Digital Society (2025) 4:8
https://doi.org/10.1007/s44206-025-00166-x

ORIGINAL PAPER

Experiences of Moderation, Moderators, and Moderating by Online Users Who Engage with Self-Harm and Suicide Content

Zoë Haime[1,2] · Laura Kennedy[3] · Lydia Grace[3] · Lucy Biddle[1,2,4]

Received: 10 May 2024 / Accepted: 22 January 2025 / Published online: 18 February 2025
© The Author(s) 2025

Abstract

Despite the growing role of content moderation online, particularly in mental health spaces, there is limited research into the effectiveness of platform practices and a lack of user-driven evidence for regulatory guidance. This study aimed to explore user accounts of moderation related to self-harm and suicide (SH/S) content online, including their experiences of being moderated and perspectives on moderation practices. Additionally, where participants were also moderators, their experiences of moderating SH/S content were explored. 14 participants were interviewed at baseline, $n=8$ at 3-months and $n=7$ at 6-months. They also completed daily diaries of online use between interviews. Thematic analysis was used to explore perspectives. Three key themes were identified: 'content reporting behaviour', exploring factors influencing decisions to report content; 'perceptions of having content blocked', exploring experiences and speculative accounts of SH/S content moderation; and 'content moderation and moderators', examining participant views on moderation approaches and their experiences of moderating. This study revealed challenges in moderating SH/S content online, and highlighted inadequacies with current procedures. Participants struggled to self-moderate online SH/S spaces, showing the need for proactive platform-level strategies. Additionally, whilst the lived experience of moderators was valued, associated risks emphasised the need for supportive measures. Policymakers and industry leaders should prioritise transparent and consistent moderation practices.

Keywords Moderation · Online harm · Thematic analysis · Suicide · Self-harm · Qualitative

Extended author information available on the last page of the article

1 Introduction

As online platforms facilitate connections between individuals on a global scale, it becomes crucial to implement practices that maintain safe environments through content moderation (Roberts, 2019; York & Zuckerman, 2019). Platform-level moderation is used to regulate user-generated content across entire platforms, and can involve automated tools, algorithms, and human moderators working to apply overarching policies and standards. Alternatively, community-level moderation involves active participation of users, through reporting or flagging content that they find inappropriate, or against the rules of specific spaces or communities within platforms. These moderation actions at platform and community level may be particularly relevant in controlling engagement with sensitive, harmful or illegal material, including self-harm and suicide (SH/S) content, and discourse around it.

1.1 Moderation Effectiveness

To regulate online spaces, moderators play a vital role in screening content, identifying problematic users, and enforcing rules (Jhaver et al., 2019). The mechanisms employed, such as blocking or removal of content, and semi- or permanent-banning of users are thought to ensure the availability of high-quality content, whilst limiting the presence of harmful material (Kiesler et al., 2012). However, experiences and outcomes of these moderating actions can vary for users. For example, when Facebook and Instagram introduced a total ban of graphic self-harm imagery, sentiment analysis revealed an increase in anger, anticipation, and sadness in the associated Twitter [renamed 'X'] discourse (Smith & Cipolli, 2022).

Complexity in moderation decision-making is particularly evident on platforms focused on mental health. Moderation in these spaces must balance the responsibility of protecting users from triggering content, with the provision of space for social support and recovery (Saha et al., 2020). The effectiveness of current moderation practices, including outright content bans in mental health spaces, have been further questioned. Multiple studies have investigated how users seek out potentially harmful pro-eating disorder content online and evade moderation systems by posting content using no text, algospeak, or alternative hashtags (Gerrard, 2018; Chancellor et al., 2016). This practice is also prevalent amongst SH related posts on Instagram, where the use of ambiguous and unrelated hashtags leads to graphic SH imagery without proper content warnings (Moreno et al., 2016). Additionally, research shows that even where platforms proactively search for and remove harmful content, this can be inefficient, as content reappears on the same or alternative platforms (Stoilova et al., 2021; Biddle et al., 2016).

1.2 Who Should Moderate Content?

There are also concerns regarding the allocation of responsibility for moderating online spaces. Platform-level moderation usually relies on paid humans, although these mechanisms are sometimes automated. However, within peer-driven communities, users themselves often moderate content, usually in voluntary roles. While sev-

eral benefits to the moderator role have been reported, including feelings of altruism, having a sense of purpose (Steiger et al., 2021) and – in the case of lived-experience moderators – receiving validation of their own experiences (Coulson & Shaw, 2013), concerns have also been raised for both community and platform level moderators' about the impact on their own mental health (Steiger et al., 2021; Saha et al., 2020).

A recent study by Spence et al. (2023), found 40.8% of 213 online platform-level content moderators were exposed to distressing content daily, and that they showed a dose-dependent relationship between frequency of exposure to distressing content, and psychological distress and secondary trauma. However, their findings also revealed, that after accounting for work factors in the analysis, including access to supportive colleagues and feedback about their work, the relationships failed to remain significant, indicating that a supportive work environment may ameliorate negative effects.

Some platforms and sites, such as Google and Facebook, have begun to implement strategies to address risks to their content moderators, such as training on working with sensitive content, recommended access to both individual and group counselling services, and suggested 'wellness breaks' (Google, 2023; Halevy et al., 2020). However, the effectiveness of these interventions for moderator wellbeing, remains unclear. Additionally, whilst large platforms are able to dedicate resources to moderator care, there are less viable options available to communities hosted on those platforms, or to smaller individual platforms.

Even within larger platforms, it may also be challenging to ensure uniform care across all moderators, due to varying work environments and individual needs. For instance, diversity in preferences of moderators was shown by Saha et al. (2020). Some moderators desired support from mental health experts and welcomed the idea of having trained professionals within the moderation team. However, other moderators emphasised that these spaces should not provide medical advice to users, and that moderators with lived experience of mental health problems can create the supportive community users are looking for, without additional help. This highlights the differing needs amongst moderators in mental health spaces and emphasises the need for further understanding of individual moderator narratives to optimise their experiences.

1.3 The Online Safety Act

Policy and regulatory frameworks also play a significant role in the expectations and responsibilities of online platforms in moderating content. In the United Kingdom, there has been a growing emphasis on addressing online harms, through the recently enacted Online Safety Act (2023). Key points within this legislation include the need for platforms to remove SH/S content, implement measures to minimise user exposure to harmful content, uphold a duty of care towards users, and provide individuals with improved mechanisms to report harmful content. Whilst these regulations attempt to minimise harm to users and enhance platform accountability, their positive impact, particularly for those purposefully engaging with mental health content online, remains unclear.

1.4 Current Gaps in the Research

To date, limited research has been conducted exploring experiences of moderation by users who engage with online mental health content, or of those who moderate it. Notably, existing findings suggest that moderation is essential to mental health online spaces (Saha et al., 2020) but that current approaches may be ineffective, and, in some cases increase vulnerability to harm for users and moderators. For instance, content posted with limited information in order to avoid automatic removal systems, increases the risk to users of unexpectantly encountering potentially triggering content (Gerrard, 2018), and moderators may experience emotional distress and burnout (Coulson & Shaw, 2013). A noticeable research gap exists in understanding user perspectives on the moderation of SH/S related content online, particularly user experiences of having their own SH/S related content moderated (reported or removed), reporting others' SH/S content, seeing others' SH/S content being removed, or of taking on lived-experience moderator roles themselves. Understanding the perspective of users engaging in SH/S online spaces is essential for informing effective, responsible, and user-centric moderation strategies, including industry guidance and legislative and policy recommendations, for the digital space.

1.5 Aim of the Study

This study aimed to gather and analyse in-depth first-hand user accounts of moderation, moderators and moderating from individuals engaging with SH/S online spaces in order to address the current gaps in understanding of user preferences and experiences. The findings provide critical insights into the effectiveness of current moderation practices and contribute to the development of industry and policy guidelines for online moderation procedures by drawing on evidence from those directly affected.

2 Method

2.1 Procedure

Data collected as part of the DELVE study (Haime et al., 2024) was analysed. Participants were interviewed at three time points (baseline, midpoint, endpoint) over a six-month study period about their engagement with SH/S content online. Interviews were semi-structured, with researchers employing an open-ended approach using probing to obtain detailed, participant-led accounts. In addition, participants were asked to maintain daily research diaries during the intervening periods. In the interview, 'moderation' was a topic area where participants were asked to share their thoughts on moderators they had encountered online, their experiences of others', and their own SH/S content being moderated, and any experiences of moderating SH/S content themselves.

The study was approved by University of Bristol Faculty of Health Sciences Ethics Committee (approval no. 117491). All participants provided written informed con-

sent prior to participation. In order to ensure the safety and wellbeing of study participants, they were required to complete a mandatory safety plan, including details of a safety contact, their general practitioner's details, and a self-care plan that was individually designed by each participant. Further, a distress protocol was developed with a clinician to manage any risks of worsening mental health or self-harm behaviour during participation in the study. A hierarchy of responses was specified in cases of distress, including the options of offering follow-up advice from a named senior clinician (see Haime et al., 2024 for more details).

2.2 Participants

Participants were 16 years and over, English-speaking, and online users who had engaged with SH/S content. Recruitment methods (outlined in Haime et al., 2024) involved multiple channels such as social media platforms, a mental health app targeting young people, and charity websites and newsletters. Participants responded to study adverts and were assessed for eligibility before being sent the study information sheet. Purposive sampling was used to promote diversity in gender, ethnicity, and age of sampled participants.

2.3 Analysis

Thematic analysis (Braun & Clarke, 2012) with deductive coding informed by interview questions was used, with several steps:

1. Interviews were transcribed, and researchers read and familiarised themselves with both the participant transcripts and their diary data.
2. ZH began coding the data. Codes were systematically assigned to quotes which captured certain concepts, or ideas about moderation or moderators. Coding was iterative, and codes were refined and revised in consultation with LK and LB.
3. Similar codes were grouped together to form themes representing patterns in concepts or ideas. Themes were then explored narratively.

3 Results

3.1 Participants

Fourteen participants took part. The sample had diversity in terms of age and ethnicity, and relatively good representation of gender (Table 1). All participants completed a baseline interview, eight completed midpoint and seven completed endpoint interviews (see Haime et al., 2024). Due to participant dropout, overall 44 diaries were anticipated, with a total of 31 (70.5%) diaries actually returned. All interview transcripts and diary data were used in this analysis.

Table 1 Demographic characteristics of participants who took part in the DELVE study

Demographic Variable	Total *N* (%)
Gender	
Female	10 (71.4)
Male	4 (28.6)
Ethnicity	
Asian British	2 (14.2)
Black British	1 (7.2)
White British	7 (50.0)
Asian Other	2 (14.2)
Black Other	0 (0.0)
White Other	1 (7.2)
Mixed Race	1 (7.2)
Age	
16–17 Years	1 (7.2)
18–24 Years	7 (50.0)
25–35 Years	0 (0.0)
36–45 Years	4 (28.6)
46–54 Years	2 (14.2)
55 + Years	0 (0.0)

3.2 Thematic Analysis

Results are visually depicted in Fig. 1 by theme and subtheme. These are explored in more detail below.

3.2.1 Content Reporting Behaviour

3.2.1.1 Mental Health State and Awareness Four participants shared how they did not consciously consider reporting SH/S content online due to their poor mental state. In fact, during challenging times, users described becoming preoccupied with deliberately seeking out this material, thereby prioritising the perceived benefits of content over reporting it:

> I think in that moment it didn't really cross my mind [to report suicide methods information]. I was too focused on the content itself and what it was giving. (IDJ)

A couple of participants thus described how their decisions to report SH/S content were multifaceted and variable:

> Part of me thinks they should be reported and another part of me thinks, well, freedom of speech and they should stay…I think that's one of those things that could change and fluctuate depending upon my own mental health. (IDC)

Participants often recalled no hesitation in reporting general online content considered morally wrong, such as racist comments or images. However, decisions to report



Fig. 1 Visualisation of themes and subthemes. Alt text: Image visualising the themes 'Content Reporting Behaviour', 'Content Moderation and Moderators', 'Perceptions of Having Content Blocked' and their subthemes, surrounding the text reading 'Online Self-Harm and Suicide Content: Experiences of Moderation, Moderators, and Moderating'

SH/S content were more selective. One user (IDD) reflected on the emotional impact of content on a pro-suicide forum affecting their decision: 'if it does affect me then … I like reduce time on it or report it to one of the moderators for example' (IDD). Another (IDM), described how their lack of knowledge about SH/S content they came across made them refrain from reporting:

> Why didn't you report it [graphic self-harm image]? (Interviewer)
>
> I didn't know much about it and so I decided maybe to just not show it to anyone. (IDM)

3.2.1.2 Effects on Other Users Considering the implications for users posting SH material, IDM also expressed reluctance to report images without first understanding the posters' circumstances, and suggested having a conversation with them may be more appropriate:

> I think you would have to talk to them, maybe see where they're at and maybe you can report, or maybe take it to the next step with them [rather] than reporting them … I don't think it would be appropriate to [just report] them. (IDM)

Despite concern for content posters' wellbeing, and attempts not to stigmatise their experience, there was general agreement amongst the other participants that images or videos of SH/S were largely inappropriate and justified reporting because of the detrimental impact they can have. One participant, reflected:

> I can get it if someone is saying oh, I'm really suicidal, I won't report that, because maybe they need someone to talk to, if they are posting images on that though, you can't be doing that…you can spiral into that…I ended up on a video and then it went to another video and then I ended up on a site and then yeah…
>
> I think I took an overdose afterwards (IDK)

One participant (IDI) also described how reporting served not only as a means of addressing the issue of SH/S content posted online, but as a way to raise awareness and educate the person responsible for the content: 'It also, hopefully, tells that person that what they were posting might have been sensitive towards someone' (IDI).

Some participants also described the consequences of reporting other users' content. While a couple described content being taken down (IDK, IDI), or the user being 'kicked off' (IDD) a group, others commented on specific platform actions, stating Facebook do 'nothing' (IDB) when reporting nuisance pages, or that reporting on Tumblr felt 'completely pointless…because nothing would happen' (IDF).

IDI acknowledged that although the automated process of reporting using Instagram options made it relatively easy to do, it left no room for nuance in describing the content: 'I don't think there's much of an opportunity to fully explain anything or anything like that' (IDI). Additionally, despite reported SH/S content being removed on Instagram, users were able to continue posting without consequence:

> [they can] carry on posting as normal anyway. So it just means that they can post it again' (IDI)

Finally, IDE acknowledged how it would be impractical to report all content by others that was offensive or inappropriate, and therefore felt they had to make careful decisions about what they reported. They did not detail how they made those choices, other than to note they were more likely to report those with significant follower counts, protecting other online users: 'Yeah, when I see something that I don't like, like self-harm and they've got lots and lots of followers then I do report stuff (IDE).

3.2.1.3 Practical Processes One participant described how their non-reporting was influenced by a lack of awareness about how and where to report on certain platforms (IDH).

However, those who had reported content described it as a straightforward process in practical terms, with some sites having specific options for recording if the content was related to SH/S.

Participant (IDE) also reflected on how following predetermined methods for reporting, like using checkboxes, allowed the process to be less biased, suggesting

an attempt to detach themselves from the feelings of guilt associated with reporting an online peer:

> What I'll do is I'll click all those [check] boxes and I'll try to desensitise it or make it less personal or not take it personally. So, I try to be as objective as possible about it. (IDE)

3.2.1.4 Influence of Research Participation Three of the participants who previously had not reported content, mentioned during interviews that they would now report SH/S content they came across online, using the platform guidelines to direct them. It is unclear what motivated this change, though there was some indication it may be explained by involvement in the research itself:

> My uses have changed massively because it's really [I've] been more aware, but also, I'm more conscious of it. So when you are a user you just look at it and think, 'Okay.' But when you do something like [taking part in this research study] you think, 'Hang on, actually there's another dimension to it and is there harm or should I report?' So you become a social media – I don't know if it's an influencer – but you become a sort of [an] active or civic user. (IDE)

3.2.2 Perceptions of Having Content Blocked

3.2.2.1 Content Blocking by Sites and Platforms Five participants were able to describe their experiences of having content moderated online, including content expressing suicidal ideation (IDB, IDK), privately posted images of SH (IDL), and blogs about depression and self-harm (IDN). Participants described receiving differing levels of information from sites when their content had been blocked, with some feeling that the rules were unclear:

> I have like one or two blocked [blogposts], but I don't know why. I don't know which words [are] the restriction. (IDN)

IDL, who used Instagram to post images privately, documenting and tracking their SH journey, recalled instances of content being blocked without warning. This interfered with their ability to monitor their well-being, and disrupted their sense of control and ownership:

> …Under [the] notification page they'll just say, your post has been removed because of a violation. And I think it's quite frustrating because I don't know what I've written so I can forget what it was about … even if they said that, I want to remember and recall how I was feeling and the tough times I'd been through. But they don't give you any warning at all, they just take it (IDL)

However, it seemed the platform was inconsistent with their moderation, leaving the user uncertain what the consequences might be when posting:

> …only sometimes they take it down, sometimes they don't. It's kind of confusing, I don't know what they are [doing] (IDL)

Users also criticised the lack of diversity in reactions by platforms when moderating content. IDK, who had once posted they were 'struggling' on Twitter [X], had this reported by another user, and criticised the platform's 'one-size-fits' all response of removing the post and sending suicide support contact details, which they felt were irrelevant to their current, money-related difficulties.

One participant who had no personal experience of their content being reported or blocked by platforms, also reflected on the potential emotional impact if it were to happen, expressing they would feel 'disempowered' and as if 'you're not being heard' (IDC).

3.2.2.2 Content Blocking by Community Moderators In some cases, participants chose to engage with SH/S content in spaces regulated by community moderators, as well as by the platforms themselves. These moderators had responsibility for enforcing community rules and guidelines, which were sometimes more specific or stricter than those set by the platform (e.g., individual Facebook groups on Facebook).

Two participants shared experiences of having content removed from these spaces. Again. both were left feeling confused and frustrated due to inconsistent and unclear moderation policies. For instance, IDG posted 'what is the point?' on a forum, which led to moderators removing their post and contacting them with resources of support. The participant expressed feeling '… just like a criminal. Well not a criminal, but like I've really done something wrong.' (IDG). They also seemed surprised to see other content on the site they deemed more inappropriate hadn't been removed.

The other participant (IDB) recalled two instances of moderators blocking content in an online private Facebook group that provided SH peer support. Initially, they posted a suicide note on the group, and moderators replaced it with a supportive post, tagging the participant. This action resulted in an influx of messages of support from other users. However, this felt intrusive, overwhelming and made the participant feel guilty, placing further strain on their mental state:

> it was driving me up the wall … I was like I've made my decision, I know what I'm doing. I don't want you to all tell me to stop doing it, I just want some help with [suicide plans], can you please just stop because you're just making me feel guilty and you're not helping me … Obviously it was nice but in that state I just couldn't take it. (IDB)

In a separate event, the participant had a post removed by moderators when in disagreement with another user about their (IDB) giving mental health advice. The participant described feeling frustrated as a result of this interaction and that moderation could be biased:

> I started to get really wound up, and then the people that were managing the page took my comment off and blocked me from commenting on it anymore

> but allowed for her [other user] to continue, which just drove me insane because I felt that was so unfair. (IDB)

3.2.3 Content Moderation and Moderators

3.2.3.1 Moderation Approaches and Strategies As participants generally perceived that some SH/S content should be accessible online, regulation and moderation of content was viewed as a necessary part of making those spaces safe:

> …there is a place for it [SH/S content] online…but it needs to come from a very supportive place and it needs to be highly monitored and regulated. That's what I've generally learnt. The safer places to be online and the safer places to deal with it are when it is more moderated and looked after. (IDB)

Participants, owing to their different encounters with SH/S content online, experienced varying interactions with moderators across platforms. Three participants (IDA, IDB and IDC) who regularly engaged with a SH support group hosted on Facebook, mentioned being drawn to their moderation methods, due to their clear, friendly approach, and close monitoring of content and rapid moderation:

> I tend to stick to [self-harm support Facebook group] because I know that it's heavily moderated (IDB)

> If anything like that [harmful posts] does crop up on [SH support Facebook group] it's gone within minutes you know. An hour is long for it to have still been on there. (IDB)

However, it is important to note that these perspectives may have been subject to bias within the sample. Notably, IDA and IDB who regularly engaged with the SH support Facebook group, also moderated for them in a voluntary capacity, which may have influenced their views.

Participants also provided insights into the moderation practices of prominent platforms such as Twitter [X] and Facebook. Interestingly, participant experiences seemed to contradict one another. For instance, one participant criticised Facebook for using 'a carpet ban' (IDC) when it came to SH/S content. In contrast, another participant (IDK) reported that moderation practices on Facebook were relaxed and 'lenient' compared to Twitter X), though it is unclear whether participants were referring to community or platform moderation in these circumstances.

3.2.3.2 Improvements to Moderation Approaches Participants identified areas of moderation that they would like to see improved across platforms. One common concern was the lack of transparency and consistency in what content would be blocked or removed:

> In terms of restrictions for social media overall, I personally don't feel they're very open about exactly how they moderate posts. So I think that it's difficult to

> understand what their thought processes [are] behind blocking posts or restricting posts. (IDI)

Specifically, for SH/S content, participants emphasised that current platform regulations were too reactive, resulting in punitive measures such as immediate bans or content removal, whereas a more nuanced approach would ensure support is provided to those who need it:

> …they seem to throw the baby out with the bath water. They do a formal, catch you all sort of ban process there or suspend here…But where it comes to suicide and self-harm maybe it's a point of that actually this is a serious situation where we need to be looking at this openly … rather than actually someone's put themselves on the line to say 'I'm feeling suicidal.' All of a sudden it's all guns blazing, red lights etcetera.…more safety-conscious rules and policies in place, that's what we need to have. (IDG)

Specifically highlighting the portrayal of SH scars, IDI echoed the call for a more thoughtful and empathetic approach to moderating online SH/S content. They emphasised a need to recognise scars can become a part of someone's image and identity, and so removing the content would be disregarding their experiences and self-expression:

> it's like they've got a one size fits all rule thing where any posts about self-harm, any scars at all, they'll just remove the pictures, whereas I've seen some people talk about how their scars are like recovery scars and just a part of who they are. So I think that improvement is still needed at the moment even just to establish the baseline of what's right and what's wrong to post about. (IDI)

A more multidimensional approach was encouraged for companies making decisions around SH/S content policy:

> we need to get the social media firms to get their position right on what content they're allowing to show on their platforms. They need to be real as well and actually work with local, national governments, but they also need to be working with like voluntary and community sectors of people that live with it. And also people with lived experience as well to actually find balance. (IDG)

However, other participants favoured more rigorous moderation, particularly to protect vulnerable users, such as young people, who they acknowledged will seek out SH/S content and need safe spaces to engage:

> If it was my child, I wouldn't want them accessing that [SH/S] content (IDF)

> I feel like if teenagers or young people are gonna go on it [SH/S spaces online] then they need somewhere that will moderate it by professionals … rather than just being on Facebook or Instagram for example. (IDD)

One participant (IDG), who was exposed to graphic self-harm imagery on a mental health Facebook group, noted that approving posts before they are shared could be an easily implemented moderation method.

3.2.3.3 Professional vs. Lived Experience Moderators Some participants went on to describe whether the moderator role in online spaces should be undertaken by trained professionals or by those with lived experience. Two participants (IDB & IDC) argued for moderators with personal experience of SH/S.

The general belief was that these moderators would have a better understanding of the content posted in the online communities allowing them to make informed judgements about the appropriateness of content, in context:

> it's not like you need like a psychiatrist or a counsellor on there or something like that. You know you don't need that you just need somebody that's been accountable for it … you know that's not just some random person. (IDB)

Despite this, one individual preferred the concept of mental health professionals moderating, or for peer moderators to undergo training:

> It definitely needs to have someone who's trained in either like mental health first aid and specifically suicide and self-harm or like a proper professional (IDJ)

Participants also highlighted the importance of implementing mechanisms to prevent those with lived experience becoming too entrenched in a community's mindset, which may be harmful to them. One suggestion was to have regular changes in moderators as a preventative measure (IDH). Another was to provide support to moderators from mental health professionals, to ensure they could handle content they may encounter, and maintain their own mental wellbeing (IDB).

3.2.3.4 Participants as Moderators Three participants had become moderators or administrators on the platforms where they accessed SH/S content. The decision to take on this role was driven by a hope to help others, and ensure rules within the online spaces were adhered to, as well as an altruistic way of giving back to the community:

> If I can help support that even a little bit to make it a bit safer and keep people safe that are also vulnerable, it's just my little way of saying thank you, I guess, for the support I've got and enabling other people to be able to still continue to get that support. (IDB)

These participants reflected on the significant responsibility associated with moderating online SH/S content. IDI shared their personal recognition of the need to disengage from a moderation/administrator role when their own mental health declined. They expressed concern about their ability to effectively address someone else's

issues in such circumstances, as well as the potential negative impact of engaging with another users' distressing content:

> I think the other thing is I won't open a DM unless, again, I'm in a mood where I think I'll be able to handle it because I don't want a clash of negativity or anything to end up going wrong and … it's the whole responsibility thing. So you don't want someone to hear your point of view when you're feeling negative yourself (IDI)

However, IDB seemed less able to manage their behaviour in this way. They recalled an incident where they encountered a video in the group they moderated that presented a SH method that was novel to them:

> Obviously, it [SH method post] got taken down from the main group … I'm first line of defence. It's just that I keep getting hit by it at the moment because I'm the one that's moderating it … and since then, I've bought [SH equipment] to try and do it myself. (IDB)

IDB, when considering the best approach to moderation, drew upon their own experience of encountering strict moderation methods. They found this experience influenced their gentler approach, recognising how more forceful moderation could potentially isolate vulnerable users, making them feel worse:

> Different people do it different ways. I'm normally quite gentle because quite often, especially with the [Self-Harm Support Group] rules, the reason they're breaking the rules is from a really nice place…We have got some people on the other end of the spectrum who say, 'You agreed to the rules when you joined this group. You are breaking the rules. This is your one and only warning.' I've been on the receiving end of that one and my first warning was like that and it made me nearly leave the group entirely … (IDB)

Additionally, IDI emphasised their approach to moderation involved being open and transparent, a quality they had criticised platforms for lacking in their own moderation practices:

> I think if I thought that a comment would be damaging to someone else, then I'd delete it and then I'd message the person who sent the message to explain why I've done that. (IDI)

4 Discussion

Findings from this study suggest that although users engaging with SH/S content online favour the use of moderation and moderators in these spaces, the current methods used by platforms and communities may be inadequate to provide a safe and effective user experience due to inconsistencies, ambiguities, and biases.

Many online environments rely on user reporting to ensure moderators can successfully enforce community rules (Jhaver et al., 2019). However, despite acknowledging their role as 'civic' users (Gardenier et al., 2024), participants in this study revealed an inherent inability to undertake community moderation through reporting SH/S content, particularly during times of poorer mental health,. Participants revealed these struggles to report content stemmed from a fear of potentially stigmatising other users and a desire to access the content themselves, despite recognising the potential dangers of the material, aligning with research by Saha et al. (2020). This emphasises the complexity of decision-making processes for vulnerable users in online spaces and highlights the importance of mechanisms that empower users to report harmful content. Such mechanisms must aim to minimise potential negative consequences for reporters and reportees, whilst providing appropriate follow-up support (Haime et al., 2024).

Although participants described difficulties in reporting SH/S content, several users did take on moderating roles in these online environments, with largely altruistic intentions. However, they also found their ability to moderate fluctuated alongside their own mental health, a concern recognised in previous research (Steiger et al., 2021; Saha et al., 2020). While one user was able to identify their mental health declines and take action to protect themselves and others' by temporarily withdrawing from their moderating duties, another was unable to safeguard themselves and became at risk of harm whilst undertaking their moderating role. This highlights the necessity of platforms taking control to ensure moderator wellbeing through the implementation of appropriate and effective measures of support, rather than solely relying on moderators' ability to recognise their own wellbeing needs (Saha et al., 2020).

Participants largely agreed that lived-experience moderation was important, due to the moderators' understanding of user experiences. However, many also emphasised the potential difficulties and harmful consequences a lived-experience moderator may experience by engaging with content and users in these spaces, consistent with Spence et al's (2023) findings on the heightened psychological distress associated with moderation of distressing content. Whilst some participants supported professionals undertaking these moderator roles instead, recent research by Perry et al. (2022) indicates that professionals moderating an online suicide prevention space often felt restricted in utilising their therapeutic skills, limiting their capacity to provide meaningful support. This may suggest the way professionals are currently being utilised in the moderator role is ineffective, and their contributions need to be re-evaluated to ensure both they and platform users benefit from their input.

Alike to participants in previous research (Saha et al., 2020) and recommendations from Facebook (Halevy et al., 2020), some participants in this study encouraged the introduction of a more hybrid model, that incorporates mental health professionals supervising and training lived-experience moderators. This approach may better support lived-experience moderator wellbeing, ensuring they are empowered to manage their roles effectively and enhance their safety, whilst also providing professionals with roles where they feel able to appropriately apply their expertise.

When considering SH/S content moderation by others, participants in this study described a lack of consistency in how platforms and community moderators

responded to SH/S content, which may reflect non-standardised approaches to moderation of mental health content across online spaces (Haime et al., 2024), issues with recommendation algorithms (Goswami, 2024), and inefficient safety mechanisms (Milton et al., 2023). Additionally, poor transparency in communication of moderation methods and results not only led to disappointment, frustration, and a perceived loss of control amongst participants, but also added a speculative element to many participant accounts. Several also criticised 'catch-all' reactive policies for SH/S material, where content could be immediately removed, or users banned without warning. These punitive moderation methods were viewed as non-empathetic, harmful, and potentially stigmatising, consistent with user experiences of Facebook and Instagram's ban of graphic self-harm imagery, and of 'heavy-handed removal or banning strategies' for mental health related content on TikTok (Smith & Cipolli, 2022; Milton et al., 2023).Many 'big tech' platforms, including Facebook and TikTok, have attempted to overcome harmful mental health-related content accessibility on their platforms by using automated moderation systems. Transparency reports provided by platforms themselves reveal that the majority of SH/S content removed is flagged by these systems. This shows some platforms are making concerted efforts to tackle SH/S content and attempting to relay this data externally. Better accessibility to these transparency reports by users may therefore act towards improving insights and confidence in platform moderation abilities (Urman & Makhortykh., 2023). However, the reports also indicate that automated systems fail to capture all SH/S content, leaving some dependence on user reporting. Additionally, automated moderation has faced criticism in relation to mental health content, whereby it is not precise enough to understand the nuances in content type or intent (Milton et al., 2023). These limitations may reduce the perceived value of transparency reports in reassuring users, and undermine trust in the consistency of moderation systems.

4.1 Limitations

Using data from the DELVE research study (Haime et al., 2024) presented a valuable opportunity to explore user experiences of moderation, moderators, and moderating of SH/S content. This study is particularly relevant in the context of ongoing changes in legislature, such as the implementation of the Online Safety Act (2023) in UK law and the Digital Services Act (2022) in Europe. Despite this, it is essential to acknowledge certain limitations in our approach. This study aimed to recruit participants who had engaged with SH/S content online. It was evident that participants encountered such content via several different platforms. Although this gave us good diversity, allowing for a broad understanding of user experiences, it made more nuanced platform-specific insights and implications difficult. Another limitation of this study was that participants who undertook moderator roles, were doing so voluntarily at the community-level and therefore we lacked insights from platform-level moderators, of moderating SH/S content. Additionally, where participants considered lived experience vs. mental health professional moderators in their accounts, they did not explicitly consider mental health professionals with lived experience of mental illness, potentially overlooking a factor affecting moderator effectiveness. Finally, we also struggled to recruit any participants who publicly posted the most explicitly

harmful SH/S content (including graphic images/videos, and methods information) online. Exploring this perspective could offer valuable insights into participant perceptions regarding content reporting and removal, and how they respond to these actions.

4.2 Implications for Industry

Findings from this study should inform moderation practices by online industry leaders, with the following preliminary recommendations:

- Platforms and communities should recognise the challenges faced by mental health online community members in practicing self-moderation. Reliance on user-reporting should be considered insufficient for providing a safe environment.
- Platforms and communities should consider integrating support from mental health professionals into mental health online spaces. Specifically, lived-experience moderators should have access to training and supervision from such professionals.
- Platforms and communities should consider the ongoing well-being of moderators. In doing so, they have a responsibility to ensure moderators have the capacity to recognise their own mental health concerns. This includes re-evaluating current models to ensure adequate time and mechanisms are allocated to moderators to address their mental and physical health needs.
- Platforms and communities should adopt a balanced approach to moderation of SH/S content, prioritising the safety of the overall userbase whilst also considering the wellbeing of individuals posting content. Platforms should allow for nuance in moderator decision-making processes, and prioritise the use of clear language and messaging to users around decisions made. They should also consider provision of postvention support to users following content removal or account banning.
- Platforms and communities should consider using an open dialogue approach with users in their moderation practices. This will enable users to stay informed of the processes their content or account may be undergoing and ensure transparency from platforms and communities regarding their moderation policies and guidelines.

4.3 Implications for Policymakers

These preliminary recommendations may also provide valuable insights to policymakers, such as Ofcom, in their role in governing digital industry leaders and for government legislators in relation to the Online Safety Act (2023). The current evidence base for moderation practices by platforms intentionally or unintentionally hosting SH/S content remains limited, leaving the effectiveness and impact of existing regulations on platform users, unclear.

Additionally, whilst the Online Safety Act (2023) represents steps towards online safety by promoting proactive moderation, including requiring platforms to remove illegal content, protecting children from harmful content (e.g., SH/S and eating dis-

order content), ensuring clear reporting procedures, employing effective moderation, and mandating transparency reports, it faces several challenges.

The language used in the Online Safety Act (2023), such as 'harmful content that celebrates, glorifies, or instructs self-injurious behaviour', is subjective, which could maintain existing inconsistencies in how harmful content is interpretated by users and how guidelines are applied across platforms. Platforms are also required to conduct internal risk assessments to evaluate harmful content risks (though evidence must be submitted). This puts significant responsibility onto platforms, and the recommendation to conduct these annually (at a minimum) may not account for rapid changes in the online landscape. Additionally, it is unclear how these assessments will capture individual differences between platforms. Further, banning and removing content without offering alternative online mental health spaces, or sufficient resources and support mechanisms, could potentially increase self-censorship, encourage 'algospeak' (Gerrard, 2018), or drive users to less moderated platforms. Finally, the Act is primarily designed to target larger platforms, which may result in smaller platforms becoming risky, less moderated spaces, leading to inconsistent levels of online protection for vulnerable users.

Policymakers should therefore support the preliminary recommendations outlined above for industry leaders, promoting a transparent and sensitive approach to moderation of SH/S content online, and advocating for strategies that prioritise user safety, such as requiring mental health professional supervision of moderators and postvention support to reporters and reportees of SH/S content. Policymakers must also recognise the ongoing challenges online platforms and communities face in moderation of mental health spaces, and work alongside academic researchers, mental health experts, industry organisations and people with lived experience of mental health needs, to develop collaborative approaches and guidelines to tackle current and developing challenges in this area.

4.4 Future Research and Conclusion

Findings from this study should encourage further research into user experiences of moderation, moderators, and moderating in online mental health spaces. Exploring the perspectives of platform-level moderators overseeing SH/S and other mental health related content will also significantly aid our understanding of moderation and the challenges moderators face in these environments. Additionally, measuring the effectiveness of moderation practices including our preliminary recommendations on users, particularly those engaging with SH/S content, is necessary to establish an evidence base for future decisions around online safety.

Overall, this study highlights that current approaches to SH/S content moderation by online platforms are inadequate, and the OSA does not fully address these gaps. There is a need for proactive, platform-level strategies, that incorporate and evaluate our preliminary recommendations. Policymakers should prioritise collaborative approaches, working alongside experts, users, and industry leaders to develop effective methods for online moderation of mental health related content and spaces.

Acknowledgements We express gratitude to the DELVE participants for their contribution to this work. Appreciation also extends to networks that facilitated recruitment, including TellMi, Samaritans, SMaRteN, McPin Foundation, and MQ Mental Health. We also acknowledge the contributions of our colleagues Jane Derges and Rachel Cohen, for their involvement in the DELVE study–s conception. We would also like to acknowledge our funders: National Institute for Health and Care Research Bristol Biomedical Research Centre and Samaritans, UK.

Author Contribution Zoë Haime: Project Administration, Investigation, Methodology, Formal Analysis, Writing (Original Draft), Writing (Review & Editing), Visualisation; Laura Kennedy: Investigation, Validation, Writing (Review & Editing); Lydia Grace: Conceptualisation, Writing (Review & Editing); Lucy Biddle: Funding Acquisition, Project Administration, Conceptualisation, Investigation, Methodology, Formal Analysis, Supervision, Writing (Review & Editing), Validation.

Funding This study was supported by the National Institute for Health and Care Research Bristol Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. Original data collected as part of the DELVE study was funded by Samaritans, UK.

Data Availability Anonymised transcript and questionnaire data will be stored on the University of Bristol's Research Data Service Facility. Researchers will be able to request access to non-identifiable data upon reasonable request. Access will be subject to a data access agreement and following approval from Dr Lucy Biddle and the University of Bristol Data Access Committee.

Declarations

Ethical Approval The study was approved by The University of Bristol Faculty of Health Sciences Ethics Committee (approval no. 117491).

Consent to Participate All participants provided written informed consent to participate in the study and for data collected in the study to be published (after anonymisation). Additionally, consent was provided for sharing anonymised data with other researchers.

Competing Interests We declare that there are no competing interests to disclose for any of the authors involved in the creation of this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Braun, V., & Clarke, V. (2012). Thematic analysis. *APA handbook of research methods in psychology, Vol. 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association. (APA handbooks in psychology).

Biddle., L., Derges, J., Mars, B., Heron, J., Donovan, J. L., Potokar, J., Piper, M., Wyllie, C., & Gunnell, D. (2016). Suicide and the internet: Changes in the accessibility of suicide-related information between 2007 and 2014. *Journal of Affective Disorders*, *190*, 370–375.

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (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 (CSCW '16). Association for Computer Machinery, New York, NY, USA, 1201–1213. https://doi.org/10.1145/2818048.2819963

Coulson, N. S., & Shaw, R. L. (2013). Nurturing health-related online support groups: Exploring the experiences of patient moderators. *Computers in Human Behavior*, *29*(4), 1695–1701.

[Digital Services Act] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 Oct. 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

Gardenier, A., van Est, R., & Royakkers, L. (2024). Technological citizenship in times of digitization: An integrative framework. *DISO*, *3*(21). https://doi.org/10.1007/s44206-024-00106-1

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12). https://doi.org/10.1177/1461444818776611

Google (2023). Google Wellness Standards for Sensitive Content Moderation (2023). Available online, accessed 04/01/2024: Wellness Standards Sensitive Content Moderation - Google (about.google).

Goswami, A. (2024). Recommendation system as a Social Determinant of Health. *DISO*, *3*, 28. https://doi.org/10.1007/s44206-024-00118-x

Halevy, A., Ferrer, C. C., Ma, H., Ozertem, U., Pantel, P., Saeidi, M., Silvestri, F., & Stoyanov, V. (2020). Preserving integrity in online social networks. *arXiv*:2009.10311.

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Trans Comput Hum Interact*, *26*, 31. https://doi.org/10.1145/3338243

Haime, Z., Kennedy, L., Grace, L., Cohen, R., Derges, J., & Biddle, L. (2024). The journey of engaging with web-based self-harm and suicide content: longitudinal qualitative study. *JMIR Infodemiology*, *4*, e47699. https://doi.org/10.2196/47699

Kiesler, S., Kraut, R., & Resnick, P. (2012). Regulating behavior in online communities. In *Building successful Online communities: Evidence-based social design*. MIT Press.

Milton, A., Ajmani, L., DeVito, M. A., & Chancellor, S. (2023). I See Me Here: Mental health content, community, and algorithmic curation on TikTok. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581489

Moreno, M. A., Ton, A., Selkie, E., & Evans, Y. (2016). Secret society 123: Understanding the language of self-harm on Instagram. *Journal of Adolescent Health*, *58*(1), 78–84.

Online Safety Act (2023). Parliament: House of Commons. Bill no. HL Bill 87(Rev). London.

Perry, A., Lamont-Mills, A., du Preez, J., & du Plessis, C. (2022). I want to be stepping in more - professional online forum moderators' experiences of supporting individuals in a suicide Crisis. *Frontiers in Psychiatry*, *13*, 863509. https://doi.org/10.3389/fpsyt.2022.863509

Roberts, S. T. (2019). *Behind the screen*. Yale University Press.

Saha, K., Ernala, S. K., Dutta, S., Sharma, E., & De Choudhury, M. (2020). Understanding moderation in online mental health communities. In G. Meiselwitz (Ed.), *Social computing and social media: Participation, user experience, consumer experience, and applications of social computing*. HCII 2020 (Vol. 12195). Springer. Lecture Notes in Computer Sciencehttps://doi.org/10.1007/978-3-030-49576-3_7

Smith, H., & Cipolli, W. (2022). The Instagram/Facebook ban on graphic self-harm imagery: A sentiment analysis and topic modeling approach. *Policy & Internet*, *14*(1), 170–185.

Spence, R., Bifulco, A., Bradbury, P., Martellozzo, E., & DeMarco, J. (2023). Content Moderator Mental Health, secondary trauma, and Well-being: A cross-sectional study. *Cyberpsychology Behavior and Social Networking*. https://doi.org/10.1089/cyber.2023.0298

Steiger, M., Bharucha, T. J., Venkatagiri, S., Ridel, M. J., & Lease, M. (2021). The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 341, 1–14. https://doi.org/10.1145/3411764.3445092

Stoilova, M., Edwards, C., Kostyrka-Allchorne, K., Livingstone, S., & Sonuga-Barke, E. (2021). Adolescents' mental health vulnerabilities and the experience and impact of digital technologies: A multimethod pilot study. *London School of Economics and Political Science and King's College London*.

Urman, A., & Makhortykh, M. (2023). How transparent are transparency reports? A comparative analysis of transparency reporting across online platforms. *Telecommunications Policy*, 47.
York, J. C., & Zuckerman, E. (2019). Moderating the public sphere. *Human Rights in the age of Platforms*, *137*, 143.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Zoë Haime[1,2] · Laura Kennedy[3] · Lydia Grace[3] · Lucy Biddle[1,2,4]

✉ Zoë Haime
zoe.haime@bristol.ac.uk

[1] NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust and University of Bristol, Bristol, UK

[2] Department of Population Health Sciences, University of Bristol, Canynge Hall, Bristol, UK

[3] Samaritans, Ewell, UK

[4] The National Institute for Health Research Applied Research Collaboration West, University Hospitals Bristol and Weston National Health Service Foundation Trust, Bristol, UK