

1 **Full Title: Experiences of moderation, moderators, and moderating by online users who**
2 **engage with self-harm and suicide content.**

3 **Short Title: Experiences of moderation in self-harm/suicide online spaces.**

4 **Zoë Haime^{1,2*}, Laura Kennedy^{3,4}, Lydia Grace³, Lucy Biddle^{1,2}.**

5 ¹ NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS
6 Foundation Trust and University of Bristol.

7 ² Department of Population Health Sciences, University of Bristol, Bristol, UK.

8 ³ Samaritans, Ewell, UK.

9 ⁴ Current Role: ESRC Centre for Society and Mental Health, Kings College London, UK.

10 * Correspondence to:

11 Zoë Haime

12 Department of Population Health Sciences,

13 Canynge Hall

14 Bristol, UK

15 zoe.haime@bristol.ac.uk

16

17

18

19

20

21

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

22 **Abstract**

23 Online mental health spaces require effective content moderation for safety. Whilst policies
24 acknowledge the need for proactive practices and moderator support, expectations and
25 experiences of internet users engaging with self-harm and suicide content online remain
26 unclear. Therefore, this study aimed to explore participant accounts of moderation,
27 moderators and moderating when engaging online with self-harm/suicide (SH/S) related
28 content.

29 Participants in the DELVE study were interviewed about their experiences with SH/S content
30 online. N=14 participants were recruited to interview at baseline, with n=8 completing the
31 3-month follow-up, and n=7 the 6 month follow-up. Participants were also asked to
32 complete daily diaries of their online use between interviews. Thematic analysis, with
33 deductive coding informed by interview questions, was used to explore perspectives on
34 moderation, moderators and moderating from interview transcripts and diary entries.

35 Three key themes were identified: 'content reporting behaviour', exploring factors
36 influencing decisions to report SH/S content; 'perceptions of having content blocked',
37 exploring participant experiences and speculative accounts of SH/S content moderation;
38 and 'content moderation and moderators', examining participant views on moderation
39 approaches, their own experiences of moderating, and insights for future moderation
40 improvements.

41 This study revealed challenges in moderating SH/S content online, and highlighted
42 inadequacies associated with current procedures. Participants struggled to self-moderate
43 online SH/S spaces, showing the need for proactive platform-level strategies. Additionally,
44 whilst the lived experience of moderators was valued by participants, associated risks

45 emphasised the need for supportive measures. Policymakers and industry leaders should
46 prioritise transparent and consistent moderation practice.

47 **Keywords:**

48 **Moderation, Online Harm, Thematic Analysis, Suicide, Self-Harm, Qualitative**

49 **Author Summary**

50 In today's digital world, ensuring the safety of online mental health spaces is vital. Yet,
51 there's still a lot we don't understand about how people experience moderation,
52 moderators, and moderating in self-harm and suicide online spaces. Our study set out to
53 change that by talking to 14 individuals who engage with this content online. Through
54 interviews and diaries, we learned more about their experiences with platform and online
55 community moderation.

56 Our findings showed some important things. Firstly, individuals with declining mental health
57 struggled to use tools that might keep them safe, like reporting content. This emphasised
58 the need for effective moderation in online mental health spaces, to prevent harm.

59 Secondly, unclear communication and inconsistent moderation practices lead to confusion
60 and frustration amongst users who reported content, or had their own content moderated.

61 Improving transparency and consistency will enhance user experiences of moderation
62 online. Lastly, users encouraged the involvement of mental health professionals into online
63 moderating teams, suggesting platforms and online communities should provide training
64 and supervision from professionals to their moderation staff. These findings support our
65 recommendations for ongoing changes to moderation procedures across online platforms.

66

67 **1. Introduction**

68 As online platforms facilitate connections between individuals on a global scale, it becomes
69 crucial to implement practices that maintain safe environments through content
70 moderation [1, 2]. Platform-level content moderation is used to regulate user-generated
71 content across entire platforms, and can involve automated tools, algorithms, and human
72 moderators working to apply overarching policies and standards. Alternatively, community-
73 level content moderation involves active participation of a platforms' users, through
74 reporting or flagging content that they find inappropriate, or against the rules of specific
75 spaces or communities within platforms. These moderation actions at both platform and
76 community level may be particularly relevant in controlling engagement with sensitive,
77 harmful or illegal material, including self-harm and suicide content and discourse around it.

78 ***1.1. Moderation Effectiveness***

79 To regulate online community spaces, moderators play a vital role in screening content,
80 identifying problematic users, and enforcing rules [3]. The mechanisms employed, such as
81 blocking or removal of content, and semi- or permanent-banning of users are thought to
82 ensure the availability of high-quality content, whilst limiting the presence of harmful
83 material [4]. However, experiences and outcomes of these moderating actions can vary for
84 online users. For example, when Facebook and Instagram introduced a total ban of graphic
85 self-harm imagery, sentiment analysis revealed an increase in anger, anticipation, and
86 sadness in the associated Twitter [renamed 'X'] discourse [5].

87 Complexity in moderation decision-making is particularly evident on platforms or online
88 environments focused on mental health. Moderation in these spaces must balance the
89 responsibility of protecting users from triggering content, with the provision of space for

90 social support and recover [6]. The effectiveness of current moderation practices, including
91 outright content bans in mental health spaces, have been further questioned. Multiple
92 studies have investigated how users seek out potentially harmful pro-eating disorder
93 content online and evade moderation systems by posting content using no text, or
94 alternative hashtags [7, 8]. This practice is also prevalent amongst self-injury related posts
95 on Instagram, where the use of ambiguous and unrelated hashtags leads to graphic self-
96 harm imagery without proper content warnings [9]. Additionally, research shows that even
97 where platforms proactively search for and remove harmful content, this can be inefficient,
98 as content reappears on the same or alternative platforms [10, 11].

99 **1.2. Who Should Moderate Content?**

100 There are also concerns regarding the allocation of responsibility for moderating online
101 spaces. Platform-level moderation usually relies on paid humans, although these
102 mechanisms are sometimes automated. However, within peer-driven communities, users
103 themselves usually moderate content, often in voluntary roles. While several benefits to the
104 moderator role have been reported, including feelings of altruism, having a sense of
105 purpose [12] and – in the case of lived-experience moderators – receiving validation of their
106 own experiences [13], concerns have also been raised for both community and platform
107 level moderators' about the impact on their own mental health [6, 12].

108 A recent study by Spence et al [14], found 40.8% of 213 online platform-level content
109 moderators were exposed to distressing content daily, and that moderators showed a dose-
110 dependent relationship between frequency of exposure to distressing content and
111 psychological distress and secondary trauma. However, their findings also revealed, that
112 after accounting for work factors in the analysis, including access to supportive colleagues

113 and feedback about their work, the relationships between exposure and psychological
114 distress and secondary trauma failed to remain significant, indicating that a supportive work
115 environment may ameliorate negative effects.

116 Some platforms and sites, such as Google and Facebook, have begun to implement
117 strategies to address the risks to their content moderators, such as training on working with
118 sensitive content, recommended access to both individual and group counselling services,
119 and suggested 'wellness breaks' [15, 16]. However, the effectiveness of these interventions
120 for moderator wellbeing, remains unclear. Additionally, whilst large platforms are able to
121 dedicate resources to moderator care, there are less viable options available to
122 communities hosted on those platforms, or to smaller individual platforms.

123 Even within larger platforms, it may also be challenging to ensure uniform care across all
124 moderators, due to varying work environments and individual needs. For instance, diversity
125 in preferences of moderators was shown by Saha et al [16]. Some moderators desired
126 support from mental health experts and welcomed the idea of having trained professionals
127 within the moderation team. However, other moderators emphasised that these spaces
128 should not provide medical advice to users, and that moderators with lived experience of
129 mental health problems can create the supportive community users are looking for, without
130 additional help. This highlights the differing needs amongst moderators in mental health
131 spaces and emphasises the need for further understanding of individual moderator
132 narratives to optimise their experiences.

133 ***1.3. The Online Safety Bill***

134 Policy and regulatory frameworks also play a significant role in the expectations and
135 responsibilities of online platforms in moderating content. In the United Kingdom, there has

136 been a growing emphasis on addressing online harms, through the recently enacted *Online*
137 *Safety Bill* [17]. Key points within this legislation include the need for platforms to remove
138 self-harm and suicide content, implement measures to minimise user exposure to harmful
139 content, uphold a duty of care towards users, and provide individuals with improved
140 mechanisms to report harmful content.

141 To date, limited research has been conducted exploring experiences of moderation by users
142 who engage with online mental health content or those who moderate it. Notably, existing
143 findings suggest that moderation is essential to mental health online spaces [6] but that
144 current approaches may be ineffective, and, in some cases increase vulnerability to harm for
145 users and moderators. For instance, content posted with limited information in order to
146 avoid automatic removal systems, increases the risk to users of unexpectedly encountering
147 potentially triggering content [7], and moderators may experience emotional distress and
148 burnout [13]. A noticeable research gap exists in understanding user perspectives on the
149 moderation of SH/S related content online, particularly user experiences of having their own
150 SH/S related content moderated (reported or removed), reporting others' SH/S content, or
151 seeing others' SH/S content being removed. Understanding the first-hand perspective of
152 users engaging in mental health online spaces is essential for informing effective,
153 responsible, and user-centric moderation strategies in the digital space.

154

155

156 **1.4. Aim of the study**

157 This study involved thematic analysis [18] of qualitative interview transcripts to gain insights
158 into participant experiences with moderation and moderators in the context of engaging
159 with SH/S content online. By examining these perspectives, this research aimed to deepen
160 our understanding of how moderation practices impact user experiences and consider how
161 this can inform the development of industry guidance.

162 2. Results

163 2.1. Participants

164 Fourteen participants took part. The sample had diversity in terms of age and ethnicity, and
165 relatively good representation of gender (Table 1). All participants completed a baseline
166 interview, eight participants completed midpoint and seven completed endpoint interviews
167 (see [19]). Participants returned monthly diaries. Due to participant dropout, overall 44
168 diaries were anticipated, with a total of 31 (70.5%) diaries actually returned. All interview
169 transcripts and diary data were used in this analysis.

170 **Table 1.**

171 *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)

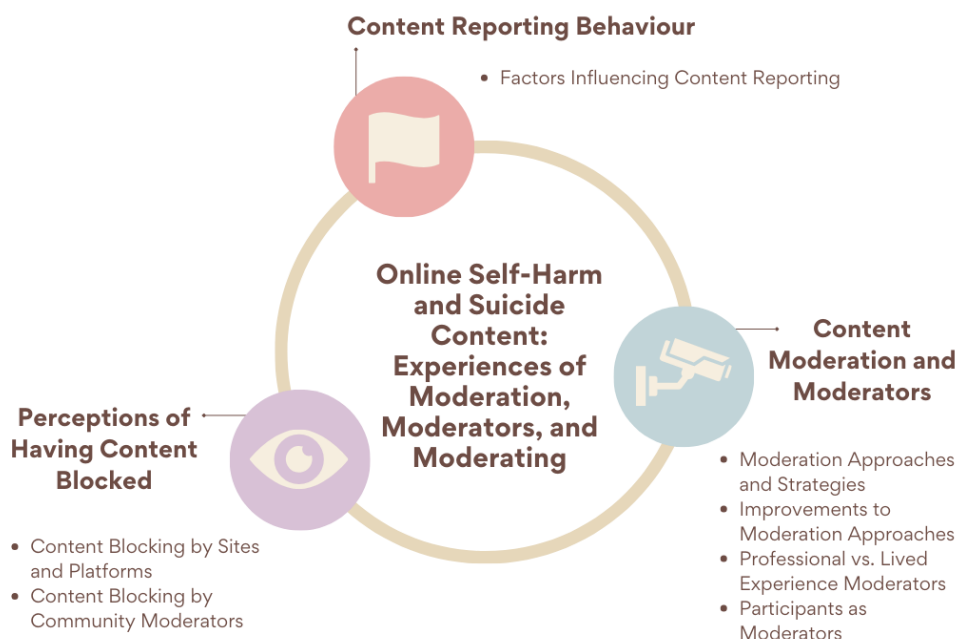
18-24 Years	7 (50.0)
25-35 Years	0 (0.0)
36-45 Years	4 (28.6)
46-54 Years	2 (14.2)
55+ Years	0 (0.0)

172

173 **2.2. Thematic Analysis**

174 Results are visually depicted in Figure 1. by theme and subtheme. These are explored in

175 more detail below.



176

177 **2.2.1. Content Reporting Behaviour**

178 **2.2.1.1. Factors Influencing Content Reporting**

179 Several participants described instances where they didn't report SH/S content. One

180 participant described how their non-reporting was influenced by a lack of awareness about

181 how and where to report on certain platforms/sites (IDH). However, participants also shared

182 that they did not consciously consider reporting SH/S content when online (IDC, IDH, IDM,

183 133), due to their poor mental state. In fact, during challenging times, users described
184 becoming preoccupied with this material or deliberately seeking it out, thereby prioritising
185 the perceived benefits of content over reporting it:

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

189 A couple of participants described how decisions to report SH/S content were multifaceted,
190 one stating how the choice to report would change depending upon their mental health
191 state:

192 'Part of me thinks they should be reported and another part of me thinks, well,
193 freedom of speech and they should stay. So I most probably feel that they should
194 stay, on balance. However, I think that's one of those things that could change and
195 fluctuate depending upon my own mental health.' (IDC)

196 Participants who recalled reporting content described no hesitation in reporting general
197 online content considered morally wrong, such as racist comments or images. However,
198 decisions to report SH/S content were often more selective. One user (IDD) reflected on the
199 emotional impact of content on a pro-suicide forum affecting their decision: 'Cause like if it
200 does affect me then... I like reduce time on it or report it to one of the moderators for
201 example' (IDD). Another (IDM), described how their lack of knowledge about SH/S content
202 they came across made them refrain from reporting:

203 'Why didn't you report it [graphic self-harm image]?' (Interviewer)

204 'I didn't know much about it [graphic self-harm image] and so I decided maybe to
205 just not show it to anyone.' (IDM)

206 Considering the implications for users posting the SH material, this participant also
207 expressed reluctance to report images of SH without first understanding the posters'
208 circumstances and suggested having a conversation with them may be more appropriate
209 than reporting:

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

214 Despite concern for content posters' wellbeing, and attempts not to stigmatise their
215 experience, there was general agreement amongst the other participants that images or
216 videos of SH/S were largely inappropriate and justified reporting, sometimes due to their
217 influence over participants' own SH/S behaviours:

218 'I can get it if someone is saying oh, I'm really suicidal, I won't report that, because
219 maybe they need someone to talk to, if they are posting images on that though, you
220 can't be doing that... And also there are lots of videos of people trying to kill
221 themselves, they're on the internet, that's another thing you can spiral into that,
222 there are sites just full of people dying...I think it was on Twitter, I ended up on a
223 video and then it went to another video and then I ended up on a site and then
224 yeah...' (IDK)

225 '...Ok and how did that make you feel sort of seeing that play out?' (Interviewer)

226 'I think I took an overdose afterwards' (IDK)

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

231 One participant reflected on how following predetermined methods for reporting, like using
232 checkboxes, allowed the process to be less biased, and suggested an attempt to detach
233 themselves from the feelings of guilt associated with reporting an online peer:

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

237 When asked, those who had reported online content did describe it as a straightforward
238 process in practical terms, with some sites having specific options for recording if the
239 content was related to SH/S. However, one participant (IDE) acknowledged how it would be
240 impractical to report all content that was offensive or inappropriate, and therefore they felt
241 they had to make careful decisions about what they reported and when, although they did
242 not detail how they made those choices, other than to note they were more likely to report
243 users with significant follower counts: 'Yeah, when I see something that I don't like, like self-
244 harm and they've got lots and lots of followers then I do report stuff' (IDE).

245 Some of the participants who previously had not reported content, mentioned during
246 interviews that they would now report SH/S content they came across online, using the
247 platform guidelines to direct them (IDH, IDJ, IDC). It is unclear what motivated this change,

248 though there was some indication it may be explained by involvement in the research itself.

249 For instance, IDE described increased awareness of content that may be harmful, and being

250 more conscious of its effects:

251 'My uses have changed massively because it's really [I've] been more aware, but

252 also, I'm more conscious of it. So when you are a user you just look at it and think,

253 'Okay.' But when you do something like [taking part in this research study] you think,

254 'Hang on, actually there's another dimension to it and is there harm or should I

255 report?' So you become a social media – I don't know if it's an influencer – but you

256 become a sort of [an] active or civic user.' (IDE)

257 Some participants also described the consequences of reporting other users' content. While

258 a couple described content being taken down (IDK, IDI), or the user being 'kicked off' (IDD) a

259 group, others commented on specific site/platform actions, stating Facebook do 'nothing'

260 (IDB) when reporting nuisance pages, or that reporting on Tumblr felt 'completely

261 pointless...because nothing would happen' (IDF).

262 IDI acknowledged that although the automated process of reporting using Instagram

263 options made it relatively easy to do, it left no room for nuance in describing the content: 'I

264 don't think there's much of an opportunity to fully explain anything or anything like that'

265 (IDI). Additionally, despite reported SH/S content being removed on Instagram, users were

266 able to continue posting without consequence:

267 '[they can] carry on posting as normal anyway. So it just means that they can post it

268 again' (IDI)

269 **2.2.2. Perceptions of Having Content Blocked**

270 2.2.2.1. Content Blocking by Sites and Platforms

271 Several participants, (IDB, IDN, IDL, IDK, and IDG) were able to describe their experiences of
272 having content reported online, including content expressing suicidal ideation (IDB, IDK),
273 privately posted images of SH (IDL), and blogs about depression and self-harm (IDN).
274 Participants described receiving differing levels of information from sites when their content
275 had been blocked, with some feeling that the rules felt unclear:

276 'I have like one or two blocked [blogposts], but I don't know why. I don't know which
277 words [are] the restriction things. I just can't post it out and it remains in the draft.
278 But I think I'm just writing a normal thing, but maybe they have like a key word ban'
279 (IDN).

280 IDK, who had once posted stating they were 'struggling' on Twitter [X], had this reported by
281 another user and criticised the platform's 'one-size-fits' all response:

282 '[It's] a bit pointless to be honest, like suicide hotlines are not going to solve your issues,
283 I think that's the problem... but it's [why they are struggling] about I can't afford this, I
284 can't afford my house, like what am I going to do, so I found it a bit hard but thanks for
285 sending me this, do you know what I mean? (IDK)

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

290 '...I think it's like under [the] notification page they'll just say, your post has been
291 removed because of a violation. And yeah I think it's quite frustrating because I don't

292 know what I've written so I can forget what it was about...because I mean even if
293 they said that, I want to remember and recall how I was feeling and the tough times
294 I'd been through. But they don't give you any warning at all, they just take it' (IDL).

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

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

299 Interestingly, one participant who had content blocked on Twitter [X] due to the use of bad
300 language found the situation 'quite funny to be honest' (IDK), perhaps highlighting the
301 difference in emotional response to content that is less personally significant.

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

305 **2.2.2.2. Content Blocking by Community Moderators**

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

311 Two participants shared experiences of having content removed from these spaces. Both
312 were left feeling confused and frustrated due to inconsistent and unclear moderation
313 policies. For instance, one had (IDG) posted 'what is the point?' on a forum, which led to

314 moderators removing their post and contacting them with resources of support. The
315 participant expressed feeling ‘... just like a criminal. Well not a criminal, but like I’ve really
316 done something wrong.’ (IDG). They also seemed surprised to see other content on the site
317 they deemed more inappropriate hadn’t been removed.

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

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

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

333 ‘...she’s (other user) like, “it’s so dangerous, you’re going to kill someone”, I’m like,
334 so are you, like it says in my post I’m suicidal, is that (users’ comment to her) really
335 helpful. And I started to get really wound up, and then the people that were
336 managing the page took my comment off and blocked me from commenting on it

337 anymore but allowed for her to continue which just drove me insane because I felt
338 that was so unfair.” (IDB)

339 **2.2.3. Content Moderation and Moderators**

340 **2.2.3.1. Moderation Approaches and Strategies**

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

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

348 ‘I don’t feel that it [online SH/S spaces] should be [taken] away because if it’s
349 moderated correctly then it’s not doing anybody any particular harm.’ (IDD)

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

355 ‘I tend to stick to [self-harm support Facebook group] because I know that it’s heavily
356 moderated’ (IDB)

357 ‘... it’s clear from the onset. They’re completely honest and they say this is a
358 moderated Facebook page, so you know where you stand...A couple of days ago,

359 somebody said about direct messaging and... very soon, a moderator came on and
360 just said, 'Can I just remind you?', and he phrased it really nicely.' (IDC).

361 Additionally, IDB expressed an advantage of using the self-harm support Facebook group
362 was its rapid content moderation, which reassured them it was safe:

363 'The moderation like I say this is why I tend to use [self-harm support Facebook
364 group] more than anything else because... it's so heavily moderated and monitored
365 we don't really find it [community argument posts]. And if anything like that does
366 crop up on [self-harm support Facebook group] it's gone within minutes you know.

367 An hour is long for it to have still been on there.' – IDB

368 However, it is important to note that these perspectives may have been subject to bias
369 within the sample. Notably, IDA and IDB who regularly engaged with the self-harm support
370 Facebook group, also moderated for them in a voluntary capacity, which may have
371 influenced their views on effective approaches.

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

378 'On Twitter [X], if you've reported... like graphic images or anything like that, it will
379 be blocked, but TikTok and Facebook they're very lenient about stuff.' (IDK).

380 **2.2.3.2. Improvements to Moderation Approaches**

381 Participants identified areas of moderation that they would like to see improved across
382 platforms. One common concern for participants was the lack of transparency and
383 consistency in what content would be blocked or removed:

384 'I think in terms of restrictions for social media overall, I personally don't feel they're
385 very open about exactly how they moderate posts. So I think that it's difficult to
386 understand what their thought processes [are] behind blocking posts or restricting
387 posts because they don't make it clear.' (IDI)

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

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

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

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

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

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

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

418 'If it was my child, I wouldn't want them accessing that content but a lot of that is
419 because I think that children and the internet are a difficult match anyway because
420 of the way that the internet is designed to keep you hooked and put so much worth
421 in algorithmic likes and comments....' (IDF)

422 'I feel like if teenagers or young people are gonna go on it [SH/S spaces online] then
423 they need somewhere that will moderate it by professionals like I had when I were a
424 teenager, rather than just being on Facebook or Instagram for example.' (IDD)

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

428 ‘...because on Facebook if you’re an admin of a group, because I do it on my group...
429 you actually have to have admin approval before you approve it.’ (IDG)

430 ‘Before other people see it’ (Interviewer)

431 ‘Before it goes on the group. Now this [mental health Facebook] group doesn’t have
432 that and this is the thing that was worrying’ (IDG)

433 **2.2.3.3. Professional vs. Lived Experience Moderators**

434 Some participants went on to describe the moderator role in online spaces and whether this
435 should be undertaken by trained professionals or by those with lived experience. Two of
436 those participants expressed the importance of moderators with personal experience of
437 SH/S:

438 ‘That’s the great thing about [Self-Harm support Facebook group is] that it’s not run
439 by people that don’t get it.’ (IDB)

440 ‘I think having people have a lived experience doing the moderation is more
441 beneficial.’ (IDC)

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

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

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

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

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

458 'I think there also needs to be somewhere to go to keep themselves in check
459 because I think you can get just indoctrinated into what you're doing, [with]
460 everybody else on the site [having] the same problem. So whether it be something
461 where moderators are changed regularly I don't know, something like that.' (IDH)

462 'it feels like you need maybe somebody [to take] a step back. It may be that [Self-
463 Harm Support Group] needs a professional that's going to regularly contact the
464 people that are moderating and the people that are running these things to make
465 sure we're okay.' (IDB)

466 **2.2.3.4. Participants as Moderators**

467 Three (IDA, IDB, IDI) had become moderators or administrators on the platforms where they
468 accessed SH/S content. The decision to take on this role was driven by a hope to help
469 others, and ensure rules within the online spaces were adhered to, as well as an altruistic
470 way of giving back to the community:

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

475 These participants reflected on the significant responsibility associated with moderating
476 online SH/S content. IDI shared their personal recognition of the need to disengage from a
477 moderation/administrator role when their own mental health declined. They expressed
478 concern about their ability to effectively address someone else's issues in such
479 circumstances, as well as the potential negative impact of engaging with another users'
480 distressing content:

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

486 However, IDB seemed less able to manage their behaviour in this way. They recalled an
487 incident where they encountered a video in the group they moderated that presented a SH
488 method that was novel to them. This resulted in them subsequently acquiring the
489 equipment for the method, and then self-harming:

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

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

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

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

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

511 **3. Discussion**

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

516 Many online environments rely on user reporting to ensure moderators can successfully
517 enforce community rules [3]. However, participants in this study revealed an inherent
518 inability to undertake community moderation through reporting SH/S content, particularly
519 during times of poorer mental health. Participants revealed these struggles to report
520 content stemmed from a fear of potentially stigmatising other users, and a desire to access
521 the content themselves, despite recognising the potential dangers of the material. This
522 emphasises the complexity of decision-making processes for vulnerable users in online
523 spaces and challenges the traditional methods of content reporting strategies employed by
524 platforms.

525 Although participants described difficulties in reporting SH/S content, several users did take
526 on moderating roles in these online environments, with largely altruistic intentions.

527 However, they also found their ability to moderate fluctuated alongside their own mental
528 health, a concern recognised in previous research [6, 12]. While one user was able to
529 identify their mental health declines and take action to protect themselves and others' by
530 temporarily withdrawing from their moderating/administrative duties, another was unable
531 to safeguard themselves and became at risk of harm whilst undertaking their moderating
532 role.

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

535 content. Additionally, poor transparency in communication of moderation methods and
536 results not only led to disappointment, frustration, and a perceived loss of control amongst
537 participants, but also added a speculative element to many participant accounts. Several
538 also criticised ‘catch-all’ reactive policies for SH/S material, where content could be
539 immediately removed, or users banned without warning. These punitive moderation
540 methods were viewed as non-empathetic, harmful, and potentially stigmatising, consistent
541 with user experiences of Facebook and Instagram’s ban of graphic self-harm imagery [5].

542 Participants largely agreed that lived-experience moderation was important, due to the
543 moderators’ understanding of user experiences. However, many also emphasised the
544 potential difficulties and harmful consequences a lived-experience moderator may
545 experience by engaging with content and users in these spaces. Therefore, alike to
546 participants in previous research [6] and recommendations from Facebook [16], some
547 participants encouraged the involvement of mental health professionals in providing
548 support through supervision or training, to lived-experience moderators.

549 **3.1. Limitations**

550 Using data from the DELVE research study (Haime et al., 2023) presented a valuable
551 opportunity to explore user experiences of moderation, moderators, and moderating of
552 SH/S content. This study is particularly relevant in the context of ongoing changes in
553 legislature, such as the implementation of the Online Safety Bill [17] in UK law and the
554 Digital Services Act [20] in Europe. Despite this, it is essential to acknowledge certain
555 limitations in our approach. This study aimed to recruit participants who had engaged with
556 SH/S content online. It was evident that participants encountered such content via several
557 different platforms. Although this gave us good diversity, allowing for a broad

558 understanding of user experiences, it made more nuanced platform-specific insights and
559 implications difficult. Another limitation of this study was that participants who undertook
560 moderator roles, were doing so voluntarily at the community-level and therefore we lacked
561 insights from platform-level moderators, of moderating SH/S content. Additionally, where
562 participants considered lived experience vs. mental health professional moderators in their
563 accounts, they did not explicitly consider mental health professionals with lived experience
564 of mental illness, potentially overlooking a factor affecting moderator effectiveness. Finally,
565 we also struggled to recruit any participants who publicly posted the most explicitly harmful
566 SH/S content (including graphic images/videos, and methods information) online. Exploring
567 this perspective could offer valuable insights into participant perceptions regarding content
568 reporting and removal, and how they respond to these actions.

569 **3.2. Conclusions**

570 Findings from this study should inform moderation practices by online industry leaders, with
571 the following considerations:

- 572 • Platforms and communities should recognise the challenges faced by mental health
573 online community members in practicing self-moderation. Reliance on user-
574 reporting should be considered insufficient for providing a safe environment.
- 575 • Platforms and communities should consider integrating support from mental health
576 professionals into mental health online spaces. Specifically, lived-experience
577 moderators should have access to training and supervision from such professionals.
- 578 • Platforms and communities should consider the ongoing well-being of moderators.
579 In doing so, they have a responsibility to ensure moderators have the capacity to
580 recognise their own mental health concerns. This includes re-evaluating current

581 models to ensure adequate time and mechanisms are allocated to moderators to
582 address their mental and physical health needs.

- 583 • Platforms and communities should adopt a balanced approach to moderation of
584 SH/S content, prioritising the safety of the overall userbase whilst also considering
585 the wellbeing of individuals posting content. Platforms should allow for nuance in
586 moderator decision-making processes, and prioritise the use of clear language and
587 messaging to users around decisions made. They should also consider provision of
588 postvention support to users following content removal or account banning.
- 589 • Platforms and communities should consider using an open dialogue approach with
590 users in their moderation practices. This will enable users to stay informed of the
591 processes their content or account may be undergoing and ensure transparency
592 from platforms and communities regarding their moderation policies and guidelines.

593 These considerations may also provide insight to policymakers, such as Ofcom, in their role
594 governing digital industry leaders. Policymakers should recognise the challenges online
595 platforms and communities face in moderation of mental health spaces, and advocate for
596 strategies that prioritise user safety. Policymakers should support the initiatives outlined
597 above for industry leaders, promoting a transparent and sensitive approach to moderation
598 of SH/S content online.

599 Findings from this study should also encourage further research into user experiences of
600 moderation, moderators, and moderating in online mental health spaces. Exploring the
601 perspectives of platform-level moderators overseeing SH/S content will further our
602 understanding of moderation and moderators in these environments.

603 **4. Method**

604 **4.1. Procedure**

605 Data collected as part of the DELVE study [19] were analysed. Participants were interviewed
606 at three time points (baseline, midpoint, endpoint) over a six-month study period about
607 their engagement with SH/S content online. Interviews were semi-structured, with
608 researchers employing an open-ended approach using probing to obtain detailed,
609 participant-led accounts. In addition, participants were asked to maintain daily research
610 diaries during the intervening periods. Participants were asked to reflect on their
611 experiences of engaging with content, having content blocked or reported, or reporting
612 others' content during their interviews and in their diaries. In the interview, 'moderation'
613 was an additional topic area where participants were asked to share their thoughts on
614 moderators they had encountered online, their experiences of others' content being
615 moderated, and of their own SH/S content being moderated.

616 The study was approved by The University of Bristol Faculty of Health Sciences Ethics
617 Committee (approval no. 117491). All participants provided written informed consent prior
618 to participation.

619 **4.2. Participants**

620 Participants were 16 years and over, English-speaking, and online users who had engaged
621 with self-harm or suicide content. Recruitment methods (outlined in [19]) involved multiple
622 channels such as social media platforms, a mental health app targeting young people, and
623 charity websites and newsletters. Participants responded to study adverts and were
624 assessed for eligibility before being sent the study information sheet. Purposive sampling
625 was used to promote diversity in gender, ethnicity, and age of sampled participants.

626 **4.3. Analysis**

627 Thematic analysis [18] with deductive coding informed by interview questions was used,
628 with several steps:

- 629 1. Interviews were transcribed, and researchers read and familiarised themselves with
630 both the participant transcripts and their diary data, to gain overall understanding.
- 631 2. ZH began coding the data. Codes were systematically assigned to quotes which
632 captured certain concepts, or ideas about moderation or moderators. Coding was
633 iterative, and codes were refined and revised in consultation with LB.
- 634 3. Similar codes were grouped together to form themes representing patterns in
635 concepts or ideas. Themes were then explored narratively in the context of the
636 research question.

637

638 **5. Availability of Data and Materials**

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

643 **6. Funding**

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

648 **7. Competing Interests**

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

651 **8. CRediT author statement**

652 **Zoë Haime:** Project Administration, Investigation, Methodology, Formal Analysis, Writing
653 (Original Draft), Writing (Review & Editing), Visualisation; **Lucy Biddle:** Funding Acquisition,
654 Project Administration, Conceptualisation, Investigation, Methodology, Formal Analysis,
655 Supervision, Writing (Review & Editing), Validation; **Laura Kennedy:** Investigation,
656 Validation, Writing (Review & Editing); **Lydia Grace:** Conceptualisation, Writing (Review &
657 Editing).

658

659 **9. Acknowledgements**

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

667 **References**

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

- 669 2. York, J. C., & Zuckerman, E. (2019). *Moderating the public sphere*. Human rights in
670 the age of platforms, 137, 143.
- 671 3. Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-Machine
672 Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM*
673 *Trans. Comput Hum Interact.* 26, 5, Article 31. <https://doi.org/10.1145/3338243>
- 674 4. Kiesler, S., Kraut, R., and Resnick. P. (2012). Regulating behavior in online
675 communities. In *Building Successful Online Communities: Evidence-Based Social*
676 *Design*. MIT Press.
- 677 5. Smith, H., & Cipolli, W. (2022). The Instagram/Facebook ban on graphic self-harm
678 imagery: A sentiment analysis and topic modeling approach. *Policy & Internet*, 14(1),
679 170-185.
- 680 6. Saha, K., Ernala, S. K., Dutta, S., Sharma, E., & De Choudhury, M. (2020).
681 *Understanding Moderation in Online Mental Health Communities*. In: Meiselwitz, G.
682 (eds) *Social Computing and Social Media. Participation, User Experience, Consumer*
683 *Experience, and Applications of Social Computing*. HCII 2020. Lecture Notes in
684 *Computer Science*, vol 12195. Springer, Cham. [https://doi.org/10.1007/978-3-030-](https://doi.org/10.1007/978-3-030-49576-3_7)
685 [49576-3_7](https://doi.org/10.1007/978-3-030-49576-3_7)
- 686 7. Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social
687 media. *New Media & Society*, 20(12). <https://doi.org/10.1177/1461444818776611>
- 688 8. Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016).
689 #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating
690 Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-*
691 *Supported Cooperative Work & Social Computing (CSCW '16)*. Association for

- 692 Computer Machinery, New York, NY, USA, 1201-1213.
- 693 <https://doi.org/10.1145/2818048.2819963>
- 694 9. Moreno, M. A., Ton, A., Selkie, E., & Evans, Y. (2016). Secret society 123:
695 Understanding the language of self-harm on Instagram. *Journal of Adolescent*
696 *Health, 58*(1), 78-84.
- 697 10. Stoilova, M., Edwards, C., Kostyrka-Allchorne, K., Livingstone, S., & Sonuga-Barke, E.
698 (2021) Adolescents' mental health vulnerabilities and the experience and impact of
699 digital technologies: A multimethod pilot study. London School of Economics and
700 Political Science and King's College London. Accessed online 13/02/2024:
701 [Stoilova et al 2021 Mental health digital technologies report.pdf \(lse.ac.uk\)](#)
- 702 11. Biddle, L., Derges, J., Mars, B., Heron, J., Donovan, J. L., Potokar, J., Piper, M., Wyllie,
703 C., & Gunnell, D. (2016). Suicide and the Internet: Changes in the accessibility of
704 suicide-related information between 2007 and 2014. *Journal of Affective Disorders,*
705 *190*, pp. 370-375.
- 706 12. Steiger, M., Bharucha, T. J., Venkatagiri, S., Ridel, M. J., & Lease, M. (2021). The
707 psychological well-being of content moderators – The emotional labor of commercial
708 moderation and avenues for improving support. In Proceedings of the 2021 CHI
709 Conference on Human Factors in Computing Systems (CHI '21). Association for
710 Computing Machinery, New York, NY, USA, Article 341, 1–14.
711 <https://doi.org/10.1145/3411764.3445092>
- 712 13. Coulson, N. S., & Shaw, R. L. (2013). Nurturing health-related online support groups:
713 Exploring the experiences of patient moderators. *Computers in Human*
714 *Behavior, 29*(4), 1695-1701.

- 715 14. Spence, R., Bifulco, A., Bradbury, P., Martellozzo, E., & DeMarco, J. (2023). Content
716 Moderator Mental Health, Secondary Trauma, and Well-being: A Cross-Sectional
717 Study. *Cyberpsychology, behavior, and social networking*. Doi:
718 <https://doi.org/10.1089/cyber.2023.0298>
- 719 15. Google (2023). Google Wellness Standards for Sensitive Content Moderation (2023).
720 Available online, accessed 04/01/2024: [Wellness Standards Sensitive Content](#)
721 [Moderation - Google \(about.google\)](#)
- 722 16. Halevy, A., Ferrer, C.C., Ma, H., Ozertem, U., Pantel, P., Saeidi, M., Silvestri, F., &
723 Stoyanov, V. (2020). Preserving integrity in online social networks. arXiv:2009.10311
- 724 17. Online Safety Bill. (2023). Parliament: House of Commons. Bill no. HL Bill 87(Rev).
725 London.
- 726 18. Braun V, Clarke V. (2012). *Thematic analysis*. In: APA handbook of research methods
727 in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological,
728 and biological. Washington, DC, US: American Psychological Association. p. 57–71.
729 (APA handbooks in psychology).
- 730 19. Haime, Z., Kennedy, L., Grace, L., Cohen, R., Derges, J., & Biddle, L. (2023). The
731 Journey of Engaging with Self-Harm and Suicide Content Online: A Longitudinal
732 Qualitative Study. *J Med Internet Research (JMIR) [Pre-Print]:*
733 [10.2196/preprints.47699](https://doi.org/10.2196/preprints.47699)
- 734 20. [Digital Services Act] Regulation (EU) 2022/2065 of the European Parliament and of
735 the Council of 19 Oct. 2022 on a Single Market For Digital Services and amending
736 Directive 2000/31/EC (Digital Services Act).
- 737

738 **Figure Captions**

739 **Fig 1: Visualisation of Themes and Subthemes.**

740

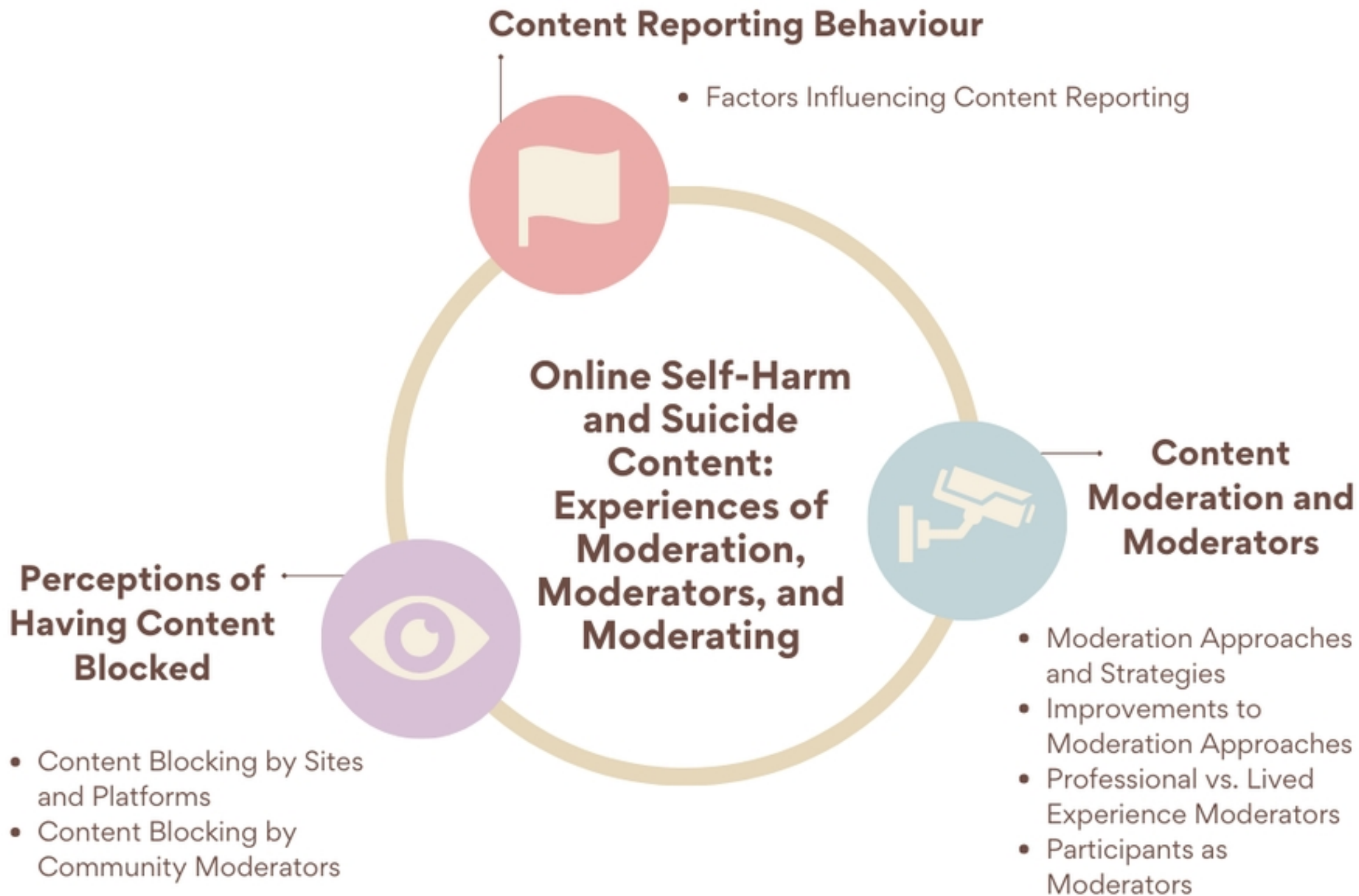


Fig 1.