=3 were unanimously
 345 agreed to be removed, and n=2 rules were consolidated and modified into a single
 346 rule that was subsequently sent to the Delphi panel for Round 3, leaving n=44 fully
 347 approved. Out of n=9 rejected rules, n=1 rule did not reach agreement for rejection.
 348 Full results of this round are provided in Supplementary Material K.

349 Following further discussion with the Steering Committee and in consideration of
 350 feedback from Delphi panels, consensus was reached to remove the categories ‘Beauty
 351 products’ and ‘Filters and edited images,’ along with their n=1 and n=2 associated
 352 rules, respectively. The category ‘Beauty products’ was removed because its inclusion
 353 extended beyond the core focus of ED-related harm and risk of expanding the harm

354 identification rules beyond their intended scope. The category ‘Filtered and edited
355 images’ was removed despite recognition of its potential harm. Committee members
356 noted mixed results on the impact of disclaimers or image alteration indicators, with
357 some studies suggesting limited or even unintended negative effects. Our Delphi partic-
358 ipants also flagged these concerns. Additionally, reliably identifying edited or filtered
359 content presents substantial practical challenges, raising concerns regarding consistent
360 detection and the risk of incorrect flagging.

361 Further discussion with the Steering Committee also led to the consolidation of two
362 rules from the ‘Food and nutrition’ category to form a single modified rule. ‘*The con-*
363 *tent suggests that nutritional advice is suitable for all viewers without personalisation*
364 *or expertise.*’ and ‘*The content suggests the amount consumed is ideal or beneficial for*
365 *health without mentioning professional credentials.*’ were merged to form ‘*The content*
366 *suggests that nutritional advice is suitable for all viewers without personalisation or*
367 *expertise. Broad, non-prescriptive suggestions (e.g., eat more vegetables) that do not*
368 *imply medical authority are acceptable.*’. Since this was a significant change to the rule
369 itself, we conducted Delphi Panel Round 3 to refine this single rule. As a result, the
370 Steering Committee approved a total of n=45 rules, which the Delphi panel members
371 agreed were the final harm identification rules.

372 **3.4.2 Delphi Panel Round 3 (refinement round)**

373 The single rule ‘*The content suggests that nutritional advice is suitable for all view-*
374 *ers without personalisation or expertise. Broad, non-prescriptive suggestions (e.g.,*
375 *eat more vegetables) that do not imply medical authority are acceptable.*’ from the
376 Steering Committee Round 4 achieved 87.5% agreement among both expert groups,
377 corresponding to 7/8 experts by lived experience and 7/8 experts by profession.

378 3.5 Cross-cutting qualitative themes

379 Across all rounds and both expert groups, three themes recurred in shaping how rules
380 were refined or excluded. *Context-dependence* was the most consistent theme: panels
381 repeatedly emphasised that dieting, exercise, and food-related content are not inher-
382 ently harmful, and that intent, framing, and audience determine whether the same
383 content is harmful, neutral, or supportive. *Definitional ambiguity* in terms such as
384 “large quantities,” “casual,” or “appearance-related” was the most common reason
385 rules failed to reach consensus, and was most often resolved by adding concrete thresh-
386 olds or explicit boundary conditions. *Tension between harm reduction and beneficial*
387 *content* arose particularly for recovery-oriented and body-neutral content, where pan-
388 els were concerned that overly broad rules risked removing supportive or educational
389 material alongside harmful content. Finally, *implementation feasibility* was raised as
390 a distinct concern from conceptual validity: some rules were considered conceptually
391 sound but “too tricky/general to enforce,” particularly where the underlying behaviour
392 (e.g., fitness content) was highly prevalent and difficult to distinguish from harmful
393 variants at scale.

394 4 Discussion

395 The current study utilised a modified Delphi methodology to develop consensus on
396 harm identification rules for identifying social media content that may be harmful
397 to individuals at risk of or experiencing ED. This process involved both experts by
398 profession (researchers, clinicians and professionals working in the body image and
399 ED field) and experts by lived experience (individuals with a past history of ED). The
400 resulting 45 harm identification rules represent one of the first systematically derived,
401 expert-informed sets of rules for identifying harmful ED-related content, integrating
402 both professional and lived experience perspectives.

403 The discussion is structured around three key themes. First, we examine the con-
404 ceptualisation of harm in ED-related social media content, highlighting its complexity
405 and context-dependent nature. Second, we consider how the harm identification rules
406 developed in this study provide a structured framework for integrating both ex-
407 plicit and context-dependent indicators of harm into social media systems. Finally,
408 we discuss the implications of these rules for content moderation and personalised
409 safeguards.

410 **4.1 Conceptualising harm in ED-related social media content**

411 Defining harm in ED-related social media content emerged as a complex, multi-
412 faceted and context-dependent challenge. Participants consistently emphasised that
413 behaviours such as dieting, exercise and food consumption are not inherently harmful,
414 but may become problematic depending on their context, intent, framing and audi-
415 ence. This resonates with existing literature, which suggests that the harm of social
416 media content does not always come from the content itself but also from how the users
417 interpret, engage and experience it [Fardouly et al. \(2015\)](#); [Holland and Tiggemann](#)
418 [\(2016\)](#). Findings from the Delphi rounds emphasised that certain types of ED-related
419 content may simultaneously offer educational or supportive benefits while also posing
420 risks of triggering or normalising disordered behaviours, depending on how they are
421 framed and consumed. For example, “what I eat in a day” content was viewed as po-
422 tentially harmful when framed around weight loss or food restriction, but supportive
423 in recovery contexts where it normalised adequate intake. Similarly, fasting-related
424 content was interpreted as harmful when associated with weight loss, but acceptable
425 within cultural or religious practices. These examples illustrate that the impact of ED-
426 related content is not fixed, but varies according to the context in which it is produced,
427 interpreted and consumed.

428 These findings highlight the need for harm identification frameworks that move
429 beyond static content classification, as harm in ED-related content cannot be under-
430 stood or addressed through content characteristics alone. Instead, it must account for
431 varying levels of risk, contextual factors and user vulnerability. The 45 harm identifi-
432 cation rules developed in this study provide a structured set of criteria, informed by
433 both professional expertise and lived experience, that incorporates explicit indicators
434 of harm as well as more context-sensitive criteria. By translating complex and often
435 implicit forms of harm into structured, identifiable indicators, these rules provide a
436 practical bridge between conceptual understanding of harm and its implementation
437 within digital systems, supporting a more nuanced approach to identifying harm in
438 ED-related content.

439 4.2 Integrating harm identification rules into social media 440 systems

441 Implementing harm identification rules on social media would require multimodal con-
442 tent analysis, which platforms already use in their existing moderation systems [Foster](#)
443 [\(2025\)](#); [Meta \(2026\)](#); [TikTok \(2025\)](#). Multimodal analysis [Alam et al. \(2022\)](#); [Arya](#)
444 [et al. \(2024\)](#) refers to the process of analysing content by integrating visual (e.g., im-
445 ages and videos), textual (e.g., captions, keywords and hashtags) and/or audio signals
446 to detect patterns associated with harmful content [Gorwa et al. \(2020\)](#). Currently,
447 platforms typically classify content into different risk categories based on an existing
448 “blacklist” of examples (e.g., depiction of thigh gaps, protruding spine combined with
449 ‘eating disorders’ term [Meta \(2026\)](#)), which then informs moderation actions, such
450 as complete removal, down-ranking or reducing visibility [Gorwa et al. \(2020\)](#). Con-
451 sistent with the findings of this study, which showed that explicit and clearly defined
452 forms of harm were more likely to reach consensus and to do so more rapidly, such
453 approaches are well suited to identifying these forms of harm. However, they fail to

capture more context-dependent content that lacks explicit indicators or falls outside previously identified forms of harm in the “blacklist” examples.

Building on the existing infrastructure for content scanning and extreme-harm signal extraction, the harm identification rules could extend these moderation systems by enabling platforms to capture and encode more nuanced, context-dependent indicators of harm at the point of upload. Such information could be stored and subsequently used to support more adaptive, audience-specific moderation strategies, allowing platforms to respond to patterns in content exposure and user vulnerability over time. This creates the potential to move beyond static, one-time content classification toward more dynamic, context-aware moderation systems. However, this approach would require robust safeguards, transparency and user consent, particularly given the sensitive nature of ED-related content. While conceptually aligned with current capabilities, its practical implementation would require careful design, governance and platform-level commitment.

The rules represent a key step towards addressing a longstanding gap by translating complex, often context- and intent-based harm into structured, interpretable indicators that can be incorporated into AI-driven moderation pipelines within the platform’s systems. They offer a clear and practical blueprint for platforms seeking to improve the detection and moderation of harmful content. While challenges remain in accurately interpreting context and intent, particularly for nuanced or recovery-related content, the greater challenge lies in adopting and implementing expert-informed approaches on social media platforms.

4.3 Content moderation, personalised safeguards and policy implications

Categories demonstrating high consensus, including explicit thin-ideal promotion, disordered eating encouragement, unsafe drug promotion, and harmful body measurements and comparisons, may warrant prioritised moderation or restriction. These findings are particularly relevant for individuals at risk of or experiencing ED, who may be susceptible to harmful messaging and algorithmic amplification of such content [Griffiths et al. \(2024\)](#).

In contrast, areas characterised by contextual ambiguity, including food- and nutrition-related content and ED recovery narratives, may require more nuanced, context-sensitive approaches rather than blanket enforcement. Our findings echoed existing work highlighting how overly restrictive measures risk removing educational, beneficial, or recovery-oriented content [Feuston et al. \(2020\)](#); [Zhang et al. \(2024\)](#). Such content is also less likely to be captured by automated systems or escalated for human review, despite its potential to cause harm among vulnerable users. This highlights the need for more advanced moderation approaches that can account for context, user vulnerability, and behavioural patterns [Gorwa et al. \(2020\)](#); [Saleous et al. \(2025\)](#).

With appropriate user consent and safeguards, personalised approaches could involve user-driven content controls, such as opting into “ED recovery” or “sensitive content” modes, muting specific categories or keywords or reducing algorithmic recommendations of potentially triggering material. This is known as personal content moderation [Jhaver et al. \(2023\)](#), which is based on users’ preferences. However, relying only on user-driven controls may be problematic in the context of ED, where individuals may experience conflicting motivations towards recovery [Zeeck et al. \(2025\)](#) or actively seek out harmful content [Ging and Garvey \(2018\)](#). In such contexts, placing the primary responsibility on the users to manage exposure to harmful content may be insufficient. The harm identification rules developed may support such approaches by

503 providing a structured framework for classifying harmful content as high- or low-risk,
504 enabling more accurate filtering beyond keyword- or hashtag-based controls.

505 From a policy perspective, there is a need to move beyond content-level mod-
506 eration towards system-level approaches that address both the identification and
507 amplification of harmful content. Integrating lived experience expertise alongside
508 professional knowledge supports the development of more inclusive and responsive
509 regulatory frameworks. In particular, the harm identification rules may inform the
510 design of evidence-based digital safety standards and platform governance strategies
511 that prioritise protecting vulnerable populations while maintaining access to benefi-
512 cial, recovery-oriented content. Such approaches should find a proper balance between
513 the need for harm reduction, and user control, expression, recovery advocacy and
514 transparency in content moderation and social media algorithms.

515 **5 Limitations**

516 This Delphi study has several limitations. First, although efforts were made to recruit
517 participants from different cultural and gender backgrounds, the majority of partici-
518 pants were female and White/ Caucasian, which resonates with broader recruitment
519 issues in the eating disorders field. While a balance between experts by profession and
520 lived experience was maintained, the most commonly represented diagnosis among
521 lived experience participants was anorexia nervosa, which may have shaped perspec-
522 tives on certain content categories. Second, participant attrition occurred across the
523 three rounds, a common occurrence in Delphi studies, which may have affected the
524 stability of consensus outcomes. Third, although the harm identification rules may
525 inform moderation practices, further research is needed to examine their practical
526 implementation in social media.

527 Despite these limitations, this study provides structured, expert-informed harm
528 identification rules that establish a foundational framework for identifying and cat-
529 egorising harmful content related to eating disorders for individuals at risk of or
530 experiencing eating disorders. Further empirical testing, cross-cultural validation and
531 examination of implementation feasibility are required to strengthen these findings.

532 **6 Conclusion**

533 This Delphi study developed expert-informed harm identification rules to guide the
534 identification of harmful ED-related content for individuals at risk of or experiencing
535 EDs. To our knowledge, this is the first study to systematically develop and reach
536 consensus on harm identification rules for ED-related content by integrating both
537 professional expertise and lived experience perspectives. The study engaged a Steer-
538 ing Committee across four rounds and Delphi panel members, including experts by
539 profession and experts by lived experience, across three rounds (two primary rounds
540 and one refinement round). Substantial consensus was achieved between the two ex-
541 pert groups, particularly in relation to explicit thin-ideal promotion, encouragement of
542 disordered eating behaviours, promotion of unregulated drugs and supplements, and
543 harmful body measurement and comparison practices.

544 At the same time, agreement varied more for food-related and ED recovery con-
545 tent, highlighting the contextual and nuanced nature of harm. The divergence observed
546 in these categories highlights the importance of context-sensitive moderation ap-
547 proaches that account for differences in intent, framing, and audience, rather than
548 blanket-ban strategies. The inclusion of both lived experience and professional per-
549 spectives strengthens the credibility and relevance of these rules and demonstrates the
550 complementary value of professional, clinical, research and experiential expertise in
551 identifying and defining harm.

Overall, this study provides a systematically derived and expert-informed set of rules that advances current approaches to identifying harmful ED-related content. By translating complex and often implicit forms of harm into structured and identifiable indicators, these rules offer a practical foundation for informing content moderation (particularly that maintained by Artificial Intelligence), digital safety strategies, and policy development. The findings emphasise that harm cannot be reduced to content characteristics alone, but must be understood in relation to context, user vulnerability, and patterns of exposure.

References

- ABC News (2023). *Eating-disorder patients confront meta bosses at federal parliament as research highlights tiktok's negative influence*. Retrieved from <https://www.abc.net.au/news/2023-09-14/eating-disorder-patients-to-confront-meta-bosses-at-parliament/102847320>
- Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Da San Martino, G., ... Nakov, P. (2022). A survey on multimodal disinformation detection. *Proceedings of the 29th international conference on computational linguistics* (pp. 6625–6643). Retrieved from <https://aclanthology.org/2022.coling-1.576/>
- Arya, G., Hasan, M.K., Bagwari, A., Safie, N., Islam, S., Ahmed, F.R.A., ... Ghazal, T.M. (2024). Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*, 12, 22359–22375, <https://doi.org/10.1109/ACCESS.2024.3361322>
- Austin, A., De Silva, U., Ilesanmi, C., Likitabhorn, T., Miller, I., Fialho, M.d.L.S., ... others (2023). International consensus on patient-centred outcomes in eating

- disorders. *The Lancet Psychiatry*, 10(12), 966–973, [https://doi.org/https://doi.org/10.1111/10.1016/S2215-0366\(23\)00265-1](https://doi.org/https://doi.org/10.1111/10.1016/S2215-0366(23)00265-1)
- Bannatyne, A.J., Hughes, R., Stapleton, P., Watt, B., MacKenzie-Shalders, K. (2018). Consensus on the assessment of disordered eating in pregnancy: an international delphi study. *Archives of Women's Mental Health*, 21(4), 383–390, <https://doi.org/https://doi.org/10.1007/s00737-017-0806-x>
- Borzekowski, D.L., Schenk, S., Wilson, J.L., Peebles, R. (2010). e-ana and e-mia: A content analysis of pro-eating disorder web sites. *American journal of public health*, 100(8), 1526–1534, <https://doi.org/https://doi.org/10.1016/10.2105/AJPH.2009.172700>
- Center for Countering Digital Hate (2024). *Youtube's anorexia algorithm how youtube recommends eating disorders videos to young girls*. Retrieved from <https://counterhate.com/wp-content/uploads/2024/12/CCDH.YoutubeED.Nov24.Report.FINAL.pdf>
- Chalmers, K.J., Bond, K.S., Jorm, A.F., Kelly, C.M., Kitchener, B.A., Williams-Tchen, A. (2014). Providing culturally appropriate mental health first aid to an aboriginal or torres strait islander adolescent: development of expert consensus guidelines. *International Journal of Mental Health Systems*, 8(1), 6, <https://doi.org/https://doi.org/10.1186/1752-4458-8-6>
- Davis, H.A., Patarinski, A.G.G., Kosmas, J.A., Roske, C., Kells, M. (2025). Does watching short-form# whatieatinaday videos impact eating disorder cognitions

and urges to engage in eating disorder behaviors? an experimental investigation
of tiktok. *Appetite*, 211, 108002, [https://doi.org/https://doi.org/10.1016/j](https://doi.org/https://doi.org/10.1016/j.appet.2025.108002)
[.appet.2025.108002](https://doi.org/https://doi.org/10.1016/j.appet.2025.108002)

Diamond, I.R., Grant, R.C., Feldman, B.M., Pencharz, P.B., Ling, S.C., Moore, A.M.,
Wales, P.W. (2014). Defining consensus: a systematic review recommends
methodologic criteria for reporting of delphi studies. *Journal of clinical epidemi-*
ology, 67(4), 401–409, <https://doi.org/https://doi.org/10.1016/j.jclinepi.2013>
[.12.002](https://doi.org/https://doi.org/10.1016/j.jclinepi.2013.12.002)

Drivas, M., Reed, O.S., Berndt-Goke, M. (2024). # whatieatinaday: The effects
of viewing food diary tiktok videos on young adults’ body image and intent
to diet. *Body Image*, 49, 101712, [https://doi.org/https://doi.org/10.1016/](https://doi.org/https://doi.org/10.1016/j.bodyim.2024.101712)
[j.bodyim.2024.101712](https://doi.org/https://doi.org/10.1016/j.bodyim.2024.101712)

Facebook (2024). *How does facebook use artificial intelligence to moderate content?*
Retrieved from <https://www.facebook.com/help/1584908458516247>

Fardouly, J., Diedrichs, P.C., Vartanian, L.R., Halliwell, E. (2015). Social comparisons
on social media: The impact of facebook on young women’s body image concerns
and mood. *Body image*, 13, 38–45, [https://doi.org/https://doi.org/10.1016/](https://doi.org/https://doi.org/10.1016/j.bodyim.2014.12.002)
[j.bodyim.2014.12.002](https://doi.org/https://doi.org/10.1016/j.bodyim.2014.12.002)

Feuston, J.L., Taylor, A.S., Piper, A.M. (2020). Conformity of eating disorders through
content moderation. *Proceedings of the ACM on Human-Computer Interaction*,
4(CSCW1), 1–28, <https://doi.org/https://doi.org/10.1145/3392845>

626

627 Foster, S. (2025). *Meta's ai bias to what extent do meta's algorithmic pro-*
628 *cesses for classifying and processing information related to acts of racism*
629 *impact fairness, bias, and social justice across its platforms?* Retrieved from
630 https://digitalcommons.fau.edu/etd_general/72

631 Ghaznavi, J., & Taylor, L.D. (2015). Bones, body parts, and sex appeal: An analysis
632 of# thinspiration images on popular social media. *Body image*, 14, 54–61,
633 <https://doi.org/https://doi.org/10.1016/j.bodyim.2015.03.006>

634

635 Ging, D., & Garvey, S. (2018). ‘written in these scars are the stories i can’t explain’: A
636 content analysis of pro-ana and thinspiration image sharing on instagram. *New*
637 *Media & Society*, 20(3), 1181–1200, [https://doi.org/https://doi.org/10.1177/](https://doi.org/https://doi.org/10.1177/1461444816687288)
638 [1461444816687288](https://doi.org/https://doi.org/10.1177/1461444816687288)

639

640 Gorwa, R., Binns, R., Katzenbach, C. (2020). Algorithmic content moderation: Tech-
641 nical and political challenges in the automation of platform governance. *Big Data*
642 *& Society*, 7(1), 2053951719897945, [https://doi.org/https://doi.org/10.1177/](https://doi.org/https://doi.org/10.1177/2053951719897945)
643 [2053951719897945](https://doi.org/https://doi.org/10.1177/2053951719897945)

644

645 Griffiths, S., Harris, E.A., Whitehead, G., Angelopoulos, F., Stone, B., Grey, W.,
646 Dennis, S. (2024). Does tiktok contribute to eating disorders? a comparison
647 of the tiktok algorithms belonging to individuals with eating disorders versus
648 healthy controls. *Body image*, 51, 101807, [https://doi.org/https://doi.org/](https://doi.org/https://doi.org/10.1016/j.bodyim.2024.101807)
649 [10.1016/j.bodyim.2024.101807](https://doi.org/https://doi.org/10.1016/j.bodyim.2024.101807)

650

- 651 Hall, D.A., Smith, H., Heffernan, E., Fackrell, K., in Tinnitus International Delphi
652 (COMiTID) Research Steering Group, C.O.M. (2018). Recruiting and retaining
653 participants in e-delphi surveys for core outcome set development: evaluating
654 the comit'id study. *PloS one*, 13(7), e0201378, <https://doi.org/https://doi.org/10.1371/journal.pone.0201378>
655
656
- 657 Hanegraaf, L., Anderson, A., Neill, E., Giddens, E., Boon, E., Bryant, E., ... others
658 (2024). Treatment targets and strategies for eating disorders recovery: A delphi
659 consensus with lived experience, carers, researchers, and clinicians. *International*
660 *Journal of Eating Disorders*, 57(12), 2494–2507, <https://doi.org/https://doi.org/10.1002/eat.24304>
661
662
- 663 Hart, L.M., & Wade, T. (2020). Identifying research priorities in eating disorders:
664 A delphi study building consensus across clinicians, researchers, consumers, and
665 carers in australia. *International Journal of Eating Disorders*, 53(1), 31–40,
666 <https://doi.org/https://doi.org/10.1002/eat.23172>
667
- 668 Hockin-Boyers, H., Pope, S., Jamie, K. (2021). Digital pruning: Agency and social
669 media use as a personal political project among female weightlifters in recovery
670 from eating disorders. *New Media & Society*, 23(8), 2345–2366, <https://doi.org/https://doi.org/10.1177/1461444820926503>
671
672
- 673 Holland, G., & Tiggemann, M. (2016). A systematic review of the impact of the use
674 of social networking sites on body image and disordered eating outcomes. *Body*
675 *image*, 17, 100–110, [https://doi.org/https://doi.org/10.1016/j.bodyim.2016.02](https://doi.org/https://doi.org/10.1016/j.bodyim.2016.02.008)
676 [.008](https://doi.org/https://doi.org/10.1016/j.bodyim.2016.02.008)

678 Instagram (2024). *How instagram uses artificial intelligence to moderate content*.
 679 Retrieved from <https://help.instagram.com/423837189385631>

680 Jhaver, S., Zhang, A.Q., Chen, Q.Z., Natarajan, N., Wang, R., Zhang, A.X. (2023).
 681 Personalizing content moderation on social media: User perspectives on mod-
 682 eration choices, interface design, and labor. *Proceedings of the ACM on*
 683 *Human-Computer Interaction*, 7(CSCW2), 1–33, [https://doi.org/https://](https://doi.org/https://doi.org/10.1145/3610080)
 684 doi.org/10.1145/3610080

685

686 Lemmer, D., Mayer, G., Schrader, P., Michelsen, I., Friederich, H.-C., Bauer, S. (2024).
 687 Experts’ views on the implementation of digital interventions for eating disor-
 688 ders: a delphi study. *BMC Public Health*, 24(1), 2486, [https://doi.org/https://](https://doi.org/https://doi.org/10.1186/s12889-024-19989-3)
 689 doi.org/10.1186/s12889-024-19989-3

690

691 McMaster, C.M., Wade, T., Franklin, J., Hart, S. (2020). Development of consensus-
 692 based guidelines for outpatient dietetic treatment of eating disorders: A delphi
 693 study. *International Journal of Eating Disorders*, 53(9), 1480–1495, [https://](https://doi.org/https://doi.org/10.1002/eat.23330)
 694 doi.org/https://doi.org/10.1002/eat.23330

695

696 Meta (2026). *Suicide, self-injury and eating disorders*. Retrieved from
 697 [https://transparency.meta.com/policies/community-standards/suicide-self-](https://transparency.meta.com/policies/community-standards/suicide-self-injury/)
 698 [injury/](https://transparency.meta.com/policies/community-standards/suicide-self-injury/)

699 Narayanan, A. (2023). Understanding social media recommenda-
 700 tion algorithms. *23-01 Knight First Amend. Inst.*, , Retrieved

701 from [https://knightcolumbia.org/content/understanding-social-media-](https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms)
702 recommendation-algorithms
703

704 Nasa, P., Jain, R., Juneja, D. (2021). Delphi methodology in healthcare research:
705 how to decide its appropriateness. *World journal of methodology*, 11(4), 116,
706 <https://doi.org/https://doi.org/10.5662/wjm.v11.i4.116>
707

708 Niederberger, M., Dieudonné, J., Schifano, J., Vogt, S. (2025). Participants' expe-
709 riences with the delphi method: an online evaluation of a three-round delphi
710 across german-speaking countries. *BMJ open*, 15(12), e105878, [https://doi.org/](https://doi.org/https://doi.org/10.1136/bmjopen-2025-105878)
711 <https://doi.org/10.1136/bmjopen-2025-105878>
712

713 Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation
714 in the facebook news feed. *Proceedings of the 33rd annual acm conference on*
715 *human factors in computing systems* (pp. 173–182).

716 Ross, A.M., Hart, L.M., Jorm, A.F., Kelly, C.M., Kitchener, B.A. (2012). Development
717 of key messages for adolescents on providing basic mental health first aid to
718 peers: a delphi consensus study. *Early intervention in psychiatry*, 6(3), 229–238,
719 <https://doi.org/https://doi.org/10.1111/j.1751-7893.2011.00331.x>
720

721 Saleous, H., Gergely, M., Shuaib, K. (2025). Exploring nlp-based solutions to social
722 media moderation challenges. *Human Behavior and Emerging Technologies*,
723 2025(1), 9436490, <https://doi.org/https://doi.org/10.1155/hbe2/9436490>
724

- 725 Schmalz, U., Spinler, S., Ringbeck, J. (2021). Lessons learned from a two-round
726 delphi-based scenario study. *MethodsX*, 8, 101179, <https://doi.org/https://doi.org/10.1016/j.mex.2020.101179>
727
728
- 729 Shrestha, P., Xie, J., Delir Haghighi, P., Byrne, M.L., Griffiths, S., McNaney, R. (2026).
730 Thigh gaps and filtered snaps: a qualitative study exploring opportunities to
731 mitigate social media harm through content moderation for people with eating
732 disorders. *Journal of Eating Disorders*, , [https://doi.org/https://doi.org/10](https://doi.org/https://doi.org/10.1186/s40337-025-01504-)
733 [.1186/s40337-025-01504-](https://doi.org/https://doi.org/10.1186/s40337-025-01504-)
734
- 735 Soubutts, E., Shrestha, P., Davidson, B.I., Qu, C., Mindel, C., Sefi, A., ... McNaney,
736 R. (2024). Challenges and opportunities for the design of inclusive digital mental
737 health tools: understanding culturally diverse young people's experiences. *Pro-*
738 *ceedings of the 2024 chi conference on human factors in computing systems* (pp.
739 1–16).
- 740 TikTok (2024). *Setting the rules for a safe tiktok experience*. Retrieved from
741 <https://www.tiktok.com/transparency/en-gb/content-moderation/>
- 742 TikTok (2025). *Mental and behavioural health*. Retrieved from
743 [https://www.tiktok.com/community-guidelines/en-GB/mental-behavioral-](https://www.tiktok.com/community-guidelines/en-GB/mental-behavioral-health)
744 [health](https://www.tiktok.com/community-guidelines/en-GB/mental-behavioral-health)
- 745 Young, B., Mughal, S., Churchill, A., Rash, J., Tee, K., Salmon, A., Shah, J. (2024).
746 Developing standards for the implementation of stepped care in child and youth
747 mental health service settings: protocol for a multi-method, delphi-based study.
748 *BMJ open*, 14(12), e096453, [https://doi.org/https://doi.org/10.1136/bmjopen](https://doi.org/https://doi.org/10.1136/bmjopen-2024-096453)
749 [-2024-096453](https://doi.org/https://doi.org/10.1136/bmjopen-2024-096453)

750

751 Zeeck, A., Gruteser, V., Lau, I., Egenolf, M., Klose, C., Lienhart, S., Hartmann, A.
752 (2025). Countertransference in the treatment of patients with eating disorders.
753 *Journal of Eating Disorders*, 13(1), 240, [https://doi.org/https://doi.org/10](https://doi.org/https://doi.org/10.1186/s40337-025-01439-z)
754 [.1186/s40337-025-01439-z](https://doi.org/https://doi.org/10.1186/s40337-025-01439-z)

755

756 Zhang, C.C., Zaleski, G., Kailley, J.N., Teng, K.A., English, M., Riminchan, A., Ro-
757 billard, J.M. (2024). Debate: Social media content moderation may do more
758 harm than good for youth mental health. *Child and Adolescent Mental Health*,
759 29(1), 104–106, <https://doi.org/https://doi.org/10.1111/camh.12689>

760

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterialA.pdf](#)
- [SupplementaryMaterialD.pdf](#)
- [SupplementaryMaterialG.pdf](#)
- [SupplementaryMaterialF.pdf](#)
- [SupplementaryMaterialB.pdf](#)
- [SupplementaryMaterialI.pdf](#)
- [SupplementaryMaterialC.pdf](#)
- [SupplementaryMaterialK.pdf](#)
- [SupplementaryMaterialE.pdf](#)
- [SupplementaryMaterialH.pdf](#)
- [SupplementaryMaterialJ.pdf](#)