Automated Content Moderation Governance in terms of body image and eating disorders

Pranita Shrestha
pranita.shrestha@monash.edu
Monash University
Melbourne, Australia

Jue Xie
jue.xie@monash.edu
Monash University
Melbourne, Australia

Pari Delir Haghighi
pari.delir.haghighi@monash.edu
Monash University
Melbourne, Australia

Michelle Byrne
michelle.byrne@monash.edu
Monash University
Melbourne, Australia

Roisin McNaney
roisin.mcnaney@unimelb.edu.au
University of Melbourne
Melbourne, Australia

Abstract

The ubiquity of social media has amplified concerns about its impact on users' body image and disordered eating behaviours, particularly for individuals at risk of or experiencing eating disorders (ED). Highly visual platforms like Facebook, TikTok, YouTube, and Instagram influence the perceptions and behaviours of millions, posing particular risks for vulnerable audiences. These platforms have implemented AI-driven content moderation systems to address harmful content. However, these systems face significant issues related to bias, context, and the nuanced nature of body image and eating disorders content. Current tools for moderating harmful content fail to detect nuanced visual, audio, and text-based cues simultaneously. Furthermore, platforms must carefully navigate the delicate balance between censorship and fostering positive, supportive content, particularly as they work to protect users from the mental health risks associated with harmful content while ensuring their moderation systems are fair and transparent.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; **User centered design**.

Keywords

social media, body dissatisfaction, eating disorders, content moderation governance

ACM Reference Format:
Pranita Shrestha, Jue Xie, Pari Delir Haghighi, Michelle Byrne, and Roisin McNaney. 2025. Automated Content Moderation Governance in terms of body image and eating disorders. In *Proceedings of Conference on Human Factors in Computing Systems (CHI '25).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

CHI '25, 978-1-4503-XXXX-X/18/06

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Each year, approximately 3.3 million people experience the profound physical and psychological effects of eating disorders (ED) [23]. These are serious mental health conditions characterized by disordered eating behaviors, such as extreme food restriction or binge eating, alongside an obsessive focus on body shape and weight. Many individuals also engage in harmful compensatory behaviors such as purging or vomiting after meals [12]. Eating disorders are strongly linked to body image issues, which encompass a person's thoughts, feelings, and perceptions about their physical appearance, including aspects like shape, size, and other attributes [18].

Social media often sets the stage for distorted beauty expectations, presenting a limited and often unrealistic idea of what it means to look attractive or desirable. [22]. Constant exposure to such content can pressure individuals to conform to these unattainable standards. For instance, viral trends like the "thigh gap challenge" encourage women to strive for a visible gap between their inner thighs while standing, reinforcing unhealthy body expectations [16]. More generally, health and fitness posts often glamorize a particular body type (lean, toned, muscular, or extremely slim), which can drive disordered eating and exercise habits in pursuit of these ideals [13, 17].

Adolescents, in particular, frequently rely on social media for information about diet, nutrition, and fitness, often turning to influencers who lack professional qualifications and promote extreme, fad-based approaches. These might include cutting out entire food groups or encouraging prolonged fasting [19]. A recent study [14] exposed the disturbing impact of TikTok's recommendation algorithm: individuals with eating disorders were shown 4343% more toxic ED-related content, 335% more dieting-related posts, and 146% more appearance-focused content than typical users. This overwhelming exposure creates a dangerous feedback loop for vulnerable individuals, intensifying body dissatisfaction and reinforcing destructive behaviors like extreme dieting, purging, or self-induced vomiting after meals [12].

Social media platforms play a crucial role in global communication, however, the governance of content moderation remains a complex and evolving challenge. While platforms implement various moderation strategies to address harmful content, gaps in transparency, consistency and enforcement continue to raise concerns.

2 Lessons from Platform-Specific Governance Approaches

2.1 Role of Artificial Intelligence (AI)

2.1.1 Current state of AI being used for governance in social media. As user-generated content continues to dominate social media, content moderation has become essential for fostering a safer online space. Major platforms like Facebook [1], Instagram [5], and TikTok [8]) have outlined policies to remove pro-ED content material that portrays eating disorders as a lifestyle choice rather than a serious mental health issue. These platforms primarily rely on hashtags and keywords, such as "pro-ana" (short for anorexia nervosa), "mia" (short for bulimia nervosa), "thinspiration," and others, to identify and moderate such content. Users searching for these terms are often redirected to helpline services specific to their country [2]. However, studies reveal that users have devised ways to circumvent moderation by avoiding hashtags altogether or using modified versions that escape detection [11].

Social media platforms employ AI-driven automated content moderation systems to enforce their community guidelines and remove potentially harmful material [6, 7, 9]. For instance, Meta's community standards explicitly call for the removal of content related to child abuse, nudity, suicide, self-harm, explicit eating disorder (ED) material, and hate speech [4]. These platforms utilize AI algorithms to analyze various content forms such as visuals, audio, and associated text, including keywords, hashtags, titles, and captions, to assess their safety [6, 7, 9]. If flagged as unsafe, content may be entirely removed or restricted from certain audiences, such as users under 18.

While this marks a crucial step toward maintaining safer online communities, concerns persist regarding the transparency of these moderation algorithms. Many decisions remain opaque, and only the most extreme content is consistently flagged, leaving a substantial amount of ED-related material accessible on social media [3]. Consequently, despite moderation mechanisms, ED-related content continues to reach large audiences.

2.1.2 Need for understanding the context. One of the primary challenges in using AI for moderation is the context and nuance surrounding body image and eating disorders. Without proper guidance, it becomes difficult for the AI to understand whether the content is genuinely harmful or a part of a broader, positive conversation about eating disorders. The viral content and trends, and the power of recommendation algorithms can also create a significant issue. For example, TikTok's format is to encourage viral challenges and trends, and its recommendation algorithm expands the reach of these viral challenges. So, when trends like "legging legs" or "A4 waist challenge" went viral and started to spread rapidly, it became a challenge for AI to regulate quickly enough. This is why introducing context understanding is important in the governance model.

2.1.3 Over-blocking and censorship. One of the challenges that AI faces in general is false positives. In content moderation, it would mean possibility of overblocking useful content related to body image and eating disorder recovery or advice. There is an ongoing battle between 'freedom of speech' and public safety and well-being. Kozyreva et al. [15] explored the critical factors that can tip the scales between these conflicting interests: the extent of harm, frequency and repetition of conducting harm, and the content category. Creating blanket bans on content categories is highly infeasible [10] and over-blocking can create a negative environment.

2.1.4 Stakeholder Involvement. There is a need to involved mental health professionals, eating disorder and body image researchers and advocates to refine the moderation framework and improve the AI's sensitivity to different categories of harmful content without unduly censoring harmful or empowering content.

2.2 Human moderation

Besides automated content moderation, platforms offer reporting and flagging options. Once the content is flagged, human moderators from the platforms review the content and make the final decision regarding breach of community guidelines [6, 7, 9]. Platforms also provide options to users to unfollow, mute and block certain content to enhance their ability to curate and moderate their feeds [21]. However, this user-driven moderation is burdensome, requires critical thinking, and exposes them to potentially harmful content in the first place, which can be triggering for those with ED. AI is effective in filtering large volumes of content, but human moderators review flagged content to ensure that there is no undue censorship. It is important to properly train the human moderators in the identification of harmful content and for their own safeguards against such content.

3 Potential solution

It is important to conduct a validated alignment of AI for moderation with the context and nuance surrounding body image and eating disorders, and then conduct auditing of the alignment to ensure that the AI for moderation effectively identifies harmful social media content related to body image and eating disorders. This approach involves four key stages: 1) Creation of alignment and auditing rule development, 2) Expert validation of developed rules via Delphi Study, 3) AI alignment, and 4) Validation and auditing process

3.1 Creation of alignment and auditing rule development

The rule development is a foundational step in alignment and auditing of the AI for moderation for harmful content detection about body image and eating disorders. This phase focuses on capturing diverse perspectives to inform the development of nuanced and context-aware factors in identifying the harmful content of this particular space. This will guide and govern what AI needs to identify and flag while considering the harmfulness factor. In this phase, it is important to involve mental health professionals who have expertise in body image and eating disorders field (experts by profession), and individuals with lived experience of eating disorders (experts by lived experience). The experts by profession will provide evidence-based perspectives on what constitutes harmful content and effective prevention strategies whereas the experts by lived experience will provide the first-hand insights into harmful content, triggers and social media experiences.

3.2 Expert validation of developed rules via Delphi Study

Delphi study can be used to achieve expert consensus on identifying harmful social media content related to body image and eating disorders. A Delphi study is a structured and systematic process that is used to gather consensus on expert opinions or certain topics [20]. It is a highly robust methodology that is generally used in healthcare research. The Delphi study can leverage the knowledge, perspectives and experience of experts by profession (researchers and professionals working in the body image and eating disorders space) and experts by lived experience (individuals with a history of eating disorders). Since the detection and mitigation of harmful content related to body image and eating disorders is complex, nuanced and emerging, the use of the Delphi study can ensure that the AI are informed by real-world perspectives and aligns with societal needs and values. The consensus-driven rules will serve as the foundation for the AI for moderation in detecting harmful content for body image and eating disorders.

3.3 AI alignment

The aim is to enhance the AI's understanding of harmful social media content by providing it with additional context about why certain material may be considered damaging. This will improve its ability to categorise social media content related to body image and eating disorders.

3.4 Validation and auditing process

The next step will involve human evaluators from diverse backgrounds (with appropriate precautions) to review and validate the AI's outcomes for the social media content. These evaluators will audit whether the classifications generated by the aligned AI (with rules) align with the gold standard. By incorporating our developed rules, the AI is expected to classify and contextualize social media content more effectively and accurately.

To audit the AI's performance (with and without alignment), there is a need to use both quantitative and qualitative methods. Quantitative analysis will measure the alignment between the AI's outcomes and the expert classifications, while qualitative analysis will explore the quality and contextual accuracy of the generated responses. Additionally, incorporating evaluators from diverse backgrounds will enhance the reliability of the validation process and help minimize bias, while also contributing to the diversity of perspectives within the dataset.

References

[1] 2021. How We're Supporting People Affected by Eating Disorders and Negative Body Image. https://doi.org/news/2021/02/supporting-people-affected-by-eating-disorders-and-negative-body-image/

[2] 2021. 'Thinstagram': Instagram's algorithm fuels eating disorder epidemic. https://doi.org/articles/thinstagram-instagrams-algorithm-fuels-eating-disorder-epidemic

[3] 2023. *Eating-disorder patients confront Meta bosses at federal parliament as research highlights TikTok's negative influence.* https://doi.org/news/2023-09-14/eating-disorder-patients-to-confront-meta-bosses-at-parliament/102847320

[4] 2024. *Facebook Community Standards.* https://doi.org/en-gb/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards

[5] 2024. Help Center. https://doi.org/252214974954612

[6] 2024. *How does Facebook use artificial intelligence to moderate content?* https://doi.org/help/1584908458516247

[7] 2024. *How Instagram uses artificial intelligence to moderate content.* https://doi.org/423837189385631

[8] 2024. Safety Center. https://doi.org/safety/en/eating-disorder

[9] 2024. *Setting the rules for a safe TikTok experience.* https://doi.org/transparency/en-gb/content-moderation/

[10] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1201–1213.

[11] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511.

[12] Sasha Gorrell and Stuart B. Murray. 2019. Eating Disorders in Males. *Child Adolesc Psychiatr Clin N Am* 28, 4 (2019), 641–651. https://doi.org/10.1016/j.chc.2019.05.012

[13] Scott Griffiths, David Castle, Mitchell Cunningham, Stuart B Murray, Brock Bastian, and Fiona Kate Barlow. 2018. How does exposure to thinspiration and fitspiration relate to symptom severity among individuals with eating disorders? Evaluation of a proposed model. *Body image* 27 (2018), 187–195.

[14] Scott Griffiths, Emily A Harris, Grace Whitehead, Felicity Angelopoulos, Ben Stone, Wesley Grey, and Simon Dennis. 2024. Does TikTok contribute to eating disorders? A comparison of the TikTok algorithms belonging to individuals with eating disorders versus healthy controls. *Body Image* 51 (2024), 101807.

[15] Anastasia Kozyreva, Stefan M Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences* 120, 7 (2023), e2210666120.

[16] Kerrie Caitlin Leonard. 2020. *The Impact of Social Media Body Challenges on Youths' Body Image.* Master's thesis. North Dakota State University.

[17] Ciara Mahon and David Hevey. 2021. Processing body image on social media: Gender differences in adolescent boys' and girls' agency and active coping. *Frontiers in psychology* 12 (2021), 626763.

[18] Ganesh Kumar Mallaram, Pragya Sharma, Dheeraj Kattula, Swarndeep Singh, and Poojitha Pavuluru. 2023. Body image perception, eating disorder behavior, self-esteem and quality of life: a cross-sectional study among female medical students. *Journal of eating disorders* 11, 1 (2023), 225.

[19] Erin L Moorman, Jennifer L Warnick, Ratna Acharya, and David M Janicke. 2020. The use of internet sources for nutritional information is linked to weight perception and disordered eating in young adolescents. *Appetite* 154 (2020).

[20] Prashant Nasa, Ravi Jain, and Deven Juneja. 2021. Delphi methodology in healthcare research: how to decide its appropriateness. *World journal of methodology* 11, 4 (2021), 116.

[21] Rachel F Rodgers, Siân A McLean, and Susan J Paxton. 2024. Enhancing understanding of social media literacy to better inform prevention of body image and eating disorders. *Eating Disorders* (2024), 1–19.

[22] Rachel F Rodgers and Tiffany Melioli. 2016. The relationship between body image concerns, eating disorders and internet use, part I: A review of empirical support. *Adolescent Research Review* 1 (2016), 95–119.

[23] Daphne van Hoeken and Hans W Hoek. 2020. Review of the burden of eating disorders: mortality, disability, costs, quality of life, and family burden. *Current opinion in psychiatry* 33, 6 (2020), 521–527.