Preventable yet pervasive

The prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest

Contents

Trigger warning and mental health resources

This report contains extensive references to suicide, self-harm and poor mental health, including feelings of intense depression. It also features examples of non-graphic content that were readily accessible and discoverable on social media platforms, but that may be distressing and triggering for some readers.

If any of the themes mentioned in this report are distressing, support is available from the following UK-based helpline services. Each of these support resources is available 24/7.

SHOUT – Text **MRF** to **85258**
Confidential crisis text line for anyone, any age - Free 24/7

Papyrus HOPELINE247 – **0800 068 4141**
pat@papyrus-uk.org
Confidential helpline for people under 35 or anyone concerned about a young person - Free 24/7

NSPCC Childline – **0800 1111**
Confidential support for young people under 19 - Free 24/7

Samaritans – Call **116 123** – jo@samaritans.org
A safe place to talk about whatever's getting to you - Free 24/7

In an emergency don't be afraid to dial **999**

Summary

'My FYP knows me better than my parents'

- a comment from a teenage boy responding to a TikTok post referencing suicide ideation

In November 2023, Molly Russell would have celebrated her 21st birthday.

Almost five years after Molly died, the Senior Coroner overseeing the inquest into her death recorded a narrative verdict that Molly died from an act of self-harm while suffering depression and the negative effects of online content. The inquest marked the first time that tech platforms had been held formally responsible for the death of a child.

Of course, the circumstances of Molly's death are far from unique. Suicide-related internet use has been reported in almost one-quarter (24%) of deaths by suicide among young people aged 10 to 19, equivalent to 43 lost lives every year.[1]

Across the UK, countless children and families continue to be exposed to inherently preventable technology-facilitated harm. As a direct result of the poorly considered commercial and design decisions taken by tech companies, children and young adults continue to be exposed to harmful suicide and self-harm related content. Too many young people are routinely algorithmically recommended large volumes of content that present a reasonably foreseeable risk of exacerbating feelings of depression, hopelessness and misery.

While the UK has made great strides to tackle the risks posed by social media platforms, most recently through the passage of the Online Safety Act, significant shortcomings remain in the ways that social media companies design and run their products. Children continue to be exposed to entirely unnecessary and unacceptable risks.

This report, a partnership between the Molly Rose Foundation and the data-for-good organisation The Bright Initiative, aims to assess the current exposure to that risk on three major social media sites, Instagram, TikTok and Pinterest.

Our research assesses the prevalence of harmful content available on each of the sites, through an analysis of the most engaged content posted to suicide and self-harm related hashtags; it explores the nature of suicide and self-harm related material, including the potential risks posed by harmful content being algorithmically recommended to children and young adults; and it goes on to emphasise the interplay between harmful content and high-risk, often poorly considered platform design choices.

The report aims to actively inform ongoing regulatory debates; and it sets out clear ways that social media platforms should reduce the risk profile of their services. Crucially, the report

---

[1] Rodway, C et al (2022) Online harms? Suicide related online experience: a UK wide case series study of young people who die by suicide. Psychological Medicine, 53 (10), pp1-12

provides a clear roadmap for Ofcom about how to approach its regulation of social media companies, and the need to adopt an effective, targeted and suitably ambitious regulatory approach.

What we found

Social media platforms continue to exhibit significant systemic weaknesses in how they identify and respond to the risks posed by harmful content. In this research, we find that substantial amounts of harmful material remain readily accessible and discoverable on TikTok, Instagram and Pinterest, including posts that promote and glorify suicide and self-harm, actively reference suicide ideation, and that contain intense themes of misery, hopelessness and depression.

On all three platforms, the risk profile is exacerbated by a set of poorly conceived design features and high-risk algorithms.

We found that:

- **Platforms are still failing to adequately tackle the clear and pervasive problem of harmful material, including how their systems recommend it:** while Instagram and Pinterest took some welcome steps to address the risks posed by suicide and self-harm content following Molly's death, the overall risk profile of all three sites remains unacceptably high.

  For example, in our quantitative analysis we found that two-thirds of the most-engaged posts on Instagram that reference suicide and self-harm, and that were posted using well-known suicide and self-harm hashtags, contained material that promoted or glorified suicide and self-harm (in clear violation of Instagram's community standards.)

  Almost half of the most-engaged posts on TikTok (49 per cent), and that were posted using suicide and self-harm hashtags, contained material that promoted or glorified suicide and self-harm, referenced suicide ideation, or otherwise contained intense themes of misery, hopelessness or depression. Each of these types of content have the potential to cause harmful effects, particularly if viewed in large amounts or cumulatively over time.

- **High-risk design choices increase the potential for negative effects:** TikTok, Instagram and Pinterest each use a range of user prompts, recommended search terms and other high-risk design features to maximise user engagement, but don't appear to have adequately assessed or mitigated the potential risks in relation to harmful material, including suicide and self-harm related content.

  We found evidence of a range of design features that could put young people at risk. For example, on Instagram we found deeply disturbing prompts that encourage the

use of suicide and self-harm hashtags, including #letmedie. TikTok recommends search results, at the end of relevant videos and in the middle of hashtag search results, including 'the quickest way to end it', 'I am going to end it soon', and 'attempt tonight.'

A range of design features that are designed to deepen user engagement, including the easy ability to save posts on all three sites, exacerbate the risk posed by algorithmic recommendations, and actively enable young people able to amass substantial collections of content relating to suicide, self-harm and mental distress. On TikTok, we observed that 30 per cent of the posts we surveyed had been saved over 30,000 times.

- **Algorithms continue to actively drive the spread of harmful content:** across all three platforms, algorithms and other high-risk design features enable harmful content to achieve remarkably high levels of reach and user engagement.

  On TikTok, the platform's algorithms enable harmful content to reach staggeringly high audiences, with more than half of the harmful posts we surveyed having been viewed at least one million times or more. 51 per cent of the harmful posts we surveyed, and that were posted using suicide and self-harm hashtags, had been liked by over 250,000 users.

  We are particularly concerned about the rapid growth of high engagement and meme accounts on TikTok and Instagram, many of which frequently post large amounts of suicide, self-harm and highly depressive material. Some of these accounts attract hundreds of thousands of views and are a primary vehicle through which harmful content is shared and amplified at scale.

  An exceptionally high volume of harmful content was algorithmically recommended on Instagram's short-form video product, Reels.

  99 per cent of the short-form videos we were algorithmically shown, through watching a set of posts recommended by the app's auto-play function, contained at least one type of harmful material, with more than half of posts referencing suicide ideation (often through graphic and slickly produced memes.) Much of this content was shared by high engagement and meme-sharing accounts, with relatively few accounts identifying as having personal experience of suicide and self-harm thoughts and behaviours.

  The research was unable to conclude if the significantly increased potential of exposure to harmful content on Reels, compared to other parts of Instagram, is predominantly driven by technical challenges in how the platform moderates short-form video content; or instead stems from a broader commercial decision to grow the product's user base, at the potential expense of user safety, and in a race for market share.

- **Harm reduction strategies map poorly map onto the dynamics of harm:** the focus of TikTok, Instagram and Pinterest has been to address the risks posed by harmful content through prioritising enforcement of often narrowly defined community standards, with particular emphasis on material that promotes or glorifies suicide and self-harm.

    This approach has arguably been at the expense of considering and responding to the longer-term, cumulative risks of viewing harmful content, including as a result of recommender algorithms and other high-risk design choices. As of today, no major social media company adequately addresses the risks - in either their community standards or other related policies – posed by the cumulative viewing of potentially harmful material. This includes posts that may not be harmful in isolation but that present a risk when consumed alongside other similar material.

    Of course, this harm reduction strategy has been further undermined by inconsistent, patchy and at times somewhat erratic content moderation.

    Ofcom must ensure that tech platforms pay adequate attention to the risks associated with how content is likely to be consumed, including as a direct consequence of algorithms and other high-risk design features. In developing its regulatory scheme, the regulator must ensure that platforms recognise and respond to the range of ways harm can take place, including the risks posed by cumulative and long-term exposure to harmful content.

    Platforms should be expected to appropriately balance the needs of young people posting suicide, self-harm and depressive material against the risks of this content being algorithmically recommended to large numbers of children and adolescents, including young people that may be at risk of or already experiencing poor mental health.

    Some platforms, including Meta, have arguably been able to effectively hide behind the inherent complexity of suicide and self-harm content, with the effect of being able to dismiss the ongoing prevalence of harmful material in this context. This strategy seems focussed on withstanding external scrutiny as much as being focussed on the needs of children and young people at risk of experiencing online-facilitated harm.[2]

---

[2] For example, in an interview with Radio 4's Today Programme last month, Meta's President of Global Affairs Nick Clegg responded to examples of harmful content, that had been algorithmically recommended to us while undertaking this project, by responding that: 'you will always be able to scour any platform anywhere in the world and find some images which you don't think [should be viewable.] The debate, and it is a difficult one, is where you draw the line and candidly people of good faith strongly disagree'. Today Programme, Radio 4, 11th November 2023

Methodology

This research aims to assess the availability and prevalence of harmful content on three major social media services, Instagram, TikTok and Pinterest.[3] Each of these platforms has had well-reported problems in identifying and removing harmful content, and as was heard during last year's inquest, both Instagram and Pinterest algorithmically recommended large volumes of content to Molly.

For each platform, our objective was to assess the prevalence of three main content types: suicide related content; self-harm related content; and material that contains themes of hopelessness, misery and worthlessness.

Across each of these categories, relevant content was deemed reasonably likely to be harmful if it promoted or glorified suicide and self-harm; referenced suicide methods; or if it referenced suicide ideation or themes of hopelessness, misery or worthlessness in a way that posed an increased risk when watched cumulatively or in large volumes, for example because of recommender algorithms or other high-risk design choices.[4]

In the first stage of the research, we undertook exploratory analysis of samples of content posted on Instagram, TikTok and Pinterest. Content was identified and scraped using hashtags that have been frequently used to post suicide and self-harm related material, and that Molly is known to have engaged with in the months before her death.

This analysis enabled us to systematically identify whether these hashtags were still widely used; and given the constantly shifting ways in which users post material to evade content moderation, to identify hashtags that have started to be used more recently for the purposes of making harmful forms of content discoverable. The analysis enabled the identification of problematic hashtags currently being used to share potentially harmful content on TikTok and Instagram.

A second sample was generated from another social media scrape in October 2023, using a slightly revised set of hashtags known to be associated with suicide and self-harm related material. The data collected dated back several years.

A set of focus hashtags for quantitative analysis were identified for each platform; seven for Instagram and eight for TikTok (see Appendix one). Posts on TikTok (617 posts in total) and Instagram (564 posts) with the highest levels of engagement were coded, to assess whether they contained potentially harmful content. These overall platform samples typically consisted of 60-80 of the most engaged posts for each focus hashtag. Note that some posts may have been analysed and categorised more than once, due to being associated with more than one focus hashtags.

---

[3] We had originally hoped to include a fourth platform, Twitter/X, in the sample. However recently implemented changes to X's API meant we were unable to analyze the platform using our intended methodological approach.
[4] Our definition of harm is drawn from Ofcom-commissioned research into risk factors that may lead children to experience harm online. Ofcom and Revealing Reality (2023) risk factors that may lead children to harm online.

A further 100 Instagram Reels were coded based upon posts algorithmically recommended to the research team via the feature's autoplay function.

We also performed extensive qualitative analysis to determine key trends and observations about the nature of harmful material; assess the effectiveness of platform content moderation efforts; and observe any potential interplay between harmful content and the design choices adopted by the platforms.

This analysis was undertaken using the generated data samples and through follow-up examination of content on the relevant sites. We examined this content using social media accounts opened in the identity of a 15-year-old girl.

In developing our methodological approach, we observed that hashtags are now used relatively infrequently to post suicide and self-harm related content on Pinterest. As such, we were unable to perform quantitative analysis using the same methodological approach adopted for TikTok and Instagram, but we were still able to perform qualitative analysis of the problematic content we were recommended or discovered.

For methodological reasons, the scope of this project only captures English language posts. Nonetheless it is important to note that we did observe considerable amounts of suicide and self-harm content in several other languages, particularly German and Hindi.

We would like to note the additional potential harm for children and young people who are able to read and comprehend more than one language, including languages where there appears to be an increased risk profile and/or reduced platform moderation capability.

Context

The risks of technology-facilitated suicide and self-harm

Suicide and self-harm are serious public health issues affecting children and young people.[5] Suicide is the third leading cause of death among 15 to 19-year-olds, and provisional figures indicate that last year 524 people aged 24 or under died by suicide in the UK.[6]

While the suicide rate in the under 20s is relatively low compared with older age groups, over the last decade recorded suicide rates across all age groups under 25 have increased and have now stabilised at these higher levels[7] (although it is important to note that the burden of proof for recording a death as suicide changed in 2019.)[8] The increase in the recorded suicide rate was particularly apparent among females aged 24 or under, with the most recent data pointing to the largest such increase among young women and girls since data collection began.[9]

Self-harm rates among children and young people have also been rising. In 2014, one in five female 16 to 24-year-olds reported non-suicidal self-harm, a three-fold increase since 2000.[10] There are an estimated 200,000 hospital presentations for self-harm per year in England, although the occurrence of self-harm in the community is likely to be considerably higher.[11]

There is emerging evidence of the relationship between exposure to harmful online content and resulting suicide and self-harm risks, with problematic internet usage an observed antecedent in suicide cases.[12] Suicide-related internet use has been reported in 24% of deaths by suicide among young people aged 10 to 19, equivalent to 43 deaths per year.[13]

---

[5] Department Of Health and Social Care (2023) Suicide Prevention in England: Five Year Cross Sector Strategy
[6] Office for National Statistics (2023) Quarterly Suicide Death Registrations in England: 2001 to 2021 Registrations and Q1 to Q4 2022provisional data
[7] Department Of Health and Social Care (2023) Suicide Prevention in England: Five Year Cross Sector Strategy
[8] At the time the change took effect, many academics including Prof Louis Appleby anticipated that this change would cause suicide numbers to rise and may make it hard to compare with previous data. The most recent analysis from the ONS suggests that the legal change did not result in any significant change in the reported rate, but did lead to a greater proportion of deaths being attributed to a cause of intentional self-harm. See Office for National Statistics (2020) Change in the Standard of Proof Used by Coroners and the Impact on Suicide Death Registrations Data in England and Wales. Appleby, L et al (2019) New standard of proof for suicide inquest in England and Wales. BMJ, Jul 29:366
[9] Office for National Statistics (2022) in England and Wales: 2021 registrations
[10] McManus, S et al (2019) Prevalence of non-suicidal self-harm and service contact in England, 2000-14: repeated cross-sectional surveys of the general population. Lancet Psychiatry, 6)7) pp573-581
[11] Department Of Health and Social Care (2023) Suicide Prevention in England: Five Year Cross Sector Strategy
[12] Susi, K et al (2023) Research Review: Viewing self-harm images on the Internet and social media platforms: systematic review of the impact and associated psychological mechanisms. Journal of Child Psychology and Psychiatry, 64(8). pp1115-1139
[13] Rodway, C et al (2022) Online harms? Suicide related online experience: a UK wide case series study of young people who die by suicide. Psychological Medicine, 53 (10), pp1-12

Suicide related internet use was recorded more frequently in the death by suicide of girls, and in cases affecting adolescents who identified as LGBTQ+.[14]

Suicide and self-harm related internet use was reported in 26% of child hospitalisations relating to self-harm.[15] Self-harm is considered a major risk factor for suicide among adolescents and young adults.[16]

Findings from multiple studies have raised concerns about the harmful effects of exposure self-harm and suicide related online content, with recent research concluding that suicide related online experience is a 'common, but likely underestimated, antecedent' to suicide in young people.[17]

While further research is needed to determine the strength of any causal relationship, and suicide and self-harm content has been found to have both harmful and protective effects, a recent systematic review concludes that harmful effects predominate.[18]

Potentially harmful impacts of engaging with self-harm and suicide content may include:

- increases in the frequency and/or severity of self-harming behaviour and suicide ideation. Arendt et al (2019) found that one-third of participants in their study carried out the same or similar types of self-harm after observing it on the social media site they studied, Instagram;[19]

- engagement behaviours such as sharing, liking, or commenting on suicide and self-harm content may reinforce the creation and sharing of self-harm images, and in turn encourage further harmful behaviours;[20]

- engaging with self-harm content may result in emotional, cognitive and physiological impacts, which may trigger or exacerbate self-harm behaviours and suicidal thoughts;[21]

- social networks may result in a possible 'assortative relating' effect, with young people experiencing suicide ideation or thoughts of self-harm being more likely to identify

[14] ibid

[15] Padmanathan, P (2018) Suicide and Self-Harm Related Internet Use: a Cross-Sectional Study and Clinician Focus Groups. Crisis, 39(6), pp469-478

[16] Hawton, K et al (2020) Mortality in children and adolescents following presentation to hospital after non-fatal self-harm in the multicentre study of self-harm: a prospective observational cohort study. The Lancet Child and Adolescent Health, 4, pp111-120

[17] Rodway, C et al (2022) Online harms? Suicide related online experience: a UK wide case series study of young people who die by suicide. Psychological Medicine, 53 (10), pp1-12

[18] Susi, K et al (2023) Research Review: Viewing self-harm images on the Internet and social media platforms: systematic review of the impact and associated psychological mechanisms. Journal of Child Psychology and Psychiatry, 64(8). Pp1115-1139

[19] Arendt, F et al (2019) Effects of exposure to self-harm on social media: evidence from a two-way panel study among young adults. New Media and Society, 21, pp2422-2442

[20] (2019) Effects of exposure to self-harm on social media: evidence from a two-way panel study among young adults. New Media and Society, 21, pp2422-2442

[21] Susi, K et al (2023) Research Review: Viewing self-harm images on the Internet and social media platforms: systematic review of the impact and associated psychological mechanisms. Journal of Child Psychology and Psychiatry, 64(8). Pp1115-1139

and build relationships with other users experiencing similar actions and thoughts.[22] Although this technology-facilitated effect may provide adolescents with much-needed immediate connection, validation help and support, it also presents significant risks (including the potential for unintended consequences.) Self-harm may become portrayed as an acceptable or normalised coping mechanism, and social support may also inadvertently preclude offline or expert-oriented forms of help seeking (establishing a sense that those who do not self-harm 'would not understand');[23]

- engaging with harmful content may result in the development of a 'self-harm' or 'suicide' identity, in some cases resulting in habituation to seeking harmful stimuli and the cementation of suicide ideation or self-harm behaviours;

- the risks of a 'contagion' effect, in which behaviours or ideation develop and deepen following exposure to harmful content, including as a result of poor platform design choices and practices that push content out to children. In the months before Molly's death, she was algorithmically recommended an increasing number of more severe and harmful posts.

Young people's mental health and well-being

Over the last decade, evidence has emerged identifying grounds for concern about the potential negative impact of social media on children and adolescents.

While the evidence base is complex and sometimes contradictory, and social media has been shown to have both positive and negative impacts for adolescents, there is growing confidence that poorly designed online platforms may adversely impact children's mental health and well-being outcomes.

Studies point to a higher relative concern in adolescent girls[24] and those already experiencing poor mental health, including for health conditions such as depression, anxiety and poor body image.[25]

---

[22] Arendt, F et al
[23] See for example Lavis, A et al (2020) #Online harms or benefits? The graphic analysis of the positives and negatives of peer support around self-harm on social media. Journal of Child Psychology and Psychiatry, 61, pp842-854
[24] Nesi, J et al (2021) Online self-injury activities among psychiatrically hospitalised adolescents: prevalence, functions, and perceived consequences. Research on Child and Adolescent Psychopathology, 49, pp519-531.
[25] For example, Meszaros et al (2020) found problematic Internet use was significantly positively correlated with symptoms relating to self injury affective disorders and anxiety. Meszaros, G et AL (2020) Non-suicidal self-injury: its associations with pathological Internet use and psychopathology among adolescents. Frontiers in Psychiatry, 11, p814

A recent systematic review found that adolescents with clinical level mental health difficulties may be particularly vulnerable to digitally mediated harm,[26],particularly in relation to exposure to harmful content. Young people diagnosed with depression reported more problematic internet use, as well as difficulties in regulating their digital engagement than their non-clinical peers.[27] Cross-sectional studies have shown higher rates of social anxiety, depression, or suicidal ideation in people who report suicide and self-harm related Internet use compared with those who do not.[28]

The algorithmic nature of social networks, and their resulting ability to personalise, curate and suggest even more extreme content to vulnerable users, is a particular driver of adverse mental health and well-being impacts, and plays an important role in driving the poorly regulated Internet use of some adolescents.[29]

Put simply, algorithmic design may not only expose vulnerable adolescents to harmful content but incentivises them to engage with it more intensively and for longer.

The Meta whistleblower, Frances Haugen, released a series of internal research reports that suggested Instagram was aware that it contributed to poor mental health and wellbeing outcomes for a substantial minority of its teenage users.

For example, one internal study of 1,282 teenage Instagram users found that one in five had thought about suicide or self-harm, with strongly observed risks in respect of social comparison, social pressure and negative interactions with other users.[30] Teenagers experiencing poor mental health, or that reported being generally unsatisfied with their lives, were more likely to see mental health-related content, and reported this made them feel worse.

Instagram's internal data also found that over one-third (35%) of teenagers who experienced poor mental health had reported feeling worse after using the site. 13.5% of UK teenage girls who had experienced suicidal thoughts said that Instagram had exacerbated or worsened their suicidal feelings.[31]

In summer 2023, the US Surgeon General issued a landmark advisory on the growing concerns about the effects of social media on young people's health and well-being. Advisories are usually reserved for urgent and significant public health challenges that require immediate awareness and action.

---

[26] Kostryke-Allchorne, K (2023) Review: Digital experiences and their impact on the lives of adolescents with pre-existing anxiety, depression, eating and non-suicidal self-injury conditions -a systematic review. Child and Adolescent Mental Health, 28(1), pp22-32

[27] See, for example, Ucar, H et al (2020) Risky cyber behaviours in adolescents with depression: a case-control study. Journal of Affective Disorders, 270, pp51-58

[28] Bell, J et al (2017) Suicide related Internet use among suicide and young people in the UK: characteristics of users, effective use, and barriers to off-line help seeking. Archives of Suicide Research, 1-15

[29] Stoilova, M et al (2021) Adolescents' health vulnerabilities and the experience and impact of digital technologies: a multi-method pilot study. Reported in Kostryke-Allchorne, K (2023).

[30] Copies of these research reports were published by the Wall Street Journal as part of its Facebook Files investigation, and are accessible on the WSJ website.

[31] ibid

In introducing the report, the Surgeon General warned that 'there are ample indicators that social media can have a profound risk of harm to the mental health and well-being of children and adolescents'. The report continued that 'at this time, we do not yet have enough evidence to determine if social media is sufficiently safe for children and adolescents to use.'[32]

[32] US Surgeon General (2023) Social Media and Youth Mental Health: the US Surgeon General's Advisory

Findings

Trigger warning

Each of the following chapters contain extensive references to suicide, self-harm and poor mental health, including feelings of intense depression.

The report also features examples of non-graphic content that are readily accessible and discoverable on social media platforms, but that may be distressing and triggering for some readers.

If any of the themes mentioned in this report are distressing, support is available from the 24/7 UK-based helpline services listed below.

Suicide and self-harm risks on Instagram

Instagram has been at the forefront of widespread public and expert concerns about the exposure of children and young people to harmful online content. Following the initial media coverage of Molly's death, the platform announced a number of changes to its how it moderates suicide and self-harm material.[33]

However, our research shows that while some of these changes have resulted in welcome targeted impacts, substantial concentrations of harmful content, including suicide and self-harm related material, continue to be freely accessible and discoverable.

The scale and prevalence of harmful suicide and self-harm content on Instagram remains unacceptably high, with material that promotes and glorifies suicide and self-harm among some of the most engaged and readily discoverable posts.

Please be aware this chapter contains extensive references to suicide and self-harm content, and examples of non-graphic but distressing content.

Our research has found:

- **problematic content remains pervasive and highly discoverable:** two-thirds of posts that reference suicide and self-harm appeared to promote and glorify these behaviours, in clear violation of the platform's community standards. We found repeated examples of content that is similar and/or identical to that Molly encountered in the months before her death;
- **systematic failures to respond to the risks of harmful suicide and self-harm content:** we found evidence of inconsistent and at times erratic content moderation, and the adoption of poorly conceived design features that amplify potential risks, including prompts that encourage the use of suicide and self-harm hashtags such as #letmedie. The risk profile on Instagram's new video feature, Reels, is unacceptably high;
- **the need for a revised harm reduction approach:** there is a clear need to move beyond a focus on enforcing often narrowly interpreted community standards towards a more overarching strategy for harm reduction. This approach should be actively informed by an understanding of how risks perpetuate, including as a result of the platform's design choices and algorithms. Instagram's current approach to risk assessment emphasises how content is produced, rather than the risks associated with how it may be consumed, but the result is a substantial volume of harmful content that is readily discoverable and ready to be fed into its algorithms;
- **limited attempts to tackle the cumulative risks of viewing harmful content:** Instagram's focus on moderating content, rather than emphasising systemic risks, means limited effort to address the risks posed by cumulative viewing of material, including through algorithms. Our research found that 1 in 5 posts referenced suicide

---

[33] Mosseri, A (2019) Instagram policy changes self-harm related content. Posted to the Instagram blog.

ideation, and almost half (48%) of posts had the potential to cause harm, including when 'binge watched' or viewed on a cumulative basis. There is an urgent need for platforms and regulators to develop an approach that is actively informed by developing research into the mechanisms for potential harm;

- **the need for sustained investment in and focus on the suicide and self-harm problem:** while Instagram has successfully reduced volume of graphic self-harm content available on the platform, there is evidence that the platform has failed to sustain and build on this progress. While many users demonstrate a well-developed understanding of how to game the platform's content moderation strategy, Instagram's grip on the problem appears to be several steps behind.

Scale and prevalence of harmful content

Our analysis of 564 of the most engaged with posts displaying hashtags linked to suicide and self-harm content found that a pervasive amount of harmful content remains publicly accessible and discoverable on Instagram.

More than one in eight posts we analysed (13 per cent) promoted or promoted suicide or self-harm behaviour, and in doing so, appear to be in breach Instagram's community standards. Strikingly, this accounts for almost three-fifths (59 per cent) of all posts in the sample that referenced suicide or self-harm.

As figure 1 shows, a significant proportion of the content we examined references suicide or self-harm ideation (19 per cent). Almost half (48 per cent) of posts contained content that displayed hopelessness, feelings of misery and highly depressive themes.

While previous research has demonstrated the potential for protective effects for vulnerable users posting this this content[34], there is growing evidence of a significant correlation between long-term exposure to online self-harm content and self-harm behaviours, suicide ideation, hopelessness, reasons for living, and suicide risk.[35]

As explored later in the report, there is a growing understanding of the potential ways in which cumulative exposure to suicide and self-harm content may result in increased levels of suicide and self-harm behaviour.

While further research is needed to better understand the potential mechanisms between exposure to harmful content and its effects,[36] there is a clear argument to adopt a

---

[34] See for example Marchant, A et al (2021) impact of web-based sharing and viewing of self-harm related videos and photographs on young people: systematic review. Journal of Medical Internet Research, 23.

[35] Arendt, F et al (2019) effects of exposure to self-harm on social media: evidence from paid to brave panel study among young adults. New Media Society, 21 (11-12)

[36] Susi, K et al (2023) Research Review: viewing self-harm images on the Internet and social media platforms: systematic review of the impact and associated psychological mechanisms. Journal of Child Psychology and Psychiatry, 64(8), pp1115-1139

precautionary principle approach to the design and operation of social media sites[37]. This should inform the regulatory frameworks that will increasingly underpin them.

Our results report results consistent with previous studies that have explored the scale and prevalence of self-harm and suicide content. For example, Carlyle et al (2018) reported suicidal intent being present in 19% of English language posts tagged as #suicide and #selfharm[38], equivalent prevalence to that identified in this study.

Overall, this would suggest that the broad dynamics of suicide and self-harm related content have remained largely unchanged over recent years; and that despite high levels of public scrutiny and multiple commitments from Instagram to improve their response, significant levels of potentially harmful suicide and self-harm content remain readily available and actively discoverable.

Figure 1: proportion of most-engaged content on Instagram, posted using suicide and self-harm related hashtags, which contains harmful material

Content type	% of posts
References suicide or self-harm	22
Glorifies or promotes suicide or self-harm	13
References suicide or self-harm ideation	19
References feelings of depression, hopelessness and misery which is likely to have a negative cumulative effect on well-being	48

Reach and engagement of suicide and self-harm posts

We found evidence that harmful content secures very high levels of engagement on Instagram, with some posts attracting a substantial number of likes and views, and a moderate amount of user comments.

Our results found that almost half of harmful posts (45 per cent) had generated more than 1,000 likes, which is strongly suggestive of the effects of algorithmic amplification. More than three-fifths of harmful posts (62 per cent) had generated at least 500 likes.

There appear to be a core number of 'power users' churning out some of the most problematic material, including numerous examples of violative content. Research into wider online harms has found similar trends: for example, analysis of mis- and disinformation by the

[37] Rodway C et al (2023) Online Harms? Suicide-related online experience: a UK wide case series study of young people who died by suicide. Psychol Med, 53(10), pp4434-4445
[38] Carlyle, K et al (2018) Suicide conversations on Instagram: contagion or caring? Journal of Communication in Healthcare, 11(1), pp12-18

Center for Countering Digital Hate[39] has found that a high proportion of problematic material comes from a relatively small number of user accounts.

Harmful content appears to attract a moderate volume of user comments, with almost one in seven posts (15 per cent) having received at least 25 comments. Five per cent received 50 comments or more.

Prevalence of harmful content on Reels

Our research has found a significantly greater prevalence of harmful content on Instagram's short form video product, Reels, than on any other part of the site.

As part of the research, we undertook an analysis of 100 algorithmically recommended videos, each of which were watched consecutively through the autoplay function (and after previously accessing suicide, self-harm and highly depressive content for the purposes of this project).

Disturbingly, almost all of the content we were algorithmically shown (99 per cent) contained material that promoted or glorified suicide or self-harm; referenced suicide or self-harm ideation; or otherwise featured highly intensive themes of depression, hopelessness and misery.

As figure 2 sets out, more than half of posts (55 per cent) actively referenced suicide ideation, often through highly produced and stylised memes. More than one in eight (13 per cent) of videos featured material that promoted or glorified suicide and self-harm behaviour.

Figure 2: prevalence of harmful content types in Instagram Reels, watched via algorithmically curated autoplay

Category type	Percentage of videos
Suicide and self-harm	57
Promotes and glorifies suicide and self-har,	13
Suicide ideation	55
Material that contains feelings of depression, hopelessness or misery	99

While it is unclear if the higher amounts of algorithmically recommended harmful material reflects technical challenges in how the platform moderates short-form video content, or a broader commercial decision to grow the product user base to take on TikTok and other

---

[39] Center for Countering Digital Hate (2021) The Disinformation Dozen: Why Platforms Must Act on 12 Leading Anti-Vaxers

competitors, the exceptionally high risk-profile reflects significant failings to design and implement the Reels product in a way that minimises the risks posed to children and young people.

The risk profile of Reels is further heightened by a set of choices and prompts that are designed to maximise user engagement.

This includes the addition of pop-up comments that are seemingly designed to encourage users to read and engage with often hundreds of comments, many of which appear to be lightly moderated, (and with a high proportion of posts referencing suicide ideation and thoughts of self-harm.)

We also received a prompt that encouraged us to produce new content on Reels, including a 'reel of saved posts' or a 'remix' of recently consumed material. In the context of the high volumes of harmful content we were being algorithmically recommended, including suicide and self-harm related material, this presents a reasonably foreseeable risk that users are being encouraged to produce ever greater amounts of potentially harmful material – all seemingly to feed unstoppable algorithmically-driven and curated demand.

Some memes were self-admittedly posted because they drove high rates of user engagement. One video, containing a well-known meme linked to a speeding motorbike, was posted with the caption: 'these types of Reels seem to get views.'

Trends and threat vectors in suicide and self-harm content

Introduction

The range of suicide and self-harm related content available on Instagram tends to be highly diverse, including predominantly text-based posts, videos and video-based compilations, and increasingly audio-based recordings.

Content is presented in a wide variety of formats, including pictures of suicide-related objects and paraphernalia, memes, short videos, animations, drawings, text-based posts, and references to films, TV programmes and popular songs.[40]

Suicide and self-harm related content is often presented interchangeably, with content being posted using a consistent set of hashtags covering both harm profiles, and that that appear to well-understood by those who post, seek and actively consume potentially harmful material.

---

[40] This is consistent with previous research into suicide and self-harm content on Instagram. For example, Picardo, P. et al (2020) suicide and self-harm content on Instagram: a systematic escaping review. PLoS One, 15(9)

Previous research has suggested that self-harm and suicide content is typically shared using primarily self-harm related hashtags,[41] but this study suggests this is no longer the case.

It is evident that online communities have emerged around suicide and self-harm related hashtags, and through the algorithmic amplification of posts containing related content, thereby allowing users with self-harm or suicide related interests to come together online.

The formation of online communities may result in both protective and harmful effects. A recent systematic review found that the harmful effects of forming online communities and viewing self-harm content predominate.[42]

The combined effects of algorithmic recommender systems, and to a lesser extent the poorly moderated use of hashtags, plays an important systemic role in facilitating the process of assortative relating, whereby individuals who possess similar difficulties and or interests are likely to identify and form relationships with each other.

While technology-facilitated assortative relating may contribute towards positive effects, for example through enabling the formation of well-being communities and peer-to-peer support, there are also a considerable set of potentially harmful impacts (including the potential for unintended consequences.)

For example, repeated exposure to suicidal and self-harm may underpin understandings of self-harm as a legitimate and normalised coping response;[43] this may result in the formation of descriptive norms and normalisation effects, through which users gain a perception that suicide ideation and self-harm behaviours are more common that they actually are;[44] and there is an increased risk of performative social learning effects that result in the modelling and imitation of behaviour based on shared and performed online characteristics.[45]

Findings

Our research demonstrates clear evidence of systemic failures in Instagram's response to harmful suicide and self-harm content on its platform, with the inconsistent application of safety-by-design measures rolled out following the initial media coverage of Molly's death.

There has also been an evident ongoing failure to respond to agile and constantly changing harm mechanisms. These are explored in more detail below.

---

[41] Moreno, M et al (2016) Secret Society 123: Understanding the Language of Self-Harm on Instagram. Journal of Adolescent Health, 58(1), pp78-84

[42] Susi, K et al (2023) Research Review: Viewing self-harm images on the Internet and social media platforms: systematic review of the impact and associated psychological mechanisms. Journal of Child Psychology and Psychiatry, 64(8). pp1115-1139

[43] Hetrick, S et al (2020) Understanding the Needs of Young People Who Engage in Self-Harm: a Qualitative Investigation. Frontiers in Psychology, 10, pp1-10

[44] See Susi, K et al (2023)

[45] Arendt, F et al (2019) Effects of exposure to self-harm on social media: evidence from a two-way panel study among young adults. New Media and Society, 21, pp 2422-2442

1. Impact of Instagram's 2019 changes

Following the initial scrutiny of Instagram following coverage of Molly's death, in 2019 the Head of Instagram Adam Mosseri announced a range of measures to limit the potential exposure to children and young people of harmful suicide and self-harm content.[46]

Our research finds that while some measures did have a positive target effect, the platform's overall risk profile remains unacceptably high. Instagram did successfully drive out graphic posts of suicide and self-harm, although examples can still be found.

Many of the promised measures have been applied inconsistently, and in some cases, it appears that initial efforts to improve safety-by-design were either ineffectively rolled out or simply not sustained.

Restricting search results

As part of its response to the scrutiny surrounding Molly's death, Instagram took welcome steps to restrict the use of some problematic hashtags. The platform introduced much needed friction into the search experience (with users having to click-through to access some sensitive and potentially triggering results) and introduced banner links listing helplines and links to online support pages.

However, as appendix two shows, Instagram's approach appears highly uneven, with the inconsistent application of safety-by-design measures. This arguably reflects Instagram's failure to sustain an understanding of changing threat vectors and the problematic use of hashtags among its users.

For example, while hashtags that were problematic in 2019 have been blocked (#_iwannakillmyself) or can only be accessed if the user clicks through one or more prompts (#secretsociety123), other problematic hashtags can be freely accessed, with limited or no evidence of steps to introduce any form of user friction or other mitigations.

Search results for some hashtags such as #suicidall lack any links to additional support resources. In the case of both the #suicidall and #letmedie hashtags, search results returned an algorithmically curated feature 'for you page' that actively promotes harmful and violative content, including in posts, videos and most frequently in Reels. An example is shown below.

---

[46] Mosseri, A (2019) Instagram policy changes self-harm related content. Posted to the Instagram blog.
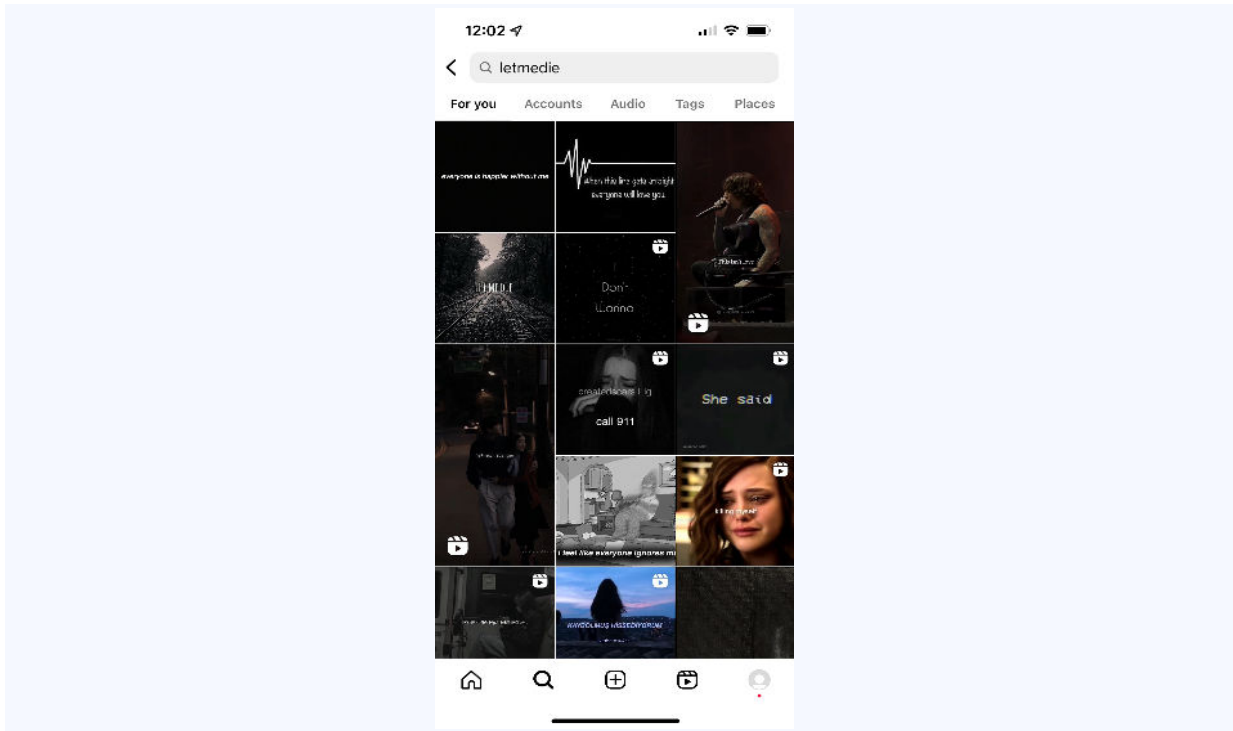
Sensitivity screens

As part of the package of 2019 announcements, Instagram announced it would consult with experts on the efficacy of using sensitivity screens. These screens were designed to ensure that posts containing non-graphic self-harm related content were not immediately visible to users.

While Instagram did proceed to introduce sensitive content screens- whereby graphic images are obfuscated with a blur and accompanied by a warning message – it appears these have been applied infrequently and inconsistently.

We identified sensitivity screens being used in less than one per cent of analysed posts. In one example, Instagram had applied a sensitivity screen to a post that contained harmful content, but when the same accounts reposted the material several months later, no sensitivity screen was apparent.

Restricting and de-indexing content

As part of its package of measures, Instagram announced it would no longer allow graphic images of self-harm. The platform also announced that it would stop displaying non-graphic,

self-harm related content in search results, hashtags and the explore tab, and that it would no longer recommend it.

Instagram appears to have had some reasonable success in identifying and removing graphic images, such as cutting. Previous research of posts tagged as #suicide and #suicidal found that 20 per cent of posts contained actual depictions of wounds.[47].However, in our results we found only a small number of graphic posts showing cutting or other related acts of self-harm.

In contrast, Instagram appears to have taken limited action to prevent non-graphic self-harm and suicide related content from being algorithmically recommended. Search results for problematic hashtags included algorithmically promoted posts that promoted and glorified suicide and self-harm, posts that referenced suicide ideation, as well as large amounts of material that may have a cumulative negative impact on mental health and well-being when viewed in volume over time.

Users can also search through algorithmically recommend lists of related hashtags and accounts that post suicide or self-harm related material. As a result, potentially harmful suicide or self-harm related content remains readily accessible and freely discoverable.

On the Discover tab, Instagram's algorithms promoted a large number of videos that contained suicide ideation and/or promoted suicide and self-harm behaviours. The results were less likely to include violative still images and text-based posts, although a substantial number referenced suicide and self-harm ideation, as well as feelings of depression, misery and hopelessness.

While it is clearly positive that Instagram has taken action to restrict the availability of graphic self-harm images, the research found a substantial number of posts that included potentially harmful suicide related content, including video compilations of suicide attempts and suicide suggestive content (for example, distressed people on railway platforms or clifftops).

Numerous posts displayed paraphernalia such as razors and pencil sharpeners often used in self-harming or glorified potential suicide methods such as pills.

Although most research to date has focused on the impact of graphic self-harm images, research suggests there may still considerable risk associated with viewing first- and third-person suicide and self-harm related imagery, with this risk heightened among people who have experienced suicide ideation or behaviours at some point in their lives. Jaroszewski et al (2020)[48] found that those who had experienced suicidal thoughts or behaviours demonstrated a reduced aversion to both first- and third-person suicide related images (aversion is potentially a natural barrier that may protect some people from harm.) This research also found evidence of a potential habituation effect, with users that had experienced suicidal thoughts more likely to experience a desensitisation effect when viewing increasing amounts of related material.

---

[47] Carlyle, K et al (2018) Suicide conversations on Instagram: contagion or caring? Journal of Communication in Healthcare, 11(1), pp12-18
[48] Jaroszweski, A et al (2020) First-person stimuli: improving the validity of stimuli in studies of suicide and related behaviours. Psychological Assessment, 32, pp663-676

Interplay between user engagement features and high-risk hashtags

Insufficient attention has been paid to date to the interplay between Instagram's user engagement features and the use of suicide and self-harm related hashtags.

Instagram has sought to maximise user engagement by using prompts in search results that encourage users to post and engage using relevant hashtags. As the example below demonstrates, this has resulted in unfortunate examples of users being encouraged to 'use hashtags' such as #letmedie and #suicidall, with prompts that contain a click-through link that automatically opens the user's phone camera.

The poorly thought out implementation of this design choice underscores the failure of the app's product teams to properly assess the potential risks and unintended consequences of making product changes and new design choices.

Instagram is already subject to regulatory requirements around responsible design, for example through the ICO's Children's Code. This failure will inevitably raise questions about whether the platform is discharging its responsibilities in a suitably thorough and consistent way.
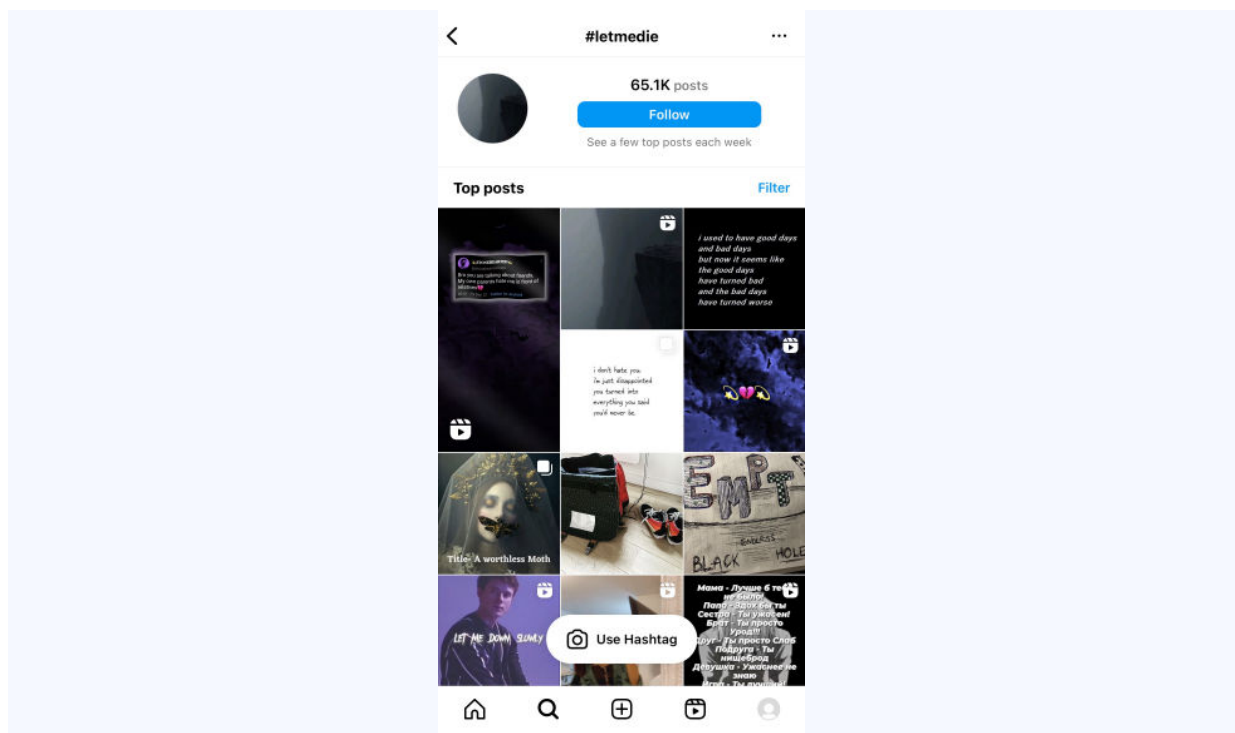
2. Strategies to game content moderation, and Instagram's response

There is substantial evidence that Instagram's moderation efforts have failed to keep up with changing and highly agile suicide and self-harm content behaviours, and that the efficacy of its moderation capabilities has lagged behind design choices that enable ever more diverse ways to post content.

The result appears to be a pronounced but inherently avoidable asymmetry between user behaviour and Instagram's threat response, with Instagram demonstrating a failure to adequately or proactively identify or respond to ways in which its platform design and moderation strategy can be readily exploited.

We identified a surprisingly diverse set of ways in which users demonstrate a strong working understanding of where Instagram's content moderation appears to be lacking, and in turn where weak links can be productively exploited. For example:

- Some users look to exploit the differential effectiveness in Instagram's moderation of video and audio material through increasingly using audio, overlaid onto video or text-based content, to post problematic content;
- embedded watermarks are used in video-based content to share links to other accounts posting suicide and self-harm content, including links to 3rd party platforms such as X;
- German-language hashtags are widely used to exploit seemingly less effective moderation arrangements in that language. Previous research has found a high proportion of potentially harmful material in German suicide and self-harm posts,[49] so it is unsurprising this has been identified and exploited by accounts seeking to post harmful and potentially violative content.

A long-term strategy to evade content moderation has been the use of 'algospeak', in which users abbreviate, misspell or substitute specific words to avoid Instagram's content moderation filters, while simultaneously ensuring that target audiences can continue to readily view and locate their content.[50]

As Steen et al (2023) noted,[51] while algospeak is primarily used to circumvent algorithmic content moderation, it may also facilitate the formation of online communities with a common interest in suicide and self-harm related behaviours. Posting using algospeak hashtags may therefore enable users to establish their identity and performatively assert their identity to this community, in doing so potentially amplifying a set of protective but also harmful effects.

Algospeak is by no means a new phenomenon, and as such Instagram's failure to effectively respond to such a well-observed and reasonably foreseeable risk is noteworthy. Users

---

[49] Arendt, F et al (2019) Effects of exposure to self-harm on social media: evidence from a two-way panel study among young adults. New Media and Society, 21, pp 2422-2442

[50] Levine, A (2022) From Camping to Cheese Pizza, Algo Speak Is Taking over Social Media. Forbes, 19/09/2022
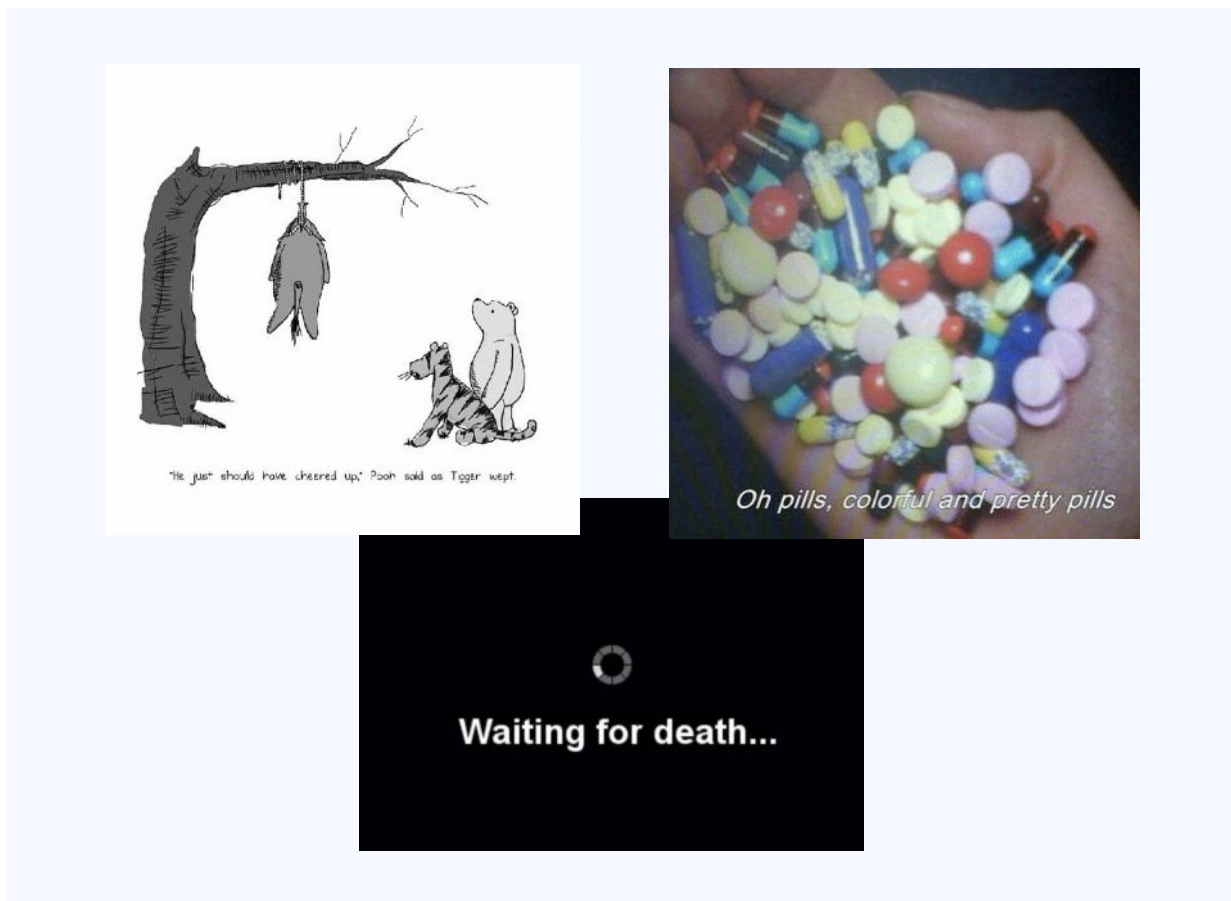
[51] Steen, E et al (2023) You can (not) say what you want: using Algo speak to contest and evade algorithmic content moderation on TikTok. Social Media and Society, 9(3)

demonstrate a high degree of algorithmic literacy,[52] with a keen understanding of constantly evolving hashtags. Some users draw on third-party sites such as Reddit to share knowledge and intelligence on weaknesses in Instagram's threat response.

Instagram's failure to keep up with such a keenly observed risk profile, and to adequately respond to a constantly evolving and highly agile set of user behaviours, demonstrates an unnecessary and preventable asymmetry between the threat profile associated with suicide and self-harm content and the platform's response.

Hashtags including #suÏcidalthoughts (13.6 posts), #selfharmm (over 5k posts) and #selfharnn (over 5k posts) are not only freely available but contain a significant number of problematic and harmful results (see below.)

In this respect, Instagram appears to be failing on the very basics.

'He just should have cheered up,' Pooh said as Tigger wept.

Oh pills, colorful and pretty pills

Waiting for death...

---

[52] Oeldorf-Hisch, A et al (2021) What Do We Know about Algorithmic Literacy? The Status Quo and a Research Agenda for a Growing Field. New Media and Society,

3. High engagement and meme accounts

A substantial proportion of harmful suicide and self-harm content is posted on high engagement accounts, accounts which frequently post memes, videos and text-based posts to quickly gain followers and maximise user engagement. High engagement accounts typically feature a wide range of material designed to appeal to those experiencing poor mental health and depression, although the material frequently veers into posts that reference suicide and self-harm content (including posts that may be violative or that present potential cumulative risks from repeated exposure).

There is clear evidence that some of these influencer accounts are driven by engagement rates and in some cases potential monetisation opportunities. For example, one account with over 100,000 followers actively invites paid features (its bio reads 'paid features DM me.')

Other accounts and Reels posts appear to offer paid promotions and link to their CashApp wallets. In one example we found an account that linked to a Telegram group offering fake followers.

We identified that accounts were readily able to identify themselves in their Instagram bios using descriptions such as 'mental health resources', 'public figures' and 'crisis prevention centres', imbuing a false sense of legitimacy and demonstrating how Instagram's design choices can be readily gamed by high engagement accounts.

Even in instances where the accounts appear genuinely committed to offering peer-on-peer support, there are obvious risks if accounts can overstate or misrepresent their status or expertise to potentially vulnerable followers.

Several accounts have made full use of recent design choices, including the introduction of Instagram Stories and broadcast channels, to rapidly build their follower base and engagement levels. A number of high engagement accounts were early adopters of broadcast channels, a new design feature in which subscribers receive posts and messages that appear alongside their DMs.

Close attention is required into the potentially high-risk ways in which broadcast channels may be used. For example, one high-engagement account with over 55,000 followers posts a daily 'mental health check-in', in which users are asked to identify with a set of options including 'I feel numb' and 'having suicidal thoughts.'

Although some users may benefit from potential protective effects of feeling part of a relatable community, such use cases also present a potential set of unintended and undesirable consequences. These include the potential for descriptive normalisation (the perception that behaviour is more common than it actually is), and the potential for social learning effects (whereby there is a risk that moods and behaviours may be modelled or imitated based on exposure to the shared characteristics of the group.)

Many accounts present themselves as offering help and support to those experiencing mental health problems, and typically feature similar and/or identical bios including an offer for

anyone experiencing distress to DM them. In some cases, accounts also offer links to suicide hotlines.

While those running high engagement accounts undoubtedly have a wide set of motivations for doing so, the consistency in how they present themselves is striking (and is unlikely to be coincidental.) There appears to be a shared and well-developed understanding of how high engagement accounts, and accounts that feature suicide and self-harm related content more generally, should present themselves to reduce the risk that their content or accounts are taken down.

Some examples of suicide-related and highly depressive content posted to high-engagement accounts are shown below.
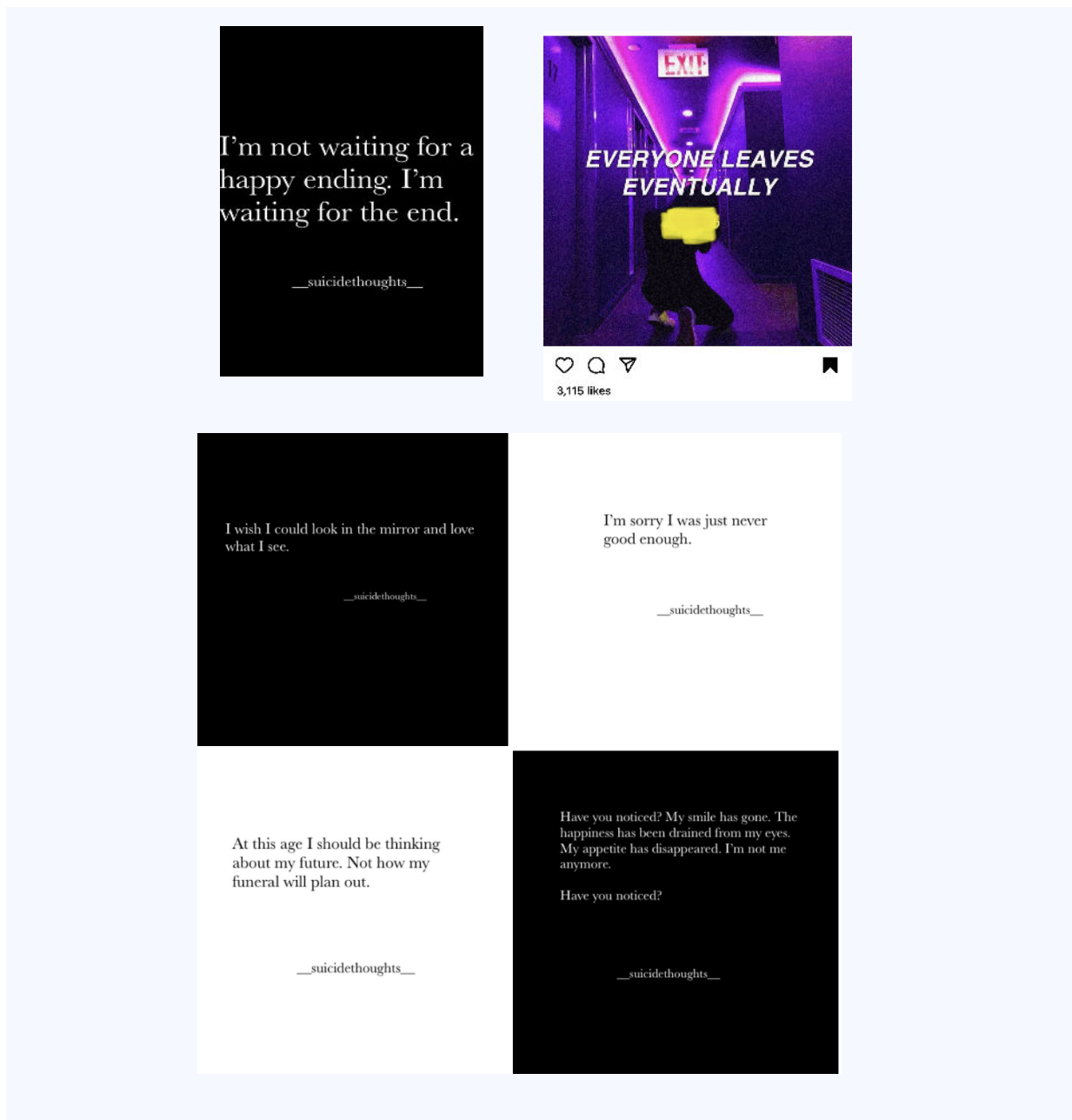
4. Instagram's approach to classifying and detecting harmful content

A significant proportion of harmful suicide and self-harm related content remains accessible on Instagram not only because of its patchy and sometimes erratic approach to content moderation, but because of the largely flawed systems and processes that the platform uses to assess and mitigate the potential for content to cause harm.

Our research found a substantial amount of content that has a cumulative potential to cause harm, particularly when recommended in volume by Instagram's algorithm; but that isn't considered harmful because of the narrowly defined and highly specific ways in which Instagram assesses harm against its community standards.

We identified numerous examples of content that could reasonably be considered harmful when viewed (or designed to be viewed) in substantial amounts. For example, there are a large number suicide and self-harm related memes videos, including multiple variations of Reels in which the user announces 'this is what I want for my birthday', before cutting to video of a high impact car crash. In another well-observed meme, video is shown of a motorbike travelling down a motorway at dangerously high speeds, with accompanying text-based or audio captions 'promising to make it look like an accident.'

Disturbingly, we found a substantial number of posts that appeared to be similar and/or identical to content seen by Molly in the weeks and months before her death, for example a post that reads 'who would love a suicidal girl?'. We also identified a significant volume of posts that appear to romanticise acts of self-harm by young people, or that normalised suicide and self-harm ideation in a disturbing and dangerous way.

If Instagram is to effectively tackle the risks posed by harmful suicide and self-harm related material on its site, it is clear the platform needs to move its current focus on a narrow set of community standards – rules that assess predominantly whether a piece of content promotes, glorifies or contains graphic material relating to suicide and self-harm (among a set of other characteristics) - towards a more sophisticated understanding of the ways in which consuming such material has the potential to cause harm, including as a result of repeated exposure or 'binge watching' driven by algorithms.

Instagram also needs to reassess the ways in which it balances the needs of those who post potentially harmful content (including those experiencing suicide or self-harm acts), alongside the risks posed by allowing such content to be made readily available to a large number of other users, including potentially vulnerable children and young people.

During Molly's inquest, Meta's Head of Health and Wellbeing Elizabeth Lagone described much of the content Molly viewed as 'admissive', and she told the court she thought that Instagram was 'safe for people to be able to express themselves.' When asked by the coroner if she thought examples of the material viewed by Molly was safe, but which her family considered to be 'encouraging' suicide and self-harm, Ms Lagone confirmed she considered the material as safe. The content would therefore be acceptable within Instagram's community guidelines.

In many respects, this testimony underscores the reasons why so much harmful suicide and self-harm related material remains freely accessible and discoverable on Instagram - and why so much has been identified in this research.

At present, Instagram's policies place considerable emphasis on creating a space in which young people experiencing suicidal self-harm ideation can express themselves, and the potential benefits to them that may result. In practice, we found these policies are operationalised to give very high levels of discretion to accounts that identify as having suicidal or self-harm ideation, and to accounts that use hashtags such as 'vent'.

This appears to be well-understood among users, and in the case of some high engagement and meme-compilation accounts appears to be exploited in order that they can game the platform's content moderation rules.

Even if Instagram could credibly claim that viewing suicide and self-harm related posts may have some protective effects when viewed in the context of a clearly vulnerable users feed, the use of algorithmic recommendations means most engagement is likely to take place stripped of this accompanying context. Content that may not necessarily pose a risk when viewed in isolation may contribute towards a substantial risk if it is algorithmically recommended in feeds, in search results, or most pressingly if consumed in volume alongside other similar material. Examples are shown below.

Other platforms are taking welcome early steps to recognise and respond to the risks posed by material that may be innocuous as a single video, but that could be problematic for some teens if viewed in repetition. For example, YouTube recently announced it would limit repeated recommendations of videos in the United States related to potentially problematic categories, including content that compares physical features and idolises some types over others, or that idealises specific fitness levels or body weights.[53]

Meta should urgently explore how it can adopt similar approaches, in a way that is sensitive to the needs of those posting content, but that also appropriately recognises and responds to the substantive and reasonably foreseeable risks that such content may become harmful or dangerous to some users when consumed.

[53] YouTube (2023) Continued support for well-being and mental health on YouTube. Posted on YouTube's Official Blog, November 2023

5. Risks and unintended consequences of self-harm, suicide and depression communities on Instagram

While much attention has been paid to the potentially protective effects of posting and engaging with suicide and self-harm related content for some users, including the direct and indirect benefits of receiving community support, there are also a set of substantial risks that Instagram needs to carefully manage, including the potential for unintended consequences.

Posts relating to suicide and self-harm related content typically attract high rates of engagement, including a significant number of comments. Many comments are supportive and highly empathetic, encouraging a sense of social connection and providing support and assistance to users experiencing mental health distress.

However, we found repeated examples of unmoderated comments where a substantial number of young people were expressing suicide ideation or sharing tips on how to conceal self-harm behaviours.

There is a risk that large volumes of comments may risk normalising self-harm as an acceptable coping strategy, trigger emotional dysregulation effects, or may encourage adolescents to understand that suicide ideation and self-harm behaviours are more common than in reality.

There is also the risk that online social support may intentionally or inadvertently preclude seeking clinical or expert help.[54] In a substantial number of posts, users expressed a sentiment that only those who experienced suicidal self-harm ideation could truly understand and offer them support. In a small number of posts, users discouraged their followers from seeking external support, including by warning of the risks that if a user contacted suicide hotlines, they might breach their confidentiality.

An underexplored but important topic is how posting or engaging with suicide and self-harm related accounts may result in young people identifying and contacting other users, and the potential risks that may result.

A large number of accounts actively encourage users experiencing distress to DM them; and by posting comments, replies or taking part in 'mental health check-in polls', an even greater number of young people may signal they are potentially vulnerable.

As part of its regulatory risk assessments, Instagram should set out what assessment it has made of how its services may be used to facilitate attempts to incite or assist in acts of suicide and self-harm, including via direct messages.

Instagram should also commit to supporting research that sets out the **potential harm pathways** that may start from engaging with or posting references to suicide and self-harm.

---

[54] This phenomenon was explored in more detail by Lavis, A (2020) #online harms or benefits? An ethnographic analysis of the positives and negatives of peer support around self-harm on social media. Journal of Child Psychology and Psychiatry, 61, pp842-854

This should include the risks that:

- Instagram is used to identify and target potentially vulnerable users to encourage or assist them in acts of self-harm and suicide, both of which may now constitute criminal offences;
- The platform is used to encourage users to migrate to, or assist the discoverability of, high-risk third-party sites, including high-risk pro suicide platforms, and;
- the risk that vulnerable children are readily identified and targeted, as a result of posting or engaging with suicide, self-harm and depression-related material, for other malign purposes such as sexual grooming.

Harmful content on TikTok, including suicide and self-harm material

TikTok demonstrates an exceptionally high-risk profile, largely driven by harmful material being algorithmically recommended to a large and potentially vulnerable teenage user base.

While TikTok appears to enforce its community standards more effectively than some other platforms, its approach to harm-reduction is narrowly defined. As a result, TikTok responds inadequately to the risks associated with how content is consumed on its platform, including the long-term and cumulative risks posed by viewing harmful content in large amounts.

Much of TikTok's risk profile is driven by a set of poorly conceived and executed design features, many of which exacerbate the risks associated with viewing harmful content in large volumes.

There is limited if any evidence that TikTok adequately risk assesses its product and design decisions to mitigate the risks posed by suicide and self-harm content, including the disproportionate risks faced by those already experiencing thoughts of suicide or self-harm ideation.

Please be aware this chapter contains extensive references to suicide and self-harm content, and examples of non-graphic but distressing content.

Scale and prevalence of harmful content

Our analysis of 617 of the most engaged with posts displaying hashtags linked to suicide and self-harm content found a significant amount of harmful content that is freely accessible and discoverable on TikTok.

Around one in five posts that referenced suicide or self-harm appeared to promote or glorify suicide and self-harm behaviours (4 per cent of all posts). Although still unacceptably high, the prevalence of suicide and self-harm content that breaches TikTok's community standards is significantly lower than the equivalent figure for Instagram.

While there is some evidence that TikTok has been reasonably effective at moderating violative content, we found a substantial amount of content that has the potential to cause harm when consumed in volume; for example, resulting from the algorithm and consumption-related design features.

Suicide and self-harm content was identified in more than one in five surveyed posts (22 per cent), with 16 per cent of posts referencing suicide ideation and intentions to self-harm. These results are broadly in line with previous prevalence estimates for social media.

Almost half (49 per cent) of posts contained content that displayed feelings of hopelessness, misery and highly depressive themes, broadly mirroring the prevalence data for Instagram.

We found particularly problematic concentrations of content using hashtags closely associated with suicide ideation and thoughts of self-harm. For example, almost 30 per cent of posts using #iwantoendit contained suicidal thoughts, and over three-fifths (62 per cent) contained material that posed a potential risk, particularly if viewed in volume.

Similarly, 54 per cent of posts featuring the hashtag #ventaccount were found to contain potentially problematic content, with 85 per cent of posts referencing suicide and self-harm containing suicide ideation or thoughts of self-harm.

Figure 3: most engaged with content on TikTok, posted using suicide and self-harm hashtags, which feature harmful material

Content type	% of posts
References suicide or self-harm	22
Glorifies or promotes suicide or self-harm	4
References suicide or self-harm ideation	16
References feelings of depression, hopelessness and misery which is likely to have a negative cumulative effect on well-being	49

Reach and engagement of suicide and self-harm related posts

Suicide, self-harm and highly depressive content generates exceptionally high levels of engagement, with content on TikTok typically generating significantly higher views, likes and comments than on any other comparable site, including Instagram.

Harmful content on TikTok achieves extraordinary levels of reach: more than half of the posts in our sample (54 per cent) received over one million views, and almost two-thirds of posts (64 per cent) were viewed more than 250,000 times.

Harmful content also generates substantial amounts of likes. More than half of posts (51 per cent) were liked at least 250,000 times. Staggeringly, one in every eight posts (12 per cent) was liked by more than a million users.

The volume of comments and replies is also considerable. 20 per cent of harmful posts received more than 5000 comments, with more than one-third of posts (36 per cent) receiving at least 2,500 replies.

In effect, some posts serve as a de facto discussion forum for users who are experiencing suicide ideation, thoughts of self-harm, and other categories of mental distress.

As figure 4 shows, posts using the eight hashtags examined in this report have so far received over 13 billion total views. Many of these posts may not necessarily contain harmful content,

but nonetheless underscore the sheer scale and reach of hashtags that are frequently used to share suicide and self-harm related material – and the need for an effective systemic response to a set of largely algorithmically-driven risks.

Figure 4: total views of high-risk hashtags used to share suicide, self-harm and highly depressive material

Hashtag	Total views
#depressedquotes	1.1 bn
#drained	1.3 bn
#icantdothisanymore	467.8 million
#iwanttoleave	12.8 million
#iwanttoendit	5.9 million
#paintok	8.4 billion
#ventaccount	3.5 billion
#SVV	132.9 million
	13.6 billion views

Trends and threat vectors in suicide and self-harm content

Diversity of content and user evasion strategies

There is a broad range of suicide and self-harm related content available on TikTok, with posts typically featuring a combination of video, audio and text-based material. Content is often presented and edited to a high standard, taking advantage of the high quality and accessibility of TikTok's well-developed editing tools.

A wide variety of formats is used to present suicide and self-harm material, including violative and/ or problematic material. These include edited videos, animations, memes, carousels and references to popular culture.

Song lyrics are often used out of their intended context, including Coldplay ('for you I'd bleed myself dry'), Adele ('should I give up…') and an edit of the Billie Eilish track 'My Future' ('I'm not coming home.') Billie Eilish's song is frequently used to accompany scenes suggesting suicide ideation and suicide acts, for example standing on the edge of a railway platform or peering over a motorway bridge.

Memes are frequently used to share suicide and self-harm material and potentially problematic content. Examples include footage of a Pac-Man style game, overlaid with audio that interchangeably expresses suicide ideation, suicide and self-harm behaviours, and

feelings of misery, hopelessness and depression. One such video references the self-harm behaviours of an 11-year-old ('putting deep lines into her skin while salty tears fall from her face [...] she cannot take it anymore'). At the time of publication, this video had received over 5.5 million views, 584,000 likes and nearly 10,000 comments.

We found evidence of numerous ways that users post content to evade TikTok's automated moderation, and to exploit the apparent differential effectiveness in its ability to detect visual and audio-based forms of harmful content.

Multiple users exploit TikTok's less developed audio moderation capability to post harmful content. For example, we found numerous videos that display innocuous captions but that feature subtly different audio transcriptions. In several examples, we found videos describing suicide ideation and thoughts of self-harm that contained text-based captions suggesting a user is an adult, when the accompanying audio says they are only 13.

Some accounts appear to have identified a weakness in TikTok's ability to detect differences between audio and text-based content. For example, one post that had attracted over 13,000 likes features the song lyric 'I don't wanna talk right now, I just want to watch TV.' A text-based caption changes the lyric to: 'I just want to leave.' (Another example more explicitly states: 'I just want to die.')

Carousels and slide decks are also routinely used to minimise the risk of automatic content detection. For example, users may post seemingly innocuous initial slides before referencing material that promotes or glorifies suicide and self-harm in subsequent slides.

In many cases, this appears to have inspired well-established formats to present suicide and self-harm material, and in some cases memes.

Ease of discoverability and algorithmic amplification

Suicide and self-harm related content is freely discoverable on TikTok, with limited if any evidence of measures to restrict or add friction to searches for hashtags associated with problematic material.

Although TikTok offers links to support resources for certain hashtags, these appear to be applied inconsistently. For example, search results for hashtags including #iwanttoendit, #drained and #depressedquotes offered no links to support resources, no click-through prompts to add friction into the search experience, nor provided any indication that search results were restricted or limited in any way.

Disturbingly, it appears that TikTok has taken limited if any steps to address the significant adverse risks associated with potentially harmful content being algorithmically recommended to its users, including children and young adults.

TikTok's For You Page (FYP) rapidly identified our interest in suicide and self-harm related material, and we were quickly presented with a range of disturbing and potentially harmful

videos. For example, we were shown a video with the text caption 'sleep can't help this type of tired anymore', with accompanying audio stating 'you've gotta die, I'll tell you why.'

Other videos included posts including: 'it's k1llling you inside, isn't it?' (15,000 views and 3,000 likes), 'no amount of sleep can fix this type of tired' (9,600 likes) and a caption that read 'I pick my poison' before showing pictures of a blade (400 likes.)

We found several examples of teenagers observing the relationship between the content being recommended on their For You Page and their overall mental health. One teenager remarked on the type of content he was being recommended by stating 'guess I'm not good in my head again', while another teenage girl seemed to plead not to keep being recommended the volumes of suicide-related and depressive material she was being shown. She replied to a post by stating: 'why does it show me so much of this when it's making me ill?'

The potential negative effects of algorithmic application are sharply heightened by TikTok's use of a set of high-risk design features, including in-video search recommendations that return highly problematic recommended hashtags and search terms ('people also search for 'quickest way to end it', 'I am going to end it soon', 'I'm going to end it').

TikTok also returns lists of recommended search terms in search results, many of which are highly problematic ('others searched for 'I feel like I'm drowning mentally', 'I'm drowning in my tears quotes'). We were also served autocomplete suggestions for search terms, with a search for 'want to…' prompting options including 'want to end it', 'want to cut', 'want to give up', and 'want to go missing.'

TikTok's failure to risk assess the impact of its design choices on users experiencing mental distress, including the appropriateness of using such user retention and engagement prompts to recommend high-risk categories of harmful content, raise substantive questions about compliance with existing regulation, including the ICO's Children's Code.

Examples of poorly considered engagement features are displayed below.
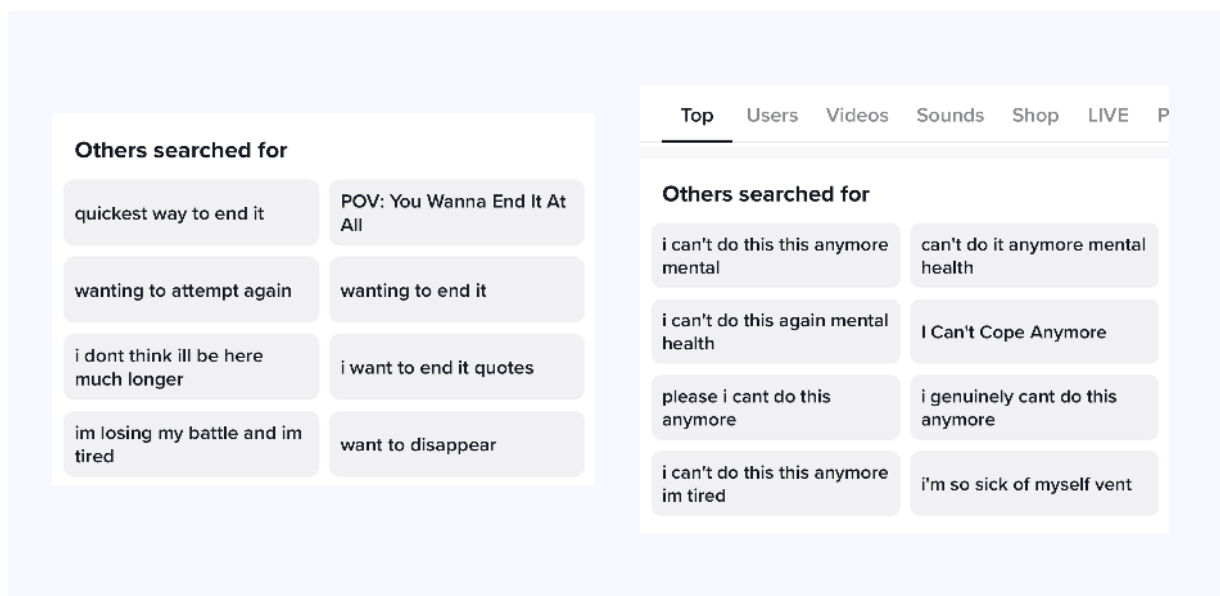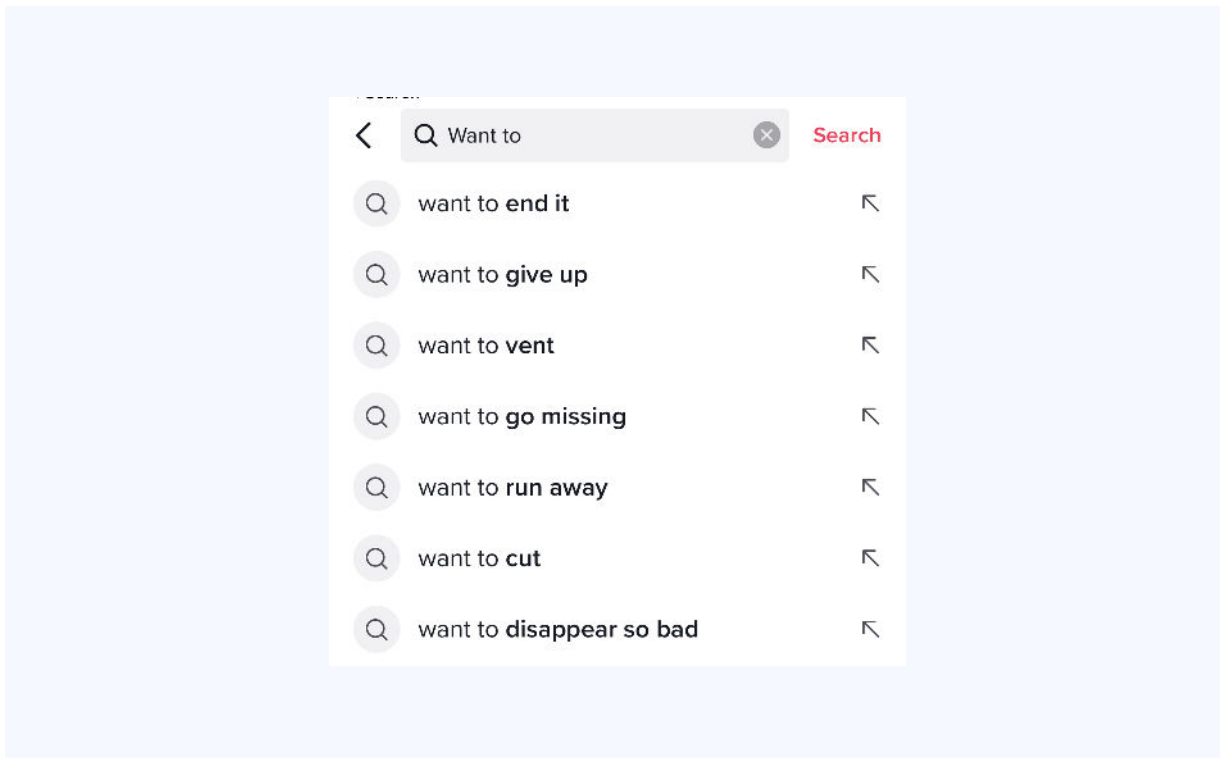
Broader impacts of algorithmic amplification

Features that reduce friction in the search and discoverability experience, such as the use of algorithms and user prompts, remove the need for users to make decisions or actively seek out content. It is therefore an evidently reasonably foreseeable risk that a combination of frictionless algorithmic recommendation, in combination with the intensive use of recommender prompts, could pose significant risks in relation to suicide, self-harm and mental health content, and to those being recommended it, unless robust and demonstrably effective mitigations are in place.

In TikTok's case, it appears that its use of algorithmic recommendations not only enable the ready discoverability of harmful material in English language posts, but also reduce the friction associated with discovering related material in other languages (where content moderation standards may be weaker).

For example, we found evidence that the platform's recommendation prompts actively guide users towards problematic self-harm content, including material displayed under German-language derived hashtags such as #SVV. There is evidence that many users posting suicide and self-harm related content now use #SVV, as well as other English and German language hashtags, interchangeably.

Our analysis of #SVV posts, focusing only on content posted or easily understood for English language speakers, identified more than one in five posts (21 per cent) that expressed suicide or self-harm ideation, with 5% of posts considered violative.

The cumulative and long-term risks of viewing harmful content

Our research suggests there is a significant potential risk posed by cumulative exposure to, and engagement with, potentially harmful content on TikTok, driven largely by large volumes of problematic material being algorithmically amplified to users.

TikTok's risk profile is heightened by the demonstrable reach that many harmful posts have, with some attracting hundreds of thousands of views; and the high-quality and stylised nature of much of the suicide and self-harm related material available, which makes it more likely young people will choose to watch and engage with substantive amounts of content.

Much of the suicide and self-harm related material available on TikTok often relies on emotionally provocative and potentially triggering imagery (including mental images). The emotionally triggering nature of images compared to textual and verbal content has been well-documented[55].

A significant proportion of potentially harmful content is shared by a set of high-reach, high-engagement accounts that specialise in sharing memes and other clips relating to suicide, self-harm and feelings of depression, misery and extended hopelessness. Many of these accounts appear geared towards serving online communities that seek out and engage with harmful content, and actively use hashtags such as #vent, #paintok and #3amthoughts that are widely understood and used by those experiencing poor mental health, anxiety and depression.

The algorithmic reach of high engagement accounts is extraordinary: the 'top post' using the hashtag #3amthoughts has received 1.3 million likes, been saved 220,000 times and shared by more than 102,000 users. A #paintok post that references suicide ideation and feelings of hopelessness has received 773,000 likes 133,000 saves.

Another #paintok post that references suicide ideation has received 1.1 million views and over 142,000 likes. The post culminates with the message 'when everyone else leaves, why the fuck are you still here?'

The case for expanding TikTok's approach to harm reduction

---

[55] Susi, K et al (2023) Research Review: Viewing self-harm images on the Internet and social media platforms: systematic review of the impact and associated psychological mechanisms. Journal of Child Psychology and Psychiatry, 64(8). pp1115-1139

As is the case with Instagram, it appears that TikTok's approach to tackling the risks of problematic content has been focused on the enforcement of narrowly defined community standards, focusing primarily on material that explicitly promotes and glorifies self-harm. As a result, TikTok has arguably failed to recognise the pervasive and potentially harmful effects of longer-term, cumulative exposure to problematic content, including as a result of algorithmic recommendation.

There is a clear and persuasive case for urgent action to address the risks posed by problematic content on TikTok, including the swift adoption of a precautionary principle approach. TikTok should adopt a corresponding set of measures that aim to reduce or prevent the algorithmic amplification of potentially harmful categories of posts and introduce additional friction to the searchability and discoverability of harmful content.

An underexplored but hugely important aspect of TikTok's risk profile are the broader mechanisms and design features that actively enable consumption of large amounts of material, including the potential for binge watching.

While considerable attention is justifiably paid to TikTok algorithms, attention should also be given to the other ways in which users can save, store, engage or share suicide and self-harm related material.

In particular, we found substantial evidence that users are saving significant amounts of harmful content, including suicide and self-harm related material. 30 per cent of harmful posts we surveyed had been saved by at least 10,000 separate users, and 4 per cent had been saved more than 50,000 times.

While further research is needed to understand this set of consumption patterns, and to better assess the potential for resulting negative effects, it is clearly disturbing to see potentially vulnerable children and young adults being able to build albums or collections of harmful content.

Given the reasonably foreseeable risk this could facilitate 'binge watching', and the resulting potential for emotional dysregulation, triggering thoughts, and even the onset of thoughts of self-harm and suicide ideation in some more vulnerable adolescents, there is a clear need for TikTok to address and respond to the potential risks, with an urgent need to assess the particular risks faced by children and young people.

Algospeak and user strategies to game content moderation

While TikTok has demonstrated clear systemic issues in how it classifies and algorithmically recommends harmful content, we also found evidence that it has been slow to systematically identify and respond to the agile and constantly changing ways in which users post and discover harmful content, including through 'algospeak'.

We found evidence of extensive and commonly understood terms relating to suicide and self-harm, including the acronym 'KYS' ('kill yourself'), 'sewer slide' and 'unalivicide.' Search

results for each of these terms produced a range of content which included potentially harmful material, including the search terms 'attempt tonight' and 'attempt at school.'

Disturbingly, TikTok's failure to close the asymmetry between user behaviour and its moderation resources is particularly apparent in user discussions around self-harm behaviours and acts.

Users have developed a set of well-understood terms to refer to self-harm behaviours, including references to the depth of self-harm cuts ('styrofoam', 'beans' and 'laffy taffy' are used as graphic descriptions of the severity of cuts, while references to 'it going white' or 'white beans' reference cuts that reach into the deepest layers of skin.)

While some posts using these terms aimed to provide peer on peer support and reassurance, others have the effect of normalising or reassuring users they could self-harm in relatively severe ways ('if it goes white it'll be fine, it's the styro.') In some cases, it was left to other users to express concerns about the potential impacts of such content: as one teenager put it, 'self-harm is so competitive, this might make someone try to cut deeper.'

TikTok's failure to effectively moderate this particularly sensitive content is exacerbated by its algorithm recommending related search posts, for example 'styro cut shh' and 'hitting the beans.'

Search results contained harmful content and poorly moderated comments that promote and glorify self-harm acts. For example, some comments read 'seeing the blood run down is so satisfying', and 'I love it.' In response, one girl replied: 'I miss it, [this] made me relapse.'

Examples of problematic algospeak suggestions that relate to self-harm are shown below.
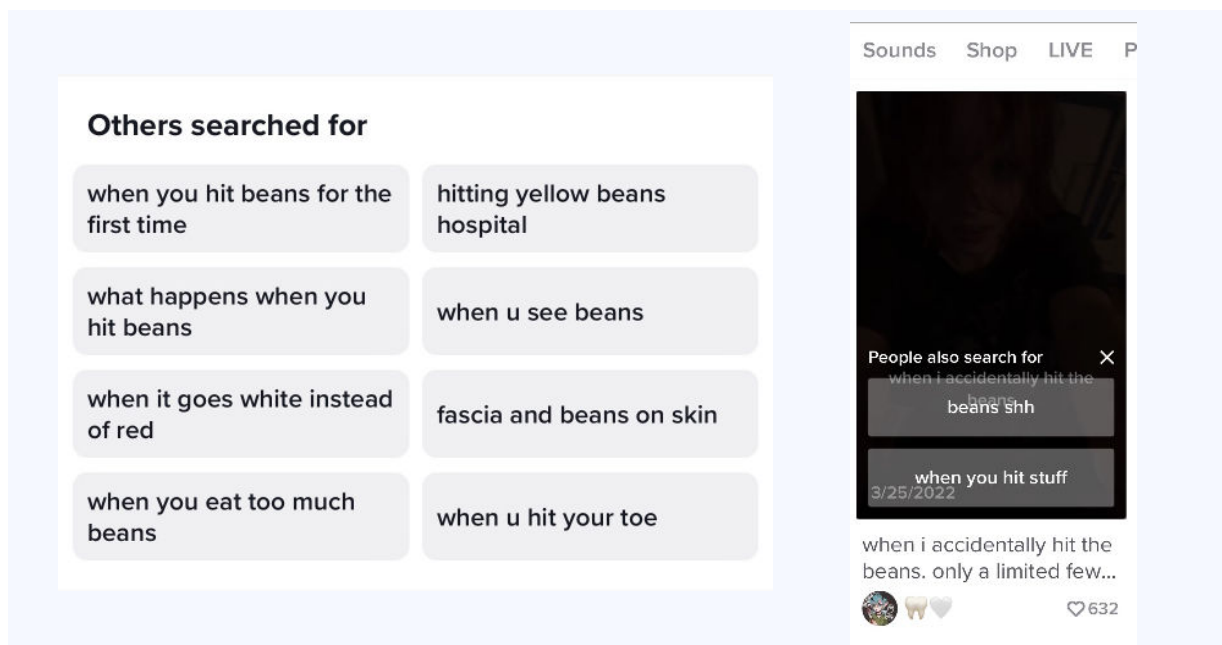
Harmful content on Pinterest, including suicide and self-harm material

At Molly's inquest, Pinterest acknowledged that the platform wasn't safe for her at the time she was using it[56]. The inquest heard evidence that Molly had been bombarded with material, including through algorithms and other poorly conceived and often high-risk design features.

Our research found that although some improvements have been made, harmful content is still readily available and continues to be algorithmically recommended at scale.

This includes material that promotes suicide and self-harm, references suicide ideation, and that contains themes of misery, hopelessness and depression, each of which have the potential to cause harmful effects, particularly when viewed in large amounts.

Please be aware this chapter contains extensive references to suicide and self-harm content, and examples of non-graphic but distressing content.

Trends and threat vectors in suicide and self-harm content

We found significant evidence that content on Pinterest continues to pose a risk to the mental health and well-being of children and young people, with much of this content readily available and discoverable through the platform's algorithms.

Pinterest's algorithms continue to recommend harmful content at scale, including through recommendations on the homepage, search results, and via algorithmically curated lists of suggested search terms and related interests.

In December 2022, Pinterest responded to the Prevention of Future Deaths Report issued following the conclusion of Molly's inquest. In its response, Pinterest stated that 'Molly's case has reinforced that depressive content merits careful treatment' and it set out a range of planned mitigations that included:

- **measures to restrict the distribution and recommendation of depressive content**, for example through preventing 'more like these' recommendations in cases where a user views or engages with depressive posts;
- **taking steps to limit the discoverability of depressive content**, including through no longer algorithmically recommending searches for depressive material;
- ensuring that Pinterest **no longer recommends searches for depressive content as autocomplete suggestions,** and;

---

[56] Under questioning from the Russell family's barrister Oliver Sanders KC, Pinterest Head of Community Operations Justin Hoffman was asked if he agreed that Pinterest was not safe when Molly used it in 2017. He replied: 'That's correct, there was content that should have been removed that was not'.

- partnering with a third-party agency to perform **independent testing of Pinterest's moderation efforts** of self-harm and suicide content.

While Pinterest promised to fully implement these measures by December 2023, our research has found limited evidence that improvements had so far been made. The results of any third-party review have not been made public, and it is unclear whether this has actually been carried out.

Algorithmic recommendation of suicide and self-harm content

Pinterest's algorithms still return substantial amounts of suicide and self-harm related material. For example, its 'more to explore' recommendations return a range of problematic material, including posts with captions including: 'life is the art of dying', 'maybe in another life, huh?', and 'memories stay, people don't' (this caption accompanies an image of a person about to walk into a road.)

Suicide and self-harm related content was actively recommended on Pinterest's home page, with a carousel that recommended 'ideas for you', including a video with the caption 'one day I will leave and never come back anymore.'

The homepage also recommended a substantial amount of harmful and problematic posts, including graphic and suggestive images of pools of blood in a sink, blood-soaked towels, and an image of a bloodied hand with the caption: 'mood ☺'.

The homepage recommended a large number of posts that reference suicide ideation and potential methods of suicide and acts of self-harm.

For example, we were actively recommended multiple pictures of people standing on cliff tops, drowning, stylised images of people in freefall through the air, pictures of pills, and an image that appeared to suggest a soul leaving a person's body and ascending upwards.

In some cases, it appears that Pinterest algorithmically recommends content that is likely to be consumed as suicide and self-harm related material (and that directly responds to our interest in it), but that appears to have been posted for a broader range of contexts and purposes, including innocuous use cases.

For example, some posts featuring people standing on the edge of cliffs, and that it is reasonable to conclude had been algorithmically recommended in relation to an interest in suicide, had been posted by a Christian centre for reconciliation and peace. Others had been posted in the context of seemingly motivational quotes.

Examples of highly stylised content that was recommended for us on Pinterest's main page are shown below.
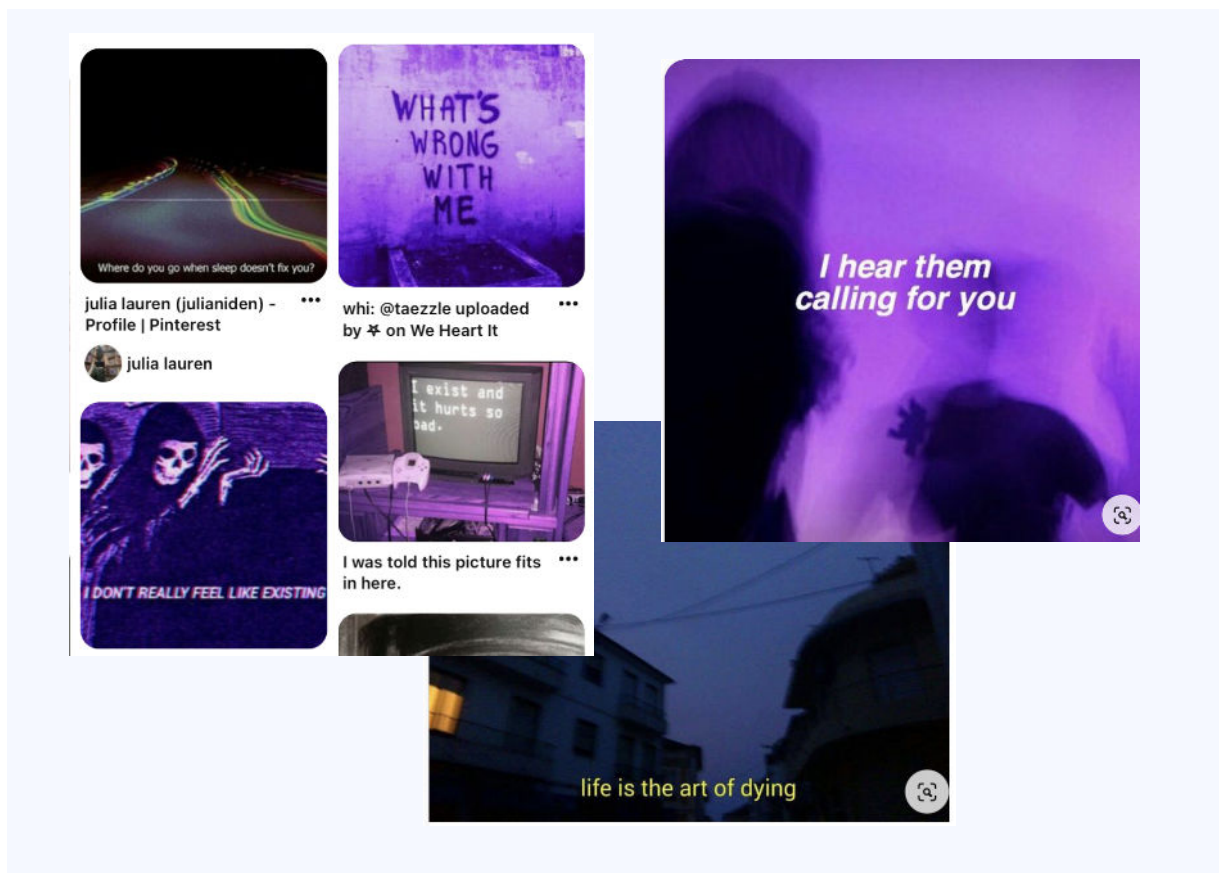
Pinterest's use of high-risk design features

Pinterest continues to recommend harmful material, including suicide and self-harm related content, through a range of poorly conceived design features and user engagement prompts.

For example, through the app's 'updates' feed, we received regular content recommendations, including 'pins inspired by you', 'pins you might like' and content recommendations titled 'your home feed has new pins'.

Each of these routinely contained a substantial proportion of posts that promoted or glorified suicide and self-harm, referenced suicide ideation, or featured themes of misery, hopelessness and depression.

Examples of 'pins inspired by you' included a stylised image of an exit sign, a caption that reads 'alive or just breathing?', and an image of a person who had drowned in the bath, with pills and a knife dropped on the bathroom floor.

Pinterest continues to recommend a substantial amount of harmful material through recommendations of suggested related content.

For example, through recommendations titled 'more to explore' and 'more posts you might like', we were repeatedly recommended a meme of a coffin being carried by skeletons, with an accompanying caption that read 'I don't really feel like existing today.'

Other recommended posts included a stylised image with the caption: 'the person you are looking for no longer exists'; and a picture of a car parked by the ocean with the accompanying captions 'I don't care what happens anymore' and 'you have a decision to make.'

The nature of much of the suicide and self-harm related material recommended to us was highly stylised and visually striking. This is likely to exacerbate the risks resulting from it being algorithmically recommended at scale, with research demonstrating that visual representations of first- and third-person suicide representations can be triggering and potentially desensitising for young people experiencing suicide ideation and/or mental distress[57].

Although it is difficult to determine definitively, it appears that some of the posts we were recommended had been AI-generated. These included scenes showing apparent suicide attempts, with images of a young woman falling backwards from a height. Given the anticipated reduction in the technical and cost barriers to producing artificially generated images, and to produce such images at scale, it appears that Pinterest is poorly placed to respond to a potential rapid growth of such material.

Examples of how Pinterest design features recommended harmful content, often through innocuous themes such as 'aesthetic pictures' and 'pretty pictures' are shown below.
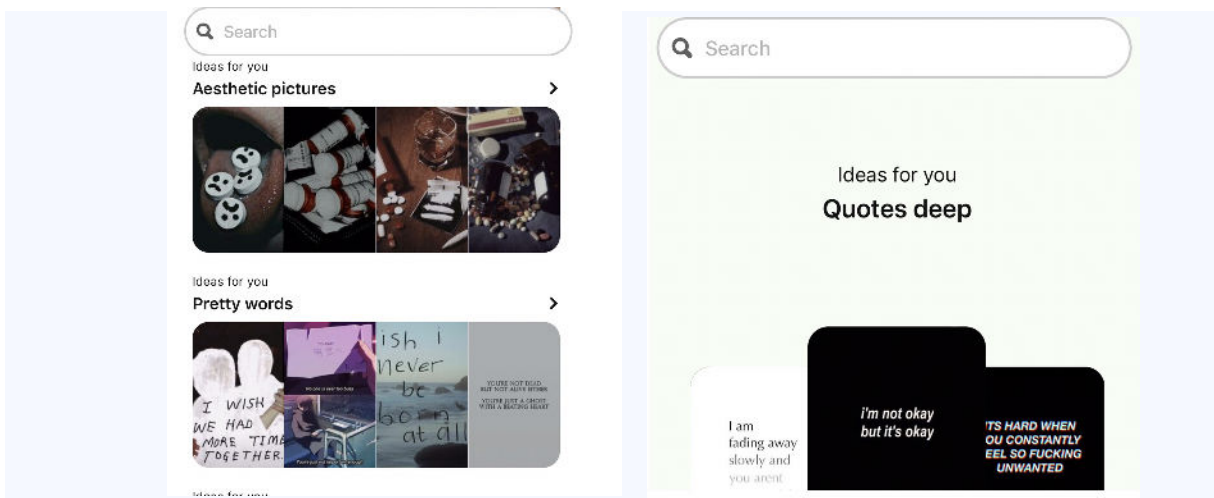
Search terms and depressive content

While Pinterest has blocked some search terms relating to suicide and self-harm, and claims that over 50,000 search terms are featured on its block list, we were still easily able to search for posts that were likely to produce suicide and self-harm related results.

---

[57] Jaroszweski, A et al (2020) First-person stimuli: improving the validity of stimuli in studies of suicide and related behaviours. Psychological Assessment, 32, pp663-676

There is limited evidence that Pinterest has responded to the use of algospeak, with no restrictions on searching for well-understood terms such as 'unaliving' or 'unaliving myself.' Search results for 'unaliving myself' produced recommended search terms such as 'need[ing] a break from life.'

Search results for these terms included content that expressed suicide ideation, with first-person video taken looking over a motorway bridge; posts that contained chains such as 'sleep won't help this sort of tired'; and a range of suicide and self-harm related memes, seemingly not posted by a person in distress, including a video stating: 'I don't want to wake up tomorrow.'

We were recommended substantial amounts of, and readily able to search for, content that referenced feelings of depression, visibility and hopelessness (material that Pinterest refers to as 'depressive content'.)

We were freely able to search the term 'depression' and 'depression quotes', in some cases but not consistently receiving a banner message signposting to advice services.

Top search results for 'depression quotes' included posts that referenced suicide ideation; pictures of a girl drowning in the ocean; an image of a girl about to perform an act of self-harm, with annotations on her arm of how and where to cut to end up in hospital or a morgue respectively; and a seemingly endless amount of content that contained disturbing themes of misery, hopelessness and depression.

There is a clear risk associated with the easy accessibility and discoverability of such a large volume of disturbing and highly depressive material, with harmful content being algorithmically recommended (including through recommended search terms), and users able to save and view it on- demand, including in large volumes and through potential 'binge watching'.

The sheer volume of material, and the ease in which it can be accessed and consumed in a largely frictionless way, appears highly likely to appeal to appeal to, and actively reinforce, a sense of 'perceived burdensomeness' and 'thwarted belongingness' (characteristics identified by Joiner that may drive young people to seek support online but then have these negative feelings and concerns amplified.) [58]

Posts that reference themes such as 'my head is a very dark place', 'I felt so much, that I started to feel nothing' and 'why would they care about you? You're a nobody' were actively recommended alongside, and interspersed with, highly stylised content that references and encourages suicide and self-harm acts, for example posts stating 'I'll use a knife' and 'I hear them calling for you.'

As part of their requirements under the Online Safety Act, Pinterest and other services should be required to assess and mitigate the potential risks posed by cumulative viewing of harmful content, including the interplay with the potential risks posed by how the platform is designed.

---

[58] Joiner, T (2005) Why people die by suicide. Harvard University Press, Cambridge, MA

Platforms should be required to treat the cumulative potential for harm as a reasonably foreseeable risk, with the active possibility of enforcement action if they fail to address design choices that put children at risk.

Further examples of recommended content are shown below.
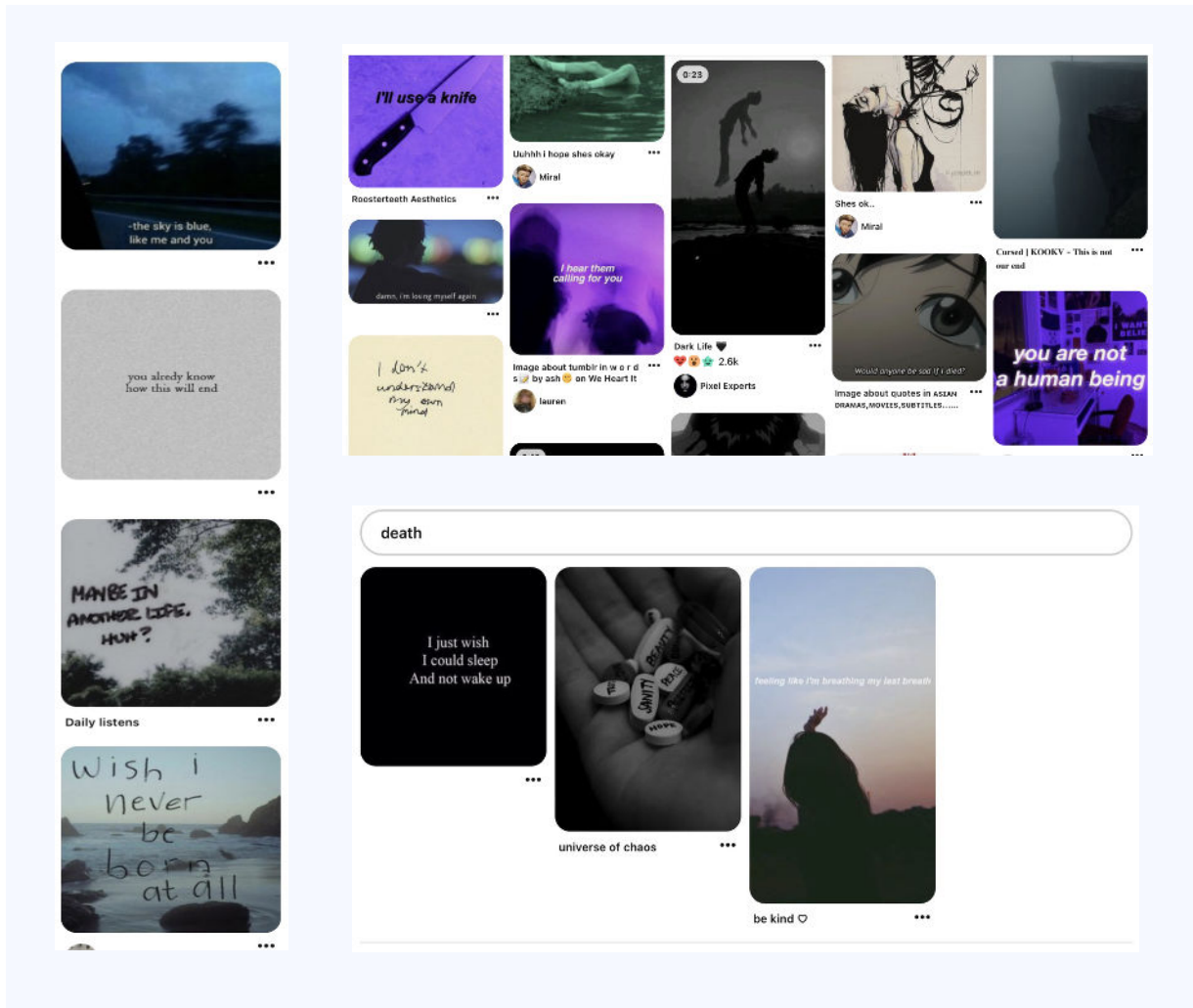
Autocompletes

Pinterest has successfully prevented its search bar from recommending certain search terms as autocomplete suggestions.

However, while terms such as 'suicide' or 'self-harm' have been effectively prevented, a range of suicide and self-harm related terms continue to produce a set of autocompleted suggestions. For example, when we searched for 'unaliving', we were suggested a number of related search terms including 'underlying letters', 'unaliving jokes' and 'unaliving myself.' We

were also presented with a banner encouraging us to 'open [the Pinterest app] for more ideas about unaliving myself.'

The restrictions on autocompleted entries only appears to apply to public search terms, but disturbingly, do not seemingly apply if a user wishes to locate suicide or self-harm related content in their own saved collections. For both 'suicide' and 'self-harm' search entries, we were successfully offered autocomplete entries, along with a thumbnail of a previously saved post.

Given the underexplored but deeply concerning ways in which users can save and curate potentially vast collections of harmful suicide and self-harm material, it is deeply concerning to see little if any concerted attempt to introduce friction into the user experience when searching and retrieving saved posts.

Content with links and relationships to third-party sites

We found substantial amounts of problematic content seemingly posted to cross-promote content hosted on third-party sites, with substantial amounts of harmful material linked to the creative writing platform Wattpad.[59] Harmful content was also posted to drive traffic towards a number of other creative writing and blog sites, including one post that contained the handwritten message: 'maybe I wasn't made for this world.'

Substantial amount of harmful content appears to have been originally produced and shared on other social media platforms, particularly TikTok.

Examples include fast, subliminal-like videos, with several posts featuring a montage of suicide and self-harm related images; so-called 'vent accounts', in which users express thoughts such as 'I don't wanna wake up tomorrow'; and a number of posts seemingly shared on meme accounts, including a TikTok video with the caption 'POV: no one thinks you'll actually do it', shared with the accompanying message 'want me to prove it by attempting?'

Pinterest actively encourages users to share pins on third party sites, including through an animated WhatsApp logo that offers posting links to a number of messaging and social media apps, including WhatsApp, Messenger, iMessage, Facebook and X.

While further research is needed to understand cross-platform consumption habits relating to suicide and self-harm content, this design feature significantly reduces the friction associated with the sharing and discovery of potentially harmful material; and as a result, further increases the risk profile associated with it.

---

[59] Wattpad has previously been linked to poorly-moderated suicide and self-harm material, including in the case of Frankie Thomas who died by suicide aged 15 after reading harmful content on the site.

Next steps and recommendations

This report has demonstrated the substantial prevalence, and disturbing nature, of harmful content that remains freely accessible and discoverable on major social media sites.

It has also set out the range of ways in which poorly conceived platform design choices, including algorithms, amplify the exposure of children and young people to content that promotes or glorifies suicide and self-harm; references suicide or self-harm ideation; or that contains distressing and often relentless themes of hopelessness, depression and misery.

Each of these content types has the potential to cause harmful effects, particularly when viewed in large amounts or on a cumulative basis, with the most substantial risks borne by children and young people experiencing suicide ideation, thoughts of self-harm or poor mental health.

Our findings present a compelling case for swift and ambitious regulatory action, including through the UK's Online Safety Act and the EU's Digital Services Act. It also underscores the urgency of legislative action in the United States, including the swift passage of the Kids Online Safety Act.

While we can't say with confidence the extent to which effective online safety regulation - and in turn the unlocking of safer online environments - may translate into reductions in deaths by suicide, suicide ideation and self-harm behaviours, the scale and nature of inherently preventable technology-facilitated harm is clear for all to see. It underscores the importance of a bold and ambitious regulatory response.

Recommendations

This report identifies four priority areas where measures should be taken to ensure a clear and comprehensive response to the risks of technology-facilitated harm, including the risks posed by suicide and self-harm related harmful content.

1. Delivering the potential of the Online Safety Act

The UK's online safety regulatory regime is a crucial opportunity to reset the risk profile associated with social media sites; and for the regulator Ofcom to ensure that tech companies are incentivised and actively required to prioritise safety-by-design in the design and delivery of their products.

If the regulatory regime is to succeed, Ofcom must:

- **adopt a bold and ambitious regulatory response, with codes of practice and regulatory guidance that respond appropriately to the scale and complexity of the risks posed by harmful content.** The regulatory scheme must be capable of tackling the systems and processes through which harmful forms of suicide and self-harm related are readily amplified;
- **require companies to assess and mitigate the risks posed by their platforms through a broad-based and systemic approach.** Companies should be required to adopt a robust and demanding risk assessment framework that emphasises the risk posed by the interplays between harmful content, high risk design choices, and algorithmic amplification;
- **underscore the importance of tackling the risks posed by cumulative exposure to harmful material**. When it publishes its risk profiles, Ofcom should set out the range of harm archetypes that companies should tackle, including the risks posed by isolated exposure to harmful content, active engagement with potential online hazards, but also the longer-term cumulative risks associated with exposure to harmful content;
- **commit to active supervision of and enforcement with the regulatory regime**. Companies must no longer be able to hide behind the complexity associated with suicide and self-harm material. Ofcom must actively supervise the quality and consistency of how platforms moderate harmful content; track the quality of platform risk assessments and the efficacy of implementing required improvements; and demonstrate a willingness to prioritise enforcement action, particularly in cases where companies continue to expose children and young people to preventable and reasonably foreseeable harm.

2. Ensuring open access to data

If civil society is to be able to effectively support the regulator, and to continue its crucial work in holding companies to account, regulators and governments must stand ready to protect and expand the ability of civil society bodies and researchers to access relevant data, including that held by social media companies.

We recommend that:

- **Ofcom recognises the importance of civil society and researcher access to data**. As part of its requirement to publish a report within two years of Royal Assent, the regulator should set out clear and comprehensive measures to safeguard and expand data access for academia, civil society and the tech accountability sector;
- **Ofcom and Government commit to resolving barriers to effective civil society research**. Civil society faces a range of challenges in being able to support the regulator effectively, including funding and capacity challenges and ongoing barriers to data access. The Government and Ofcom should identify and develop an action plan to overcome such challenges, and ensure the broader regulatory settlement, with civil society at its core, is capable of functioning effectively;

- **Ofcom signals the importance of unimpeded civil society activity and research.** As part of its ongoing supervisory approach, Ofcom should signal to tech companies that it strongly discourages any efforts to prevent or frustrate civil society research. This should include but not be limited to measures that may reasonably have the effect of making research projects cost prohibitive; seeking to prevent legitimate research through amending terms and conditions; and/or through imposing technical or other restrictions to frustrate legitimate methodologies and research approaches.

3. Building the evidence base on harmful online content

The evidence base around the nature and effects of exposure to harmful online content remains highly nascent, yet this is the basis upon which decisions on the future regulatory framework will be taken.

Our understanding of the risks posed by social media has arguably been frustrated by the clear incentives for social media companies to refuse to share their data. In a seeming attempt to skew the evidence base, some companies have opted to grant selective and arguably preferential access to data sets, with privileged access for some closely aligned and largely favourable researchers.

In order to improve our understanding of the risks associated with harmful online content, and neutralise the risks associated with industry influence, we recommend that:

- **Ofcom establishes a clear set of research priorities**, with a particular focus on the mechanisms through which exposure to suicide and self-harm content may result in immediate or longer-term risks. This should include funding and supporting studies that explore the potential cognitive and behavioural effects;
- **Ofcom recognises the importance of adopting a precautionary principle approach**. This is particularly important given the lag time between new product launches and academic publications that can study their effects;
- Ofcom, civil society and academia develop **new mechanisms to share research findings, data and other emerging insights more effectively**;
- **Civil society plays a crucial role in assessing the ongoing potential risks and impacts**, including rolling analysis of the impacts of new product features on the risk profile of social media sites.

4. Strong action by tech companies

Tech companies do not have to wait for regulation before addressing reasonably foreseeable safety risks. Social media companies should be ready to commit to immediate action to

address the risks posed by harmful content, including how their design choices and algorithms contribute to both immediate and longer-term risk profiles.

We recommend that:

- **Social media companies urgently invest in the development and implementation of new safety and well-being initiatives**, with a demonstrable commitment to report on their effectiveness;
- platforms prioritise **measures to address the safety risks associated with algorithms and other high risk design features**, building on YouTube's early leadership in this space. Tech companies should recognise and respond to the substantial risks associated with children and young people being encouraged to view suicide, self-harm or highly depressive content in large amounts or on a cumulative basis;
- **platforms should commit to improving the quality and relevance of their voluntary transparency data**. For example, Meta could usefully shift from reporting data on the exposure of suicide and self-harm content at a general population level, towards much more useful metrics such as the exposure of teenage girls to material that promote or glorify suicide and self-harm, that references suicide ideation, or contains intense themes of depression, hopelessness and misery;
- tech platforms should **work in partnership with civil society to develop industry-wide best practices**, with investment in the delivery of a more efficacious, targeted and transparent approach to harm reduction.

Appendix one:

Focus hashtags used for each platform

TW: this list contains hashtags used to distribute harmful content, including material that references suicide, self-harm and highly depressive content.

Instagram

#_iwannakillmyself

#selfharmmm

#suicidall

#iwantotdie

#letmedie

#depressionquote

#sucidalmind

TikTok

#drained

#depressionquotes

#icantdothisanymore

#iwanttoleave

#iwanttoendit

#paintok

#ventaccount

#SVV

Appendix two:

analysis of search results for Instagram hashtags containing suicide and self-harm material

Hashtag	Links offered to support resources	Friction built into the search experience e.g. 'click through to see results'	Search results	Search results
#suicidall	X	X	- Results freely available - 'for you page' recommends algorithmically curated results - Recommended list of related problematic hashtags	- FYP recommends a compilation video of suicide pact and methods - multiple violative videos that promote or glorify suicide and self-harm - posts include 'I'm not scared or suicide' and 'my eyes ran out of tears so now my wrists cry blood'
#selfharmmm	✔	✔	- Click through required to access results - recommended list of related hashtags	- Top posts recommend violative videos that promote or glorify self-harm
#letmedie	X	✔	- Click through required to access results	- FYP recommends videos that promote or glorify suicide and self-harm - user prompt to 'use the hashtag' by taking and uploading a photo (with link to open camera)
#secretsociety123	✔	✔	- Click through required to access results	
#iwannakillmyself	✔	✔	- Posts blocked	
#iwanttodie	✔	✔	- Click through required to access results - Recommended list of accounts and related hashtags - Posts blocked	- Recommended accounts contain harmful content
#depressionquotes	✔	✔	- Click through required to access results - 'for you page' recommends algorithmically curated results - Recommended list of related hashtags	- Results include a large amount of material expressing hopelessness, misery and worthlessness
#suicidalmind	X	X	- Results freely available	- 'top posts' include posts and videos referencing suicide ideation and the promotion of self-harm

Note: this analysis was undertaken in late October and early November 2023. We have observed that some hashtags or search functionality were temporarily unavailable during parts of mid-November, but were subsequently restored, the reasons for which were not readily discernible.