MOLLY ROSE FOUNDATION

the bright initiative | by Bright Data

# Pervasive-by-design

Suicide, self-harm and intense depression content on TikTok and Instagram, and how their algorithms recommend it to teens

# Contents

# Summary

**Almost 8 years after Molly Russell's death, how much has really changed?**

In this report, we set out to explore and answer that question. Our detailed analysis suggests that the risks of exposure to harmful forms of suicide, self-harm and intense depression-related content on two major social media platforms, Instagram and TikTok, remain unacceptably pervasive and widespread.

At Molly's inquest, the Senior Coroner found that exposure to harmful content on social media played a 'more than minimal' contributory role in her death.[1] Molly's inquest marked the first time that social media platforms have been held directly responsible for the death of a child.

The circumstances of Molly's death are far from unique. Technology plays a role in almost one-quarter (24%) of deaths by suicide among young people aged 10 to 19,[2] equivalent to a lost young life every week.[3]

This research aims to provide a comprehensive assessment of the nature and prevalence of harmful content, and to assess the platform design choices that allow suicide, self-harm and intense depression material to be recommended to and accessed by children and young people at scale.

## What we found

In the immediate period before regulation took effect, considerable volumes of harmful suicide, self-harm and intense depression material continued to be readily accessible and discoverable on both TikTok and Instagram, with each of the platforms algorithmically recommending substantial amounts of harmful posts to under 18s.

With regulatory enforcement of the Online Safety Act now underway, our findings point to a deep and systemic set of failures in how these major social media platforms are designed and run – and raise substantive questions about the likely efficacy of Ofcom's current approach.

The results show that risk profile on both platforms has not improved, and in TikTok's case it has actively deteriorated, since we undertook previous comparable research in 2023.[4]

1 Prevention of Future Death for Molly Rose Russell, issued 13 October 2022
2 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
3 MRF analysis, applying Rodway et al's modelling to the most recent Office for National Statistics on death registrations in England and Wales
4 Molly Rose Foundation (2023) Preventable yet pervasive: prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest

Our research finds that:

- **TikTok and Instagram continue to algorithmically recommend substantial volumes of harmful content to teens:** TikTok and Instagram's personalised recommender systems continue to suggest overwhelming volumes of harmful forms of suicide, self-harm and intense depression content to teens.

  On accounts opened with a registered age and identity of a 15-year-old girl, which had previously engaged with suicide, self-harm and depression material, almost all (96%) of the algorithmically recommended videos watched on TikTok's For You Page contained content that was likely to be harmful, particularly when viewed cumulatively or in large amounts.

  97% of Instagram short-form videos (known as Reels) contained themes likely to be harmful, particularly when being recommended and consumed in large amounts.

  The findings suggest a fundamental lack of safeguards were still not in place on either TikTok and Instagram, and that in the immediate period before regulation took effect, children could still be exposed to a substantial risk of reasonably foreseeable but preventable harm.

- **High-risk design choices continue to exacerbate the potential for cumulative harm:** TikTok and Instagram continue to use a range of user interfaces, recommended search terms, and other high-risk design features that seek to maximise user engagement and dramatically increase the risk of exposure to cumulative harm, the risks that result from harmful content being recommended and consumed in large volumes over time.

  TikTok has significantly increased its risk profile by introducing new AI-generated search prompts which regularly appear when a user watches or engages recommended content on the platform's For You Page. Our analysis found this feature enables users to immediately and more readily than ever enter 'rabbit holes' that significantly increase the risk of exposure to immediate or longer-term forms of cumulative harm.

  During the research, this feature extensively recommended topics relating to extreme suicide, self-harm and depression material; topics which encouraged and romanticised suicide among teens; 'rabbit hole' topics ranging from self-harm tutorials to meme-style content promoting suicide methods; and a near-constant barrage of topics exacerbating feelings of intense misery, hopelessness and despair.

- **Algorithms continue to enable harmful content to achieve extraordinary levels of reach:** on both TikTok and Instagram, personalised recommender systems and other high-risk design choices enable harmful content to achieve remarkably high levels of reach and user engagement, substantially increasing the potential for exposure to cumulative harm.

  On TikTok, the platform's algorithms enable harmful material to reach staggeringly high audiences, with harmful posts that were algorithmically recommended on the For You Page liked by an average of over 226,000 accounts. One in ten harmful posts (9%) had been liked at least one million times.

We also found evidence that more extreme forms of harmful material, including content that promoted suicide and self-harm or that explicitly referenced suicide methods, were three to eight times respectively more likely to appear in our For You Page than when analysing the most popular content publicly accessible and discoverable through hashtags.

- **Both platforms appear to be 'gaming' their requirements under the Online Safety Act:** During our analysis of both Instagram and TikTok, we found evidence that the platforms have been starting to introduce new design features to comply with their requirements under the Online Safety Act – but that the companies actively appear to be gaming some of the measures required of them.

  In essence, it appears that both Instagram and TikTok are seeking to 'game' the lack of specificity in Ofcom's requirements by introducing a version of a new mandated feature that enables younger users to offer negative feedback on content being algorithmically recommended to them, but that also goes one step further – by allowing users to also offer *positive* feedback about the content they're being shown.

  During our research, this meant we were able to repeatedly offer 'positive' feedback about harmful forms of content, including algorithmically recommended Reels and For You Page content that referenced suicide ideation, suicide behaviours, and multiple examples of content that reinforced and normalised intense feelings of hopelessness, misery and despair.

  After submitting positive feedback, the user is informed they will be recommended additional amounts of similar content for the following 30 days.

## Implications and next steps

Our findings show that deep and pervasive problems remain in the ways that TikTok and Instagram design and run their products. In the immediate period before regulation took effect, there were evident and widespread systemic failures in how both platforms were protecting children and young people from the palpable risks of harmful content.

The research deepens existing concerns about the likely impacts of Ofcom's existing regulatory approach, which has now taken effect, with a substantial disconnect between the magnitude of high-risk design features identified in this research and the limited scope and ambition of the measures being set out to address them.

This research should send a powerful message to platforms, regulators and the UK Government that they need to do better.

- **Social media platforms must no longer allowed to pay lip service to safety-by-design**, while their designers and engineers remain actively incentivised to introduce increasingly insidious new engagement-based features that will inevitably deepen and extend the risk profile for young people.

- **Ofcom must substantially revisit the scope and ambition of its regulatory approach.** Despite the regulator's superficially impressive claims that its Protection of Children measures will 'tame toxic algorithms',[5] its detailed codes suggest that tech companies will only need to invest £80,000 in one-off costs to fix the algorithms that cost Molly's life[6] – and that continue to threaten many more. Our research shows that Ofcom's current approach is clearly poorly placed to address the scale and magnitude of harm on major platforms; and that platforms are already successfully 'gaming' the ambiguity in its initial approach.

  The regulator's ambition continues to be dwarfed by the magnitude and persistence of high-risk design features, underpinned by platform business models that actively prioritise user engagement over meaningful improvements to safety outcomes.

- **The findings must also increase pressure on the UK Government to strengthen the Online Safety Act.** While the Technology Secretary has described the Act as 'uneven and unsatisfactory',[7] the Government has been slow to commit to fixing systemic weaknesses in the Act's design, and as a result its overall potential impact, for primarily structural reasons, continues to be substantially curtailed.

- **Advertisers must also be prepared to step up and do more to ensure that social media platforms are made fundamentally safer-by-design.** During our research, we were repeatedly shown adverts adjacent to harmful material, including ads from multiple well-known brands, at the rate of one in every 9.5 videos watched on TikTok's For You Page and 1 in 10 on Instagram Reels. Advertisers must be prepared to use their substantial leverage to demand better from tech companies – managing reputational risks for them, while simultaneously delivering meaningful improvements for young people.

Almost eight years after Molly's death, our research shows social media platforms continue to expose vulnerable young people to entirely preventable online risk. Our message to the Prime Minister is therefore clear – it is time to act, decisively, to strengthen the UK's regulatory regime.

Every month we wait means that more young lives will be lost.

5 Ofcom press release issued to coincide with its Protection of Children Code of Practice in early 2025
6 Ofcom's Protection of Children Codes Volume 4 specify that the measures to prevent Primary Priority Content appearing in children's feeds are likely result in one-off costs of £80,000, and annual ongoing costs ranging from £3,000 to £20,000 thereafter.
7 Comments made on Sunday with Laura Kuenssberg on 12 January 2025, responding to comments made by Ian Russell on the same programme in which he described Ofcom's implementation of the Online Safety Act as a 'disaster.'

# Aims and methodology

This research assesses the availability and prevalence of harmful suicide, self-harm and intense depression content suicide on TikTok and Instagram. MRF previously undertook similar research in November 2023.[8]

This updated analysis aims to provide a detailed understanding of the scale and nature of these types of harmful content, prior to Ofcom's Protection of Children Code coming into effect; and to better understand the design choices and user engagement features that result in harmful content being recommended to and discoverable by children and young people.

In this project, content was analysed and classified as likely to be harmful if it promoted or glorified suicide and self-harm; referenced suicide methods; or if it referenced suicide ideation, self-harm ideation, or themes of intense hopelessness, misery and despair in a way that poses an increased risk when watched cumulatively or in large volumes, for example because of recommender algorithms or other high-risk design choices.

A three-stage methodology was undertaken:

1. **Scraping and analysing of harmful content:** In this first stage of the project, we analysed content from both Instagram and TikTok using hashtags that have been frequently used to post suicide, self-harm and depression-related material.

   Content was collected over several months, from November 2024 to March 2025. The most engaged with posts from TikTok (514 posts) and Instagram (450 posts) were analysed and coded to determine if they contained potentially harmful content.

   Analysis was undertaken on accounts registered as a 15 year-old UK girl.

2. **Analysis of algorithmically recommended Reels and For You Page posts:** In the second stage of the project, 644 short-form videos were algorithmically recommended by the Instagram and TikTok recommendation engines and were watched consecutively. Analysis was undertaken in May and June 2025. This included 300 Instagram Reels and 342 TikToks viewed via the platform's For You Page.

   Posts were once again analysed and coded to determine if they were potentially harmful, using the coding framework set out above, with the first stage of the project having enabled the platform's algorithms to be 'trained' on our interest in these topics.

3. **Analysis of platform design choices and engagement features:** We also undertook extensive qualitative analysis to assess the broader set of ways in which harmful content can be recommended and consumed on these platforms. This included an in-depth analysis of the relationship between platform design choices and exposure to harmful content; and the role of user engagement features and other prompts in the discovery and consumption of harmful posts.

8 Molly Rose Foundation (2023) Preventable yet pervasive: prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest

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

This research does not claim to represent what a typical child with interests in suicide, self-harm or intense depression content will necessarily be recommended, in so much as that is possible – for example, it has not sought to mimic teenage usage and behaviour patterns, nor did we actively leave comments on posts or send direct messages to other users.[9] However, the research is an accurate reflection of the content that Instagram and TikTok's core recommendation engines and user engagement features proactively recommended or facilitated discovery of, and as such is a robust methodological approach to test the design features and safeguards built into each platform's product.

We excluded livestream recommendations when undertaking our analysis of TikTok For You Page content.

9 We have drawn on the ethical framework suggested by Revealing Reality in their assessment of avatar methodologies undertaken by Ofcom. Revealing Reality (2023) avatar methodology: pilot study. A report for Ofcom

# Trigger warning and access to 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 readers.**

**If you need it, support is available from the following UK-based helpline services. Each of these support resources are 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**

# Context

## The scale and impact of technology facilitated suicide and self-harm

Technology facilitated suicide and self-harm are serious public health issues affecting children and young people.[10]

There is growing 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 youth suicide cases.[11] Suicide-related Internet use has been reported in 24% of deaths by suicide among young people aged 10 to 19, equivalent to a young life being lost every single week.[12]

Suicide related Internet use is recorded more frequently in the death by suicide of girls, and in cases affecting adolescents identified as LGBTQ+.[13] In recent years, the suicide rate among women and girls has risen noticeably.[14] Recent official data suggests that gay, lesbian and bisexual people are 2.2 times more likely to die by suicide than people identifying as straight.[15]

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.[16] Suicide and self-harm related Internet use was reported in 26% of child hospitalisations relating to self-harm, with self-harm considered a major risk factor for suicide among adolescents and young adults.[17] Recent figures suggest that young people aged 16 to 24 identifying as gay, lesbian or bisexual have experienced a larger relative increase in the risk of intentional self-harm than all other age groups.[18]

## The harmful effects of exposure to suicide, self-harm and intense depression content

Findings from multiple studies have raised concerns about the harmful effects of exposure to suicide, self-harm and intense depression-related content. Recent research found that suicide-related online experience is a 'common but likely underestimated antecedent' to suicide in young people.[19]

10 Department of Health and Social Care (2023) Suicide Prevention in England: Five Year Cross-Sector Strategy
11 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
12 MRF analysis, applying Rodway et al's modelling to Office for National Statistics on death registrations in England and Wales
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
14 ibid
15 Office for National statistics (2025) Self-Harm and suicide by sexual orientation, England and Wales: March 2021 to December 2023
16 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
17 Padmanathan, P (2018) Suicide and Self-Harm Related Internet Use: a Cross-Sectional Study and Clinician Focus Groups. Crisis, 39(6), pp469–478
18 Office for National statistics (2025) Self-Harm and suicide by sexual orientation, England and Wales: March 2021 to December 2023
19 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

While further research is needed to determine the strength of any causal relationship, and suicide and self-harm material has been found to have both harmful and protective effects, a recent landmark systemic review found that harmful effects currently predominate.[20]

The harmful effects of engaging with suicide, self-harm and intense depression-related content may include (but are likely not limited to):

- increases in the frequency and/or severity of suicide and self-harm ideation and behaviours. 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 Instagram;[21]

- 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;[22] while engaging with material may result in a range of emotional, cognitive and physiological impacts, for example increased rumination, which may go on to trigger or exacerbate themes of despair, hopelessness and in some cases suicide and self-harm ideation;

- Algorithms and other engagement-based design features may result in an 'assortative relating' effect, with young people experiencing suicide ideation or thoughts of self-harm being more likely to identify and build relationships with other users experiencing similar actions and thoughts.[23] While technology-facilitated peer support can offer powerful protective effects, it also presents significant risks, particularly where feelings of despair, hopelessness or poor mental health can become unhealthily normalised;

- Children may actively search for and/ or consume harmful content as a maladaptive coping response. Many young people lack much-needed skills in critical online and algorithmic literacy,[24] with previous research finding that some teenagers may view the amount and type of harmful content being recommended to them as a proxy for their mental health;[25]

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

- Online communities of practice may form, which in some cases may offer protective effects but that can also result in young people experiencing an exaggerated perception of suicide, self-harm and intense depression behaviours; a risk that self-harm is seen as normalised or beneficial coping mechanism; and a risk that these online cultures may inadvertently preclude offline or expert oriented forms of help seeking (cementing a sense that those who do not self-harm ''cannot understand.')[26]

20 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
21 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
22 ibid
23 Ibid
24 Gagrcin, E. et al. 2024. Algorithmic media use and algorithmic literacy: An integrative literature review. New media and society.
25 Molly Rose Foundation (2023) Preventable yet pervasive: prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest
26 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

## The impact of cumulative harm and implications for platform design

In recent years, there has been increasing concern about the risks associated with viewing large amounts of harmful content, typically driven by high-risk platform design choices.

Social media platforms have typically designed and run their services to prioritise user engagement and user growth, with the success of key products assessed by standardised metrics, including whether it leads to increased time spent on a company's product.

The introduction of engagement-based design choices, including Instagram and TikTok's personalised recommender systems, have resulted in some young people being effectively bombarded with large amounts of harmful content ('cumulative harm'). Molly's inquest heard that she had viewed over 2,000 items of harmful content on Instagram in the months before her death, with a pronounced increase in the volume and severity of the material being shown.

Platform algorithms effectively serve as the start of well-established harm pathways, with teenagers being recommended large amounts of harmful content and then quickly able to amass large catalogues of saved material. This can readily facilitate 'binge watching', a maladaptive coping response often seen among vulnerable teens and young adults being exposed to harmful content.[27]

Effective legislative and regulatory solutions will address cumulative harm through the adoption of a two-fold approach. Firstly, platforms must be actively incentivised to shift away from an emphasis on user growth and commercial metrics towards a more balanced approach that emphasises both product safety and safety- and wellbeing-by-design.

Secondly, platforms will need to move away from their current narrowly defined understanding of harm, assessed at the content level and primarily operationalised by their terms of service, towards a more dynamic and user-centered understanding of risk, in which platform and regulatory architectures are required to assess and mitigate harm through a nuanced understanding of how, when and to whom content is being shown, rather than simply what it contains.

27 Research has found that binge watching is significantly associated with depression, loneliness, sleep problems, anxiety and stress. Stronger associations between binge watching and two types of mental health problems, depression and sleep problems, were observed during the pandemic. Alimoradi, Z et al (2022) Binge Watching and Mental Health Problems: a Systematic Review and Meta-Analysis. Int J Environ Res Public Health, 19(15)

# FINDINGS

## Instagram: Key findings

Our analysis of Instagram found that the platform remains exceptionally high-risk for children and young people, with its algorithms recommending a near-constant supply of problematic suicide, self-harm and depression material.

While the platform has made some welcome strides to make it harder for young people to proactively search for harmful content, Instagram's personalised recommender systems continue to algorithmically recommend appallingly high volumes of harmful material, and to expose its teenage users to unacceptable levels of inherently preventable harm.

We found that:

- on an account which had previously been used to access suicide, self-harm and intense depression content, 97% of recommended short-form videos were classified as harmful – a near-continuous barrage of harmful posts that would inevitably increase the risk of cumulative harm to young users.
- harmful short-form videos were achieving deeply concerning levels of virality and reach, with one in five posts being liked over 250,000 times. Harmful posts are also being shared extensively, with algorithmically recommended harmful Reels being privately shared 55,227 times.
- Instagram appears to be actively gaming its requirements under Ofcom's Children's Code.[28] While, as required, it has introduced a new design feature introduced that enables young users to provide negative feedback on content being recommended to them, this also gives them the option to provide positive feedback, and in turn to be recommended similar types of content for the following 30 days, including suicide and depression-related material.
- There appears to be an asymmetrical response to harm across Instagram's multiple product surfaces. While Instagram has taken some positive steps to make it harder for users to proactively search for and discover harmful posts using hashtags, there is a striking lack of emphasis on the more commercially strategic parts of its product, including where content is being algorithmically recommended on its Reels page, homepage and Discover tab.
- We were substantially more likely to be algorithmically recommended harmful content on Reels than on any other part of the platform. While this may in part reflect the technical challenges associated with moderating short form video content, it may also reflect a commercial decision to grow the product user base, at the expense of user safety, driven by a race for market share. We therefore strongly encourage Ofcom to more substantively investigate the relationship between platform risk profiles and their business models, and to fully use its transparency and information disclosure powers to support this crucial upstream focus.

28 Measure PCU E3 in Ofcom's Protection of Children Code

# Prevalence and reach of algorithmically recommended harmful content

## Algorithmic recommendation of harmful short-form videos (Reels)

In recent years, Instagram's short form video product, Reels, has become a core part of its product. Meta claims that Reels now accounts for over half of all user time spent on the app.[29] Our research finds that harmful content is significantly more prevalent on Reels than any other part of the platform, with a staggeringly high proportion of videos containing harmful content, including suicide, self-harm and highly depressive material.

As part of the research, we analysed 300 short-form videos, each of which were algorithmically recommended and watched consecutively on the Reels tab. This stage of the research was undertaken after previously accessing suicide, self-harm and highly depressive content for earlier stages of the project, enabling the algorithm to understand and respond to our interest in these topics.

Disturbingly, almost all of the videos we were algorithmically recommended (97%) were found to be harmful. Videos were classified as harmful if they either promoted or glorified suicide or self-harm; referenced suicide or self-harm ideation; or otherwise featured highly intense themes of hopelessness, misery and despair.

While some videos were ostensibly harmful in their own right, in most cases videos were classified as harmful because their contents were likely to likely to exacerbate the risk of cumulative harm to users.

Examples of harmful content included a large quantity of videos that exacerbated feelings of intense misery, hopelessness and despair; videos that appealed to or were actively likely to reinforce a sense of 'perceived burdensomeness' and 'thwarted belongingness';[30] and a near-uninterrupted barrage of intense depression material and posts which seemed likely to increase the risk of unhelpful rumination and other adverse cognitive and physiological effects among young users.

## The limited impact of Teen Accounts

In September 2024, Instagram announced the rollout of Teen Accounts, which the company claimed would offer under 16s a safer and more age-appropriate experience on the platform.[31] As a result of the changes, Instagram claimed that children would no longer be algorithmically recommended certain types of material, including suicide and self-harm content.

Our findings suggest that Teen Accounts have had a modest overall impact on the volumes of harmful content being algorithmically recommended to teens, with only a slight reduction in the proportion of harmful content being algorithmically recommended overall.

29 Comments made by Mark Zuckerberg in Meta's Q2 2024 earnings call. Time spent watching Reels as a proportion of all time spent on Instagram has more than doubled since Q1 2022, when it accounted for just over 20% of time spent.
30 Characteristics which are observed antecedents in deaths by suicide. Joiner, T (2005) Why people die by suicide. Harvard University Press, Cambridge, MA
31 Announcing the launch of Teen Accounts, Sir Nick Clegg, Meta's President of Global Affairs at the time, said: 'this is a very significant moment […] This really tries to very meaningfully shift the balance.'

Compared to our previous analysis in 2023, the overall proportion of harmful content we were recommended on the Reels tab fell by only two percentage points, from 99% to 97%. Clearly by any reasonable definition this volume remains stubbornly and unacceptably high.

As table 1 shows, we did observe that Teen Accounts appear to have driven a shift in the *types* of harmful content being algorithmically shown, with a decline in the proportion of posts that promote or glorify suicide and self-harm, and a substantial drop in videos referencing suicide ideation. However, the proportion of videos referencing suicide methods increased, from 14% to 20% of recommended content.

**Table 1: prevalence of harmful content being algorithmically recommended on Reels**

| Content type | 2023 research | 2025 research |
|---|---|---|
| Promotes and glorifies suicide or self-harm | 13% | 8% |
| References suicide methods | 14% | 20% |
| References suicide and/or self-harm ideation | 55% | 44% |
| Contains feelings of intense depression, misery and despair | 99% | 97% |
| Total content likely to be harmful | 99% | 97% |

n=300 (2025), n=100 (2023)

Notably, we saw almost no change in the proportion of algorithmically recommended videos containing themes of deep and intense misery, hopelessness and despair. Given the adjacency of these videos to posts that reference suicide methods, suicide ideation and self-harm ideation, we remain deeply concerned about the potential of this content to trigger a range of adverse outcomes, including suicide and self-harming behaviours and ideation in teens, as well as depression, anxiety and mental health conditions.

Earlier this year, MRF raised concerns about the impacts of Meta's decision to cease proactive detection of many types of harmful content.[32] At the time, Meta gave public assurances that proactive detection of violative suicide and self-harm material would continue, but it declined to give similar assurances in respect of intense depression and despair-related content. We call upon regulators to actively explore whether this policy position has directly contributed to the continued high volume of harmful short-form video content being algorithmically recommended to teens on the platform.

We also strongly encourage Ofcom to reconsider its decision not to require platforms to use proactive technology to identify and action all forms of Non-Designated Content, including depression-related posts.[33] At present, Ofcom will require platforms to prevent non-designated content from being shown in children's feeds where it has been identified, but the regulator does not intend to set measures that actually require platforms to proactively detect it to a required standard in the first place.

32 in a letter to Ofcom, we voiced deep concerns about Meta's 'bonfire of safety measures' and called on the regulator to fast-track additional measures to address the resulting risks in its Protection of Children Code. The regulator declined this request.
33 Ofcom's consultation on additional safety measures, published on 30 June 2025 from

## The reach of algorithmically recommended Reels

Our research has found that Reels containing harmful suicide, self-harm and highly depressive content achieve deeply concerning levels of reach and engagement. Among the harmful videos we analysed, one in five had been liked more than 250,000 times, with each harmful video attracting an average of over 226,000 likes.

While the overall reach of relevant harmful content is less than we observed on TikTok, the findings are no less disturbing for it; and clearly powerfully underscore the extraordinary and seemingly unchecked velocity and virality of Instagram's highly personalised recommendation algorithms.

Harmful videos also generate substantial amounts of comments and shares. Each post received an average of 3,340 comments, with many videos serving as a de facto but poorly moderated discussion space for users, most of whom were children. Discussions typically centered on suicide ideation, thoughts of self-harm, and expressions of deep despair and dismay (typically being reinforced by dozens and in some cases hundreds of other users.)

We are also particularly concerned about the large volume of posts being privately shared, with each harmful post being privately shared on average 55,227 times.

While there are several potential reasons why harmful posts are being shared at such high levels, this finding suggests the presence of active communities of practice on Instagram, formed around common interests of suicide, self-harm, despair and depression. It appears that these communities of practice are not only passively consuming harmful content, but also actively sharing and exchanging it in effectively unmoderated end-to-end encrypted spaces, including direct messages and groups.

While the dynamics and risk implications of sharing and discussing harmful material in private messages remains only partially understood, there is a clear case for regulators and lawmakers to adopt a presumptive precautionary approach to the resulting potential risks. These risks may include but not be limited to a broad range of adverse psychological and cognitive effects, including an exaggerated and overly normalised sense of collective despair ('everyone feels the same'), increased levels of rumination, reduced aversion to harmful imagery, and for the routine sharing of harmful content to become firmly entrenched as a maladaptive coping response.

The regulator must also remain live to the risks that these 'communities of practice' can be readily infiltrated and exploited by threat actors, including the growing risks associated with children being identified for the purpose of suicide and self-harm related grooming. Over the last 18 months, multiple law enforcement agencies including the NCA,[34] FBI,[35] and Europol[36] have warned about the growing threat posed by sadistically-motivated malign actors, including Com groups.

There is growing confidence that threat actors actively target social media, chat forums and other user-to-user spaces to identify and make contact with their victims, with young people experiencing poor mental health and/or identifying as LGBTQ+ significantly more likely to be targeted.[37]

34 FBI (2023) Public Service Announcement: Violent Online Groups Extort Minors to Self-Harm and Produce Child Sexual Abuse Material. Published September 12, 2023
35 National Crime Agency (2025) National Strategic Assessment 2025
36 Europol (2025) Intelligence Notification: the Rise of Online Cult Communities Dedicated to Extremely Violent Child Abuse. Published January 2025
37 MRF analysis of the threat dynamics and risk profile associated with Com groups. Further information is available on request.

### How AI investment may increase the risks posed by Meta's content recommender systems

In its most recent earnings call, Meta signalled that AI is delivering substantial improvements in its core recommendation system, leading to a 6% increase in time spent on Instagram in just the last quarter.[38]

Senior leadership including Mark Zuckerberg and CFO Susan Li also signalled that further changes to the recommendation system were on the way, including to make content recommendations 'more adaptive to what a person is engaging with during their session'; ensuring that recommended content 'is the most relevant to what they're interested in *at that moment'*; and 'improving the ability for [recommender] systems to discover more diversified and niche interests through interest exploration and learning explicit user preferences.'[39]

In short, Meta is betting big on AI to turbo-charge its personalised recommendations – and its core recommendation system has been identified one of five major priorities for the company's future growth.

The impact upon children's safety may be considerable. With its renewed and explicit focus on how its core recommendation systems can drive user engagement metrics and time spent on the platform, and its renewed focus on high-risk hyper-personalisation, the Leadership Team's priorities have now been made clear.

Unless strengthened and more effective regulatory safeguards can be quickly established, it seems difficult to conclude that regulatory incentives can reverse what many will conclude is yet another example of Meta pursuing a reckless rush for growth, with damaging implications for the under 25s.

Safety, not solely profit, should be the guiding principle for the next generation of Meta's content recommendation models.

38 Meta Q2 2025 Earnings call.
39 Meta Q2 2025 Earnings call. Comments made my Mark Zuckerberg, Meta's CEO, and Susan Li, their Chief Financial Officer.

## Harmful content accessible and discoverable through hashtags

The growth of Reels and Stories has significantly shifted the ways in which users' access and discover content types on Instagram. Nonetheless, posting and engaging with 'feed' posts via hashtags remains a primary way in which young people use the platform – and in turn, it remains a major potential vector through which they can be exposed to harmful content.

Our analysis of 460 of the most engaged with 'feed' posts displaying hashtags linked to suicide, self-harm and depression content found that a substantial amount of harmful content still remains publicly accessible and discoverable.

Almost half of the posts we analysed were found to be harmful (47%), effectively unchanged from our previous analysis undertaken in 2023 (48%).

One in 20 posts (5%) either promoted or glorified suicide or self-harm behaviour, and in doing so, appears to be in breach of Instagram's terms of service.

While the overall prevalence of harmful content posted using hashtags appears largely unchanged, our analysis found fewer examples of posts compared to our previous analysis that violate Instagram's terms of service; that explicitly referenced suicide methods; and/or that discuss suicide and self-harm ideation in ways that are likely to result in cumulative harm (see table 2). Instagram deserves credit for the reduced exposure to more explicit and immediately harmful suicide and self-harm posts.

**Table 2: proportion of most-engaged content on Instagram, posted using suicide, self-harm and depression related hashtags, containing harmful material**

| Content type | 2023 research | 2025 research |
|---|---|---|
| Promotes and glorifies suicide or self-harm | 13% | 5% |
| References suicide methods | 4% | 3% |
| References suicide and/or self-harm ideation | 19% | 14% |
| Contains feelings of intense depression, misery and despair | 48% | 47% |
| Total content likely to be harmful | 48% | 47% |

n=460 (2025), n=564 (2023)

Nonetheless, this data also speaks to the platform adopting an overly narrow focus on more explicit forms of suicide and self-harm content, while continuing to place insufficient emphasis on the potential impacts of broader cumulative harms. This includes the potential ways in which being algorithmically recommended large volumes of broader categories of harmful content may result in increased risk of poor mental health outcomes; and when algorithmically recommended interchangeably with suicide-related posts, in an increased risk of suicide ideation and behaviours among young people.

Instagram has made some positive strides to introduce additional friction in the search and discoverability of harmful content. It is now noticeably harder to proactively search for and access a number of hashtags that were previously widely associated with suicide and self-harm material, and all of the hashtags analysed in our 2023 research can now no longer be proactively searched.

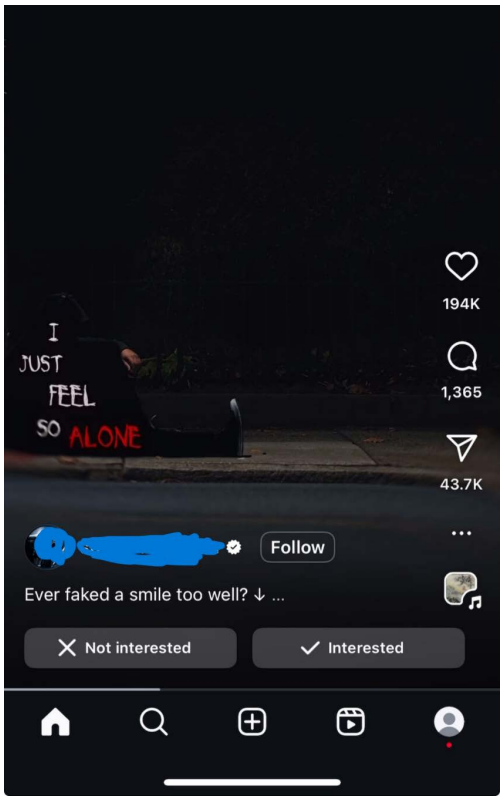
## Is Instagram 'gaming' its responsibilities under the Online Safety Act?

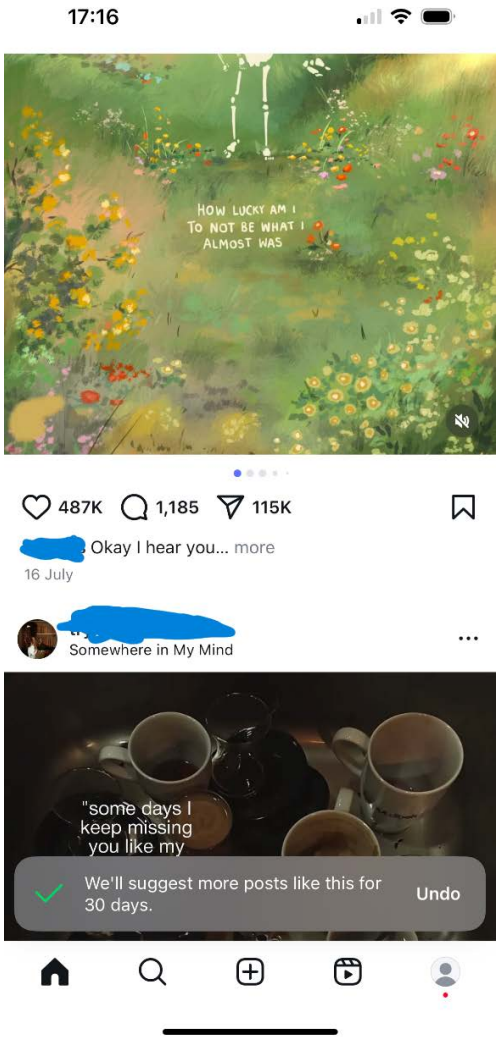
During our analysis of Instagram's product features, we found evidence that Instagram has been starting to introduce new design features to comply with its requirements under the Online Safety Act, but that it also appears to be actively gaming some of the measures required by Ofcom's Codes.

Specifically, Instagram has proceeded to introduce a new prompt which allows children to provide feedback on content they have algorithmically recommended. While this measure was proposed by Ofcom to enable children's negative feedback on content they encounter, and in turn to gain a greater agency and influence over their feeds, Instagram has instead chosen to design this prompt in a way that enables a child to either offer positive or negative feedback about the content they are being recommended.

As the example below illustrates, a child is therefore able to offer positive feedback about algorithmically recommended items of content, including material that references suicide, self-harm and repeated themes of intense depression and despair. In cases where a young user chooses to offer positive feedback, the user receives a prompt saying they'll be recommended further such content for the following 30 days.

*Examples of the user interface where a user is asked to feedback on whether harmful content is interesting to them, and the notification that shows they will be shown further similar content for the following 30 days.*



While it is deeply disappointing to see Instagram choose to game its regulatory requirements in such a way, it was also entirely foreseeable. In response to Ofcom's consultation on this measure, MRF urged the regulator to strengthen its proposals,[40] flagging the extensive track record by social media companies of using 'dark patterns' and other user experience choices to actively subvert intended outcomes and the overall impact of regulatory demands.[41]

In its consultation response, Ofcom declined to act on these concerns. The regulator stated that it is important platforms 'have flexibility over the design adjustments to recommender systems algorithms', and that as a result it did not intend to prescribe how these requirements should be introduced.[42]

Platforms have therefore effectively been given free rein to implement the measures as they see fit – and in this case to actively subvert the regulator's policy intentions. Ofcom must now be prepared to urgently revisit and amend its position, with detailed guidance and expectations regarding the user choice infrastructure that should be adopted by firms, as well as comprehensive guidance on how platforms should be expected to design, report and amend their recommender systems in light of user feedback.

## Algorithmically recommended harmful content on multiple product surfaces

Our research suggests that Instagram continues to persistently recommend harmful content, including suicide, self-harm and intense depression videos and posts, across a wide range of its core product services, including its core feed and Discover page. Limited if any evidence of the application of safety-by-design principles on either of these product surfaces could be seen.

On the core feed, we were recommended a near continuous algorithmically recommended set of harmful videos and feed posts, including content that explicitly promoted or glorified suicide, depicted suicide acts, and that extensively referenced themes around death and commitment. We also observed a number of posts depicting deaths by suicide which appeared to have been AI generated.

On both the core feed and Discover page, we were fed a large volume of posts referencing intense themes of misery, hopelessness and despair, typically interspersed with suicide and self-harm related material. These posts included:

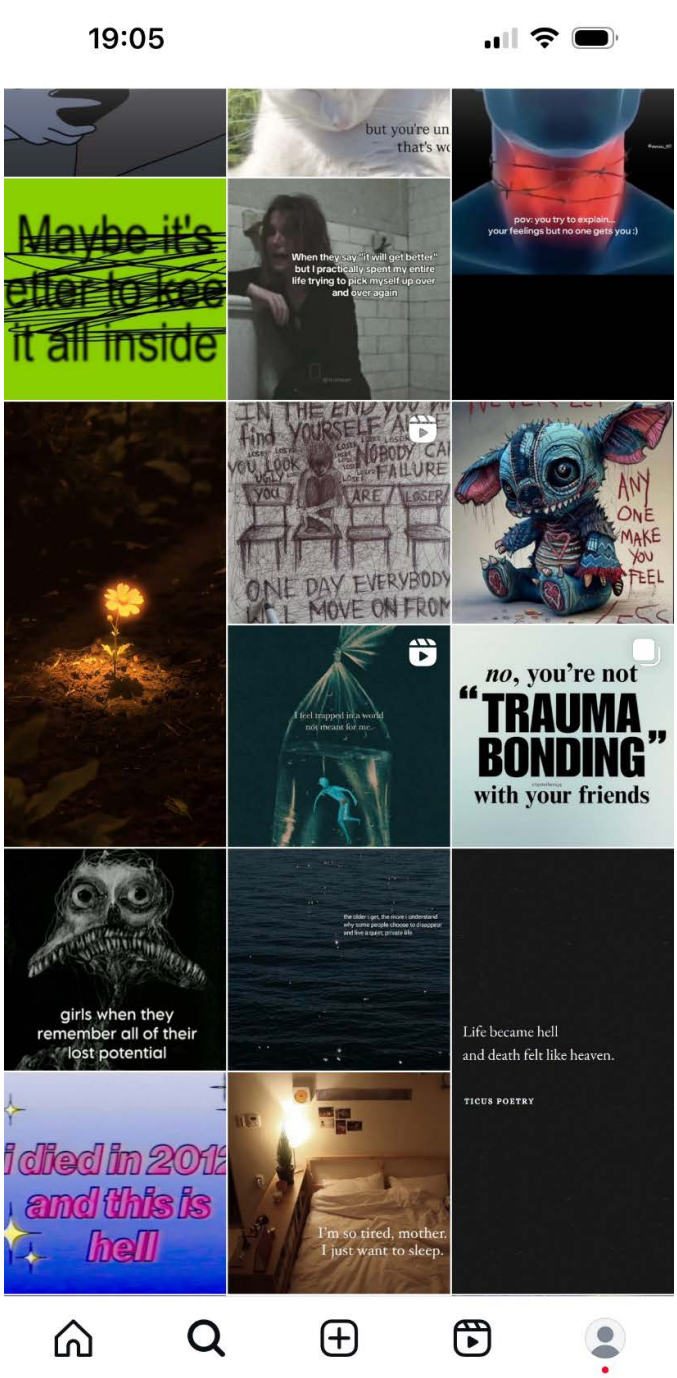
- carousel posts[43] and Reel videos actively espousing intense themes of despair and suicide, including a video that proclaimed 'they say "go where you're valued", so I disappeared'; and another with a caption reading 'this is our real home', playing over a video of a newly dug grave.
- multiple meme-style videos that reference themes relating to suicide and serious injury. For example, we were shown multiple variations of Reels in which the user declares 'this is what I want for my birthday', before switching to video of a high-impact crash. We were also shown repeated videos of a car and/or motorbike travelling down the motorway at dangerously high speeds, with accompanying text-based audio captions 'promising to make it look like an accident.'

40 Molly Rose Foundation (2024) Response to Ofcom's Consultation on its Protection of Children measures
41 OECD (2022) Dark Commercial Patterns: OECD Digital Economy Papers, No 336. Published October 2022
42 Ofcom (2025) Statement: Protecting children from harms online. Published 24th April 2025
43 Posts consisting of several screens, with the user swiping sequentially from one to the next.

- a continuous stream of posts that actively romanticised or normalised suicide ideation and feelings of intense misery, including videos that stated: ‘if I disappear, who cares? Nobody does, and that’s ok.’

- a large number of videos and posts produced to a high standard and featuring dark aesthetics, visuals and/or suicide related imagery. This included night-time shots, vantage points from high buildings or next to railway lines, and audio clips of teenage girls crying and experiencing acute emotional distress.

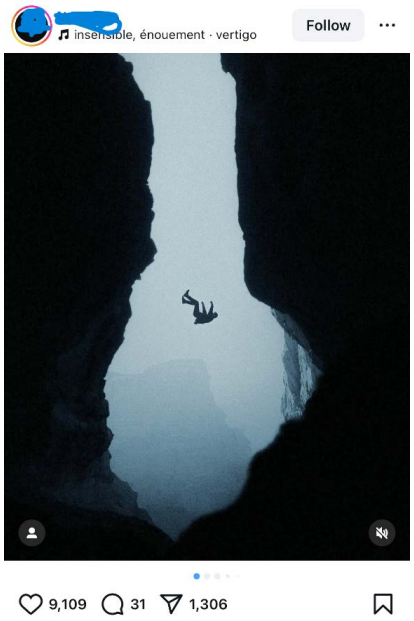
- a substantial number of posts that appeared to have clear parallels to the content seen by Molly in the weeks and months before her death. Posts included a hand drawn image with an accompanying caption ‘I am not a burden that anyone should bear’. We were routinely shown posts depicting intense themes of misery and despair, with captions including: ‘I’m getting used to being consumed by my own thoughts’, ‘I’m tired feeling this way’ (overlaid over a drawing of a dead bird), and ‘I have so much pain in my heart’ (overlaid over a picture of a broken bottle).

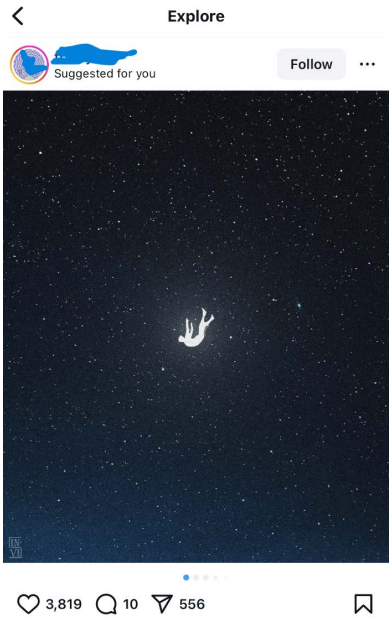



*An example of the Discover page, recommending a range of accounts and videos featuring harmful content.*

## Examples of harmful content[44]

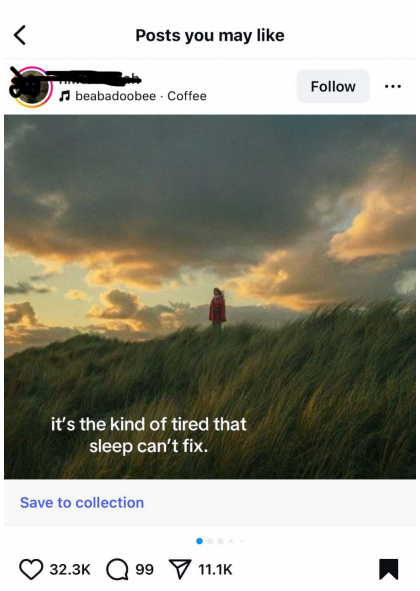
We were repeatedly shown highly stylised imagery, including images which appeared to be AI-generated, showing depictions of people falling from heights and drawing on well-observed memes that have the effect of romanticising suicide ('I'll make the stars pretty for you.')



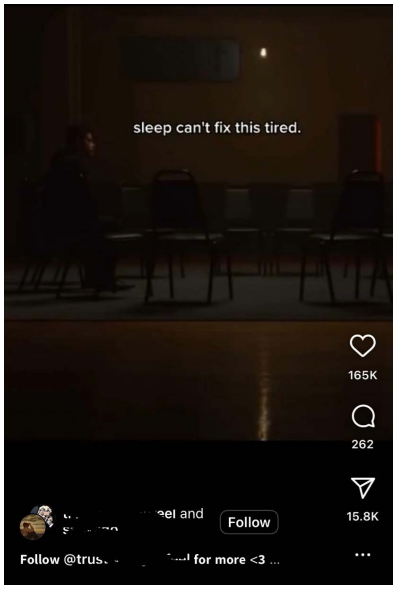





A substantial amount of algorithmically recommend posts contained content that exacerbated and reinforced feelings of intense misery, hopelessness and despair. Many of these were professionally produced, highly stylised and as a result achieved high levels of user reach and engagement.

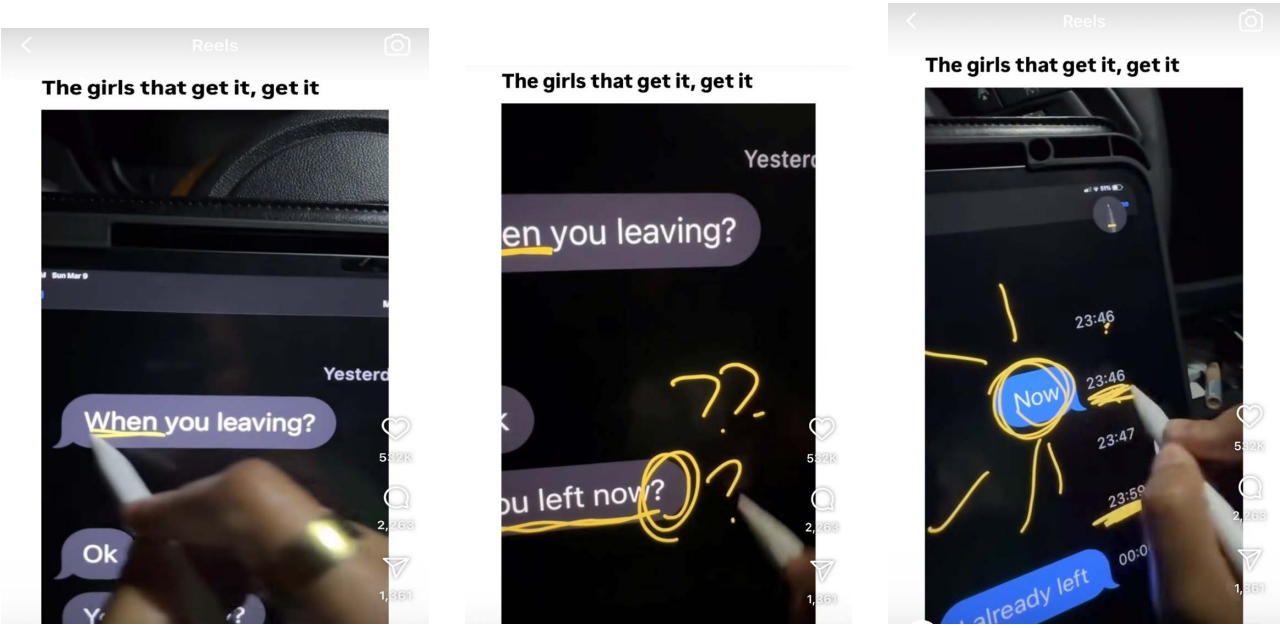







44 We are showing a selection of images that were algorithmically recommended to freely accessible to teenage accounts. While the images included in the report are distressing, we have not shown examples of much more extreme forms of material which we encountered during our research.

This short-form video actively encouraged and promoted suicide, with an annotated exchange of messages that suggested the user should take their own life without delay.




## High risk user engagement features

A number of high-risk user engagement features or product design choices first highlighted in our 2023 research continue to remain unaddressed, and in turn increase the overall risk profile associated with suicide, self-harm and depression-related risks on the platform.

We are dismayed that Instagram continues to use off-platform user engagement techniques such as **personalised content recommendation emails**. Such emails actively recommend harmful content and use a range of persuasive design techniques to encourage frequent log-ons, and bear striking hallmarks to emails that Molly received from social media platforms recommending harmful content in the months before her death.

During our fieldwork period, we were sent daily emails encouraging us to view time-limited Stories content, with messaging that emphasised that 'one of the stories expires in less than 24 hours'.[45]

We were also exposed to a number of other well-observed engagement techniques in these emails, including the use of 'red dots' which on Instagram's user interface are used to suggest that notifications and/or direct messages were waiting to be seen.[46]

Other visual cues including the use of the 'red heart' symbol, usually shown on the user interface to denote 'likes' from other users, even in cases where our account had not posted or received any engagement from other accounts.

45 This is a well-established dark pattern that emphasises a temporal limit to exploit the scarcity heuristic. OECD (2022) Dark Commercial Patterns: OECD Digital Economy Papers No 336

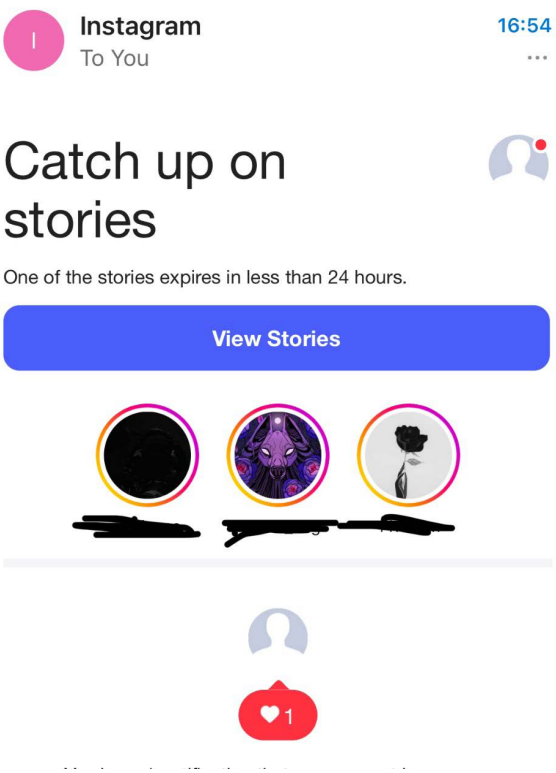
46 Similar tactics were seen in the landmark study into dark patterns undertaken by Forbrukerrådet in 2018. Forbrukerrådet (2018) Deceived by Design: how tech companies use dark patterns to discourage others from exercising our right to privacy

Instagram continues to allow accounts to readily **identify themselves in their account profiles using exaggerated and/or misrepresented descriptions** such as 'health experts', 'mental health resources', and even 'crisis prevention centres.' This option provides a false sense of legitimacy and underscores how the platforms design can be readily gamed by high-frequency, meme style accounts.

Many of these accounts also feature consistent wording encouraging other users to 'DM for support', presenting an evident risk that malign actors may be able to take advantage of Instagram's platform architecture and choices, including for criminal and malicious acts.

High frequency, meme style accounts, which continue to account for a substantial majority of the harmful content that we identified, continue to actively capitalise on more recent product services such as **Instagram Stories and broadcast channels**, with both surfaces enabling accounts to rapidly build their follower base and levels of engagement.

User engagement mechanisms further increase the risks associated with broadcast channels, with prompts that encouraged us to 'never miss creator updates' by turning on creator notifications. Some broadcast channels used the design of the surface to maximum effect, including a 'how are you?' check-in, in which users went on to express deep thoughts of despair and mental distress, which were visible to all other followers of that channel.




*An example of the personalised content emails we received, sometimes on a daily basis.*

# TikTok: Key findings

Our analysis of TikTok found that the platform continues to algorithmically recommend suicide, self-harm and intense depression content in appallingly large volumes, and that if anything its risk profile has substantially increased since our previous research in 2023.

TikTok's For You Page recommended a near continuous barrage of harmful material, with higher concentrations of the most severe types of material, including posts that promote and glorify suicide and that actively suggest suicide methods. We were more likely to be recommended more severe forms of content than on Instagram.

Our analysis found that:

- TikTok's For You Page recommended an almost uninterrupted supply of harmful material, with 96% of recommended videos containing themes that are likely to contribute to cumulative harm when viewed successively and/or in large quantities. Most of these posts achieved deeply concerning levels of reach.

- Content being recommended on the For You Page consisted of a near relentless supply of posts that were reasonably likely to reinforce intense feelings of misery, hopelessness and despair. These videos were frequently interspersed with material that explicitly promoted and glorified suicide ideation and behaviours; actively recommended suicide methods; and that used memes to actively promote and/or normalise suicidality and suicide behaviours.

- TikTok has developed several new user retention and engagement mechanisms that further extend the risk to young and vulnerable users. Related topic prompts are now routinely displayed when watching content on the For You Page, which in a single step opened up 'rabbit holes' of deeply harmful content to our account. This included memes and other videos that actively promoted and encouraged a broad range of suicide methods and dangerous stunts and challenges, including ones previously unknown to the researcher.

- We found no evidence at all that TikTok has taken *any* substantive safety-by-design measures to address the relevant risk profile for young people since our previous research, for example through introducing friction into the search and discoverability experience. Every major problematic user engagement mechanism and prompt highlighted in our 2023 research was still in operation.

- Two-thirds of the most engaged posts shared using problematic hashtags were classified as harmful (66%), a 16-percentage point increase since our previous research. While hashtags are less integral to the user experience on TikTok than on some other comparative platforms, this lack of search and discoverability friction means that hashtags and search terms can readily serve as a gateway into harmful content, and in doing so will further help to train the For You Page of a user's engagement with this material.

## Harmful content recommendations on the 'For You Page'

TikTok's 'For You Page' is the primary product surface through which users of the platform view content, with a continuous 'autoplay' stream of algorithmically recommended videos and other suggested content. TikTok's personalised recommender algorithms, which power the For You Page, are widely regarded to be integral to the platform's considerable success.

Our research has found exceptionally problematic volumes of harmful suicide, self-harm and intense depression content being algorithmically recommended on the For You Page. Almost all (96%) of the algorithmically recommended videos we were shown were found to be harmful, an almost uninterrupted barrage of harmful material that when watched consecutively and/or in large amounts is reasonably likely to increase the risk of children and young people being exposed to cumulative harm.

In this part of the research, we analysed 342 short-form videos, each of which were algorithmically recommended and watched consecutively on the platform's For You Page. Our analysis was undertaken after previously accessing suicide, self-harm and highly depressive content for earlier stages of the project. Videos were classified as harmful if they either promoted or glorified suicide or self-harm; referenced suicide or self-harm ideation; or otherwise featured highly intense themes of hopelessness, misery and despair.

TikTok's For You Page is the highest risk of any product surface analysed in this research, with a substantial proportion of recommended videos containing the highest risk category of suicide and self-harm themes. Our analysis found that:

- **8% of all harmful videos explicitly promoted or glorified suicide and self-harm acts.** TikTok generally moderates substantially higher volumes of suicide and self-harm posts than other major sites,[47] but this suggests that a considerable amount of harmful content remains undetected and is in turn algorithmically recommended to users.
- **16% of harmful videos contained explicit references to suicide methods**, including visual references to suicide acts, memes referencing the desirability of deaths by suicide (for example, 'freefall' memes showing the prospect of jumping from buildings in a positive light), and extensive references to suicide methods. As set out later in this report, this included references to suicide methods previously unknown to the researchers.
- **Over half of harmful posts (55%) actively contained references to suicide and self-harm ideation and behaviours**, including a far greater concentration of self-harm videos than observed on Instagram.

47 Molly Rose Foundation (2024) How effectively do social networks moderate suicide and self-harm content? An analysis of the Digital Services Act Transparency Database.

## The reach of algorithmically recommended harmful content

Suicide, self-harm and highly depressive content featured on the For You Page generates exceptionally high levels of reach and engagement, with the posts we analysed achieving substantially higher views, likes and comments than on any other comparable site, including Instagram.

Among the harmful videos we were recommended, posts had been liked an average of more than 303,000 times. More than a quarter of posts (26%) had been liked by more than 250,000 accounts, with almost one in 10 videos (9%) receiving at least one million likes.

As with Instagram, the volume of comments and replies is also considerable. One in five (20%) of harmful posts have received more than 5,000 comments, with each post receiving 3,810 replies on average. In our earlier research, we flagged the risks associated with poorly moderated comments acting as a de facto discussion forum for users experiencing suicide ideation, thoughts of self-harm, and other categories of mental distress.

TikTok has seemingly taken little if any action to address the considerable risks associated with either the velocity and virality of harmful content, or the frictionless ability of users to post thousands of comments on posts once they have been algorithmically spread. Comments continue to be a concerning vector for suicide and self-harm risks, with the sheer volume of comments and discussions having a number of adverse effects, including normalising and giving a distorted view of the scale of suicide and self-harm ideation among users; and encouraging many young users to actively ruminate on their poor mental health and experiences of suicide ideation.

We also remain hugely concerned about the frictionless ways in which young users are able to rapidly amass substantial 'albums' of harmful material – with young people being algorithmically recommended large amounts of harmful content, able to save it with a single click, and then access the content at a later date, including for the purposes of 'binge watching.'[48]

Our results show that more than one in three harmful posts had been saved over 39,000 times, with more than one in eight posts saved over 100,000 times. One in twenty harmful posts (6 per cent) had been saved by over at least 200,000 users.

While further research is needed to understand the 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 been able to amass such large albums and collections of harmful content.

Given the reasonably foreseeable risk that these harm pathways may facilitate 'binge watching', which has the potential to lead to emotional dysregulation, triggering thoughts, and even the onset of thoughts of self-harm and suicide ideation in some vulnerable adolescents, there is an urgent need for both TikTok and Ofcom to commit to urgent further action in the face of these well-observed risks.

48 Binge watching is understood to have been part of Molly's consumption of harmful material online.

## Harmful content discoverable using search and hashtags

Our broader analysis of posts on TikTok confirms that harmful content remains readily accessible and discoverable on the platform, with limited if any evidence of attempts to introduce search friction or other safety-by-design mechanisms that might prevent users from being able to access suicide, self-harm and intense forms of depression content.

Our analysis of 517 of the most engaged posts displaying hashtags linked to suicide, self-harm and depression content found that two-thirds of these posts (66%) had the potential to cause harm, particularly when consumed cumulatively over time.

Strikingly, the proportion of posts that contain harmful content has substantially increased compared to when we last undertook this analysis in November 2023. At that time just under half of posts (49%) were classified as harmful, as shown in table 3.

### Table 3: proportion of most-engaged content on TikTok, posted using suicide, self-harm and depression related hashtags, found to contain harmful material

| Content type | 2023 research | 2025 research |
|---|---|---|
| Promotes and glorifies suicide or self-harm | 4% | 3% |
| References suicide methods | 2% | 2% |
| References suicide and/or self-harm ideation | 16% | 22% |
| Contains feelings of intense depression, misery and despair | 49% | 66% |
| Total content likely to be harmful | 49% | 66% |

n=514 (2025), n=617 (2023)

Our findings suggest some broad differences between content discoverable through hashtags and material that is algorithmically recommended via the For You Page. For example:

- content that promotes and glorifies suicide and self-harm, and that explicitly references suicide methods, were respectively three and eight times more likely to appear in our For You Page than in our sample of the most engaged posts shared using relevant hashtags.
- posts that reference suicide and self-harm ideation were two-and-a-half times more likely to be algorithmically recommended in the FYP than to appear in our sample of content publicly accessible through hashtags.
- content that was not considered harmful, including posts that offer peer support and mental health and well-being advice (and that in turn may offer protective or beneficial effects), featured regularly in our sample of content shared using hashtags, but were only algorithmically recommended on our FYP once.

## High-risk design features and user engagement prompts

In our November 2023 analysis, we concluded that TikTok's exceptionally high-risk profile for suicide, self-harm and intense depression content was exacerbated by a set of poorly conceived and executed product design features.[49]

Almost 2 years later, our analysis finds no evidence that TikTok has taken any substantive measures to address the risk profile for teens, for example through introducing fiction into the search and discoverability experience.

Every single problematic user engagement mechanism and design feature highlighted in our previous research remains a core part of the product, while new high-risk user retention and engagement mechanisms have also been added, including the examples described below.

### Embedded search terms built into the For You Page

Since our previous research, TikTok has introduced new embedded features into its For You Page, including a new recommender prompt which routinely appears when a user watches or engages with recommended content. The recommender prompt typically recommends similar or related search topics, and it effectively enables the user to use the FYP as a springboard to access further material relating to topics and themes of interest.

Our analysis found this was an exceptionally high-risk design feature, which effectively fuses together the riskiest parts of recommended systems and frictionless search – and that enables the user to immediately find themselves immersed in 'rabbit holes' of related and typically harmful content.

During our analysis, we were repeatedly recommended search terms relating to extreme suicide, self-harm and depression material. For example, we were:

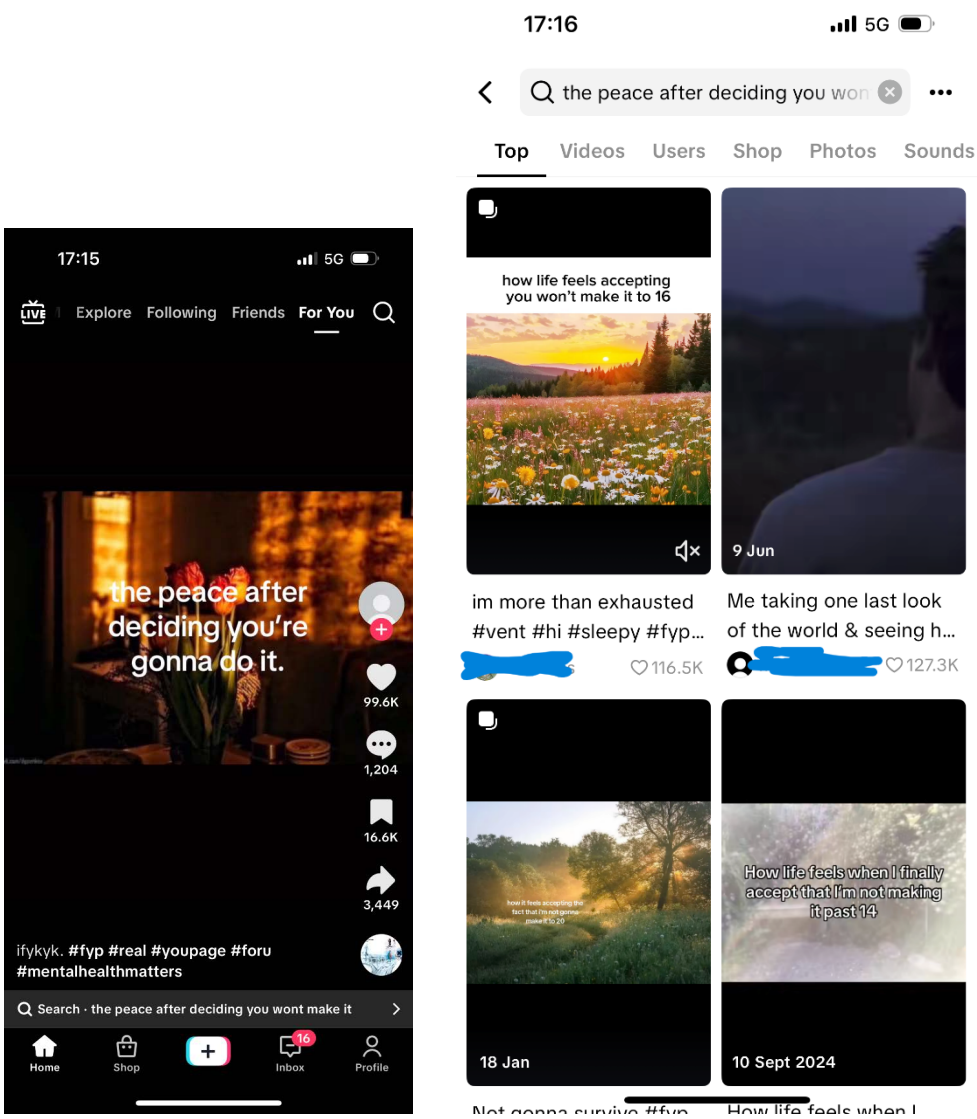
- repeatedly recommended search terms which had the effect of promoting and encouraging suicide, such as: 'I think I'm ready 2009–2025.' When we clicked on the search term, we were presented with hundreds of videos of teenagers expressing suicidal thoughts and behaviours and ruminating over their poor mental health.
- actively recommended a number of search terms relating to suicide methods, which meant that in a single click we were able to progress from a FYP video to a library of content that promoted and explained suicide methods (including suicide methods previously unknown to the research team.) In another video, we were shown a suicide method while being algorithmically recommended a search prompt: 'I'm so mentally drained I give up.'
- repeatedly shown search terms which either provided practical advice on self-harm ('blades for cutting tutorial') or that enabled unhelpful rumination on themes relating to self-harm ideation and behaviours ('why does my skin start to burn', 'self-care at night', shown adjacent to videos showing self-harm cuts, and 'blue hair cutting meaning', which introduced us to videos associated with cutting down to the vein. )

49 Molly Rose Foundation (2023) Preventable yet pervasive: prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest

- presented with recommended search terms relating to themes that romanticise and, in some cases, actively promote suicide behaviours, for example: 'I'll make the sky pretty for you' and 'I'll make the sky pretty for you meaning.'

- routinely recommended search terms relating to intense themes of depression, misery and despair. For example, 'it's getting bad again', 'why can't I cry any more' and 'I'm mentally drained.'

As a result of this design feature, at the same time as a user is able to effortlessly and immediately discover new and high-risk 'rabbit holes' of content, the prompts will also be simultaneously generating signals about their interests in this and other related types of material. This will subsequently inform further content recommendations across a range of platform services, including the For You Page, in effect fuelling TikTok's recommendation engines with a deeper and self-fulfilling understanding of a user's interest in and engagement with harmful content types.




*An example of TikTok's AI-generated recommendation prompt, and the videos that we were suggested once clicking on it.*

### Sharing and saving prompts

Since our previous research, TikTok has introduced a range of new on-screen prompts that are designed to encourage users to share content with other users. Troublingly, this includes prompts that suggest sharing content with specific named users; sharing content via third-party encrypted messaging apps such as WhatsApp and Signal; and re-posting to the account's own followers.

We were routinely presented with each of these problems while being algorithmically recommended harmful posts on the For You Page, including videos discussing suicide ideation, promoting suicide, and discussing intense depression, hopelessness and despair.

Clearly there are substantial risks associated with users being actively encouraged or incentivised to share harmful material, particularly to named users (and with the selection criteria for suggesting named users being evidently opaque.)

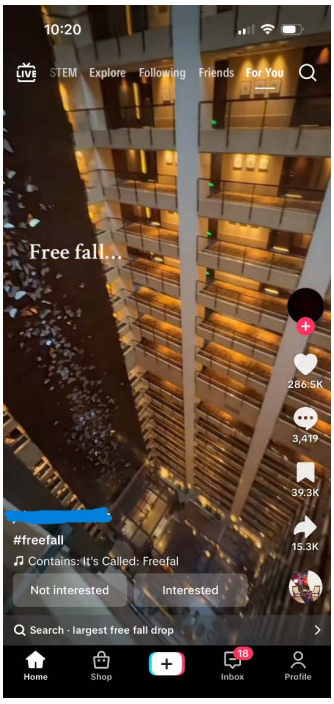
At a time when Ofcom and other regulators should be insisting on building increased friction to the sharing and discovery of harmful content, it is hugely concerning to see social media sites seemingly having free rein to extend the frictionless reach of their material, both on and off-platform.

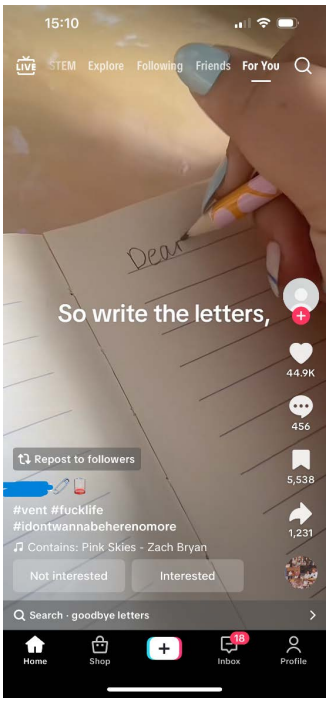
## Is TikTok gaming its responsibilities under the Online Safety Act?

Similarly to Instagram, we found evidence that TikTok appears to be introducing new features to comply with their requirements under the Online Safety Act, but that it is gaming them in a way that could result in young users being algorithmically recommended more rather than less harmful content.

The platform has followed Instagram's lead by introducing a new feature designed to offer younger users the chance to offer negative feedback on content recommendations, but that also goes one step further by allowing users to also offer *positive* feedback about the content they're being shown.

During our analysis, this meant we were able to offer positive feedback on suicide and self-harm related material. By submitting positive feedback, the user is informed they will be recommended additional amounts of similar content for the following 30 days.



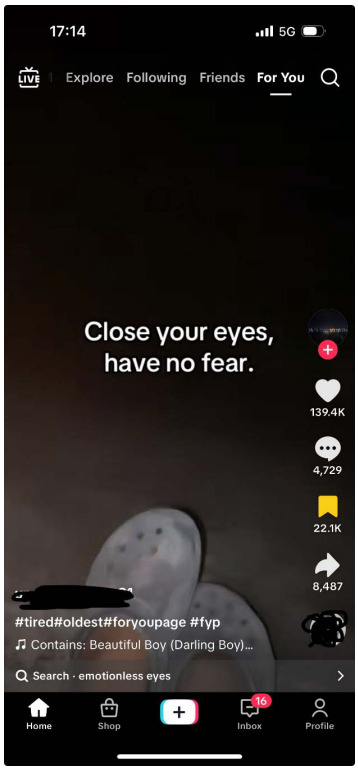

*Examples of suicide related content where we were prompted to express if we were interested in seeing similar and related content for the following 30 days.*

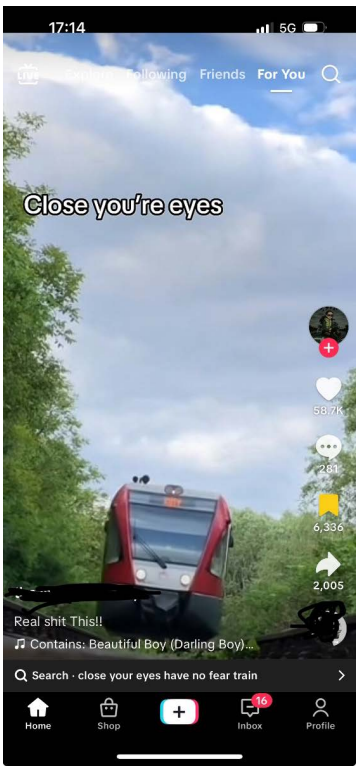
Ofcom must act quickly to close off the potential for platforms to 'game' their regulatory requirements – and must learn the lesson that platforms can and will exploit any ambiguity that is left available to them.

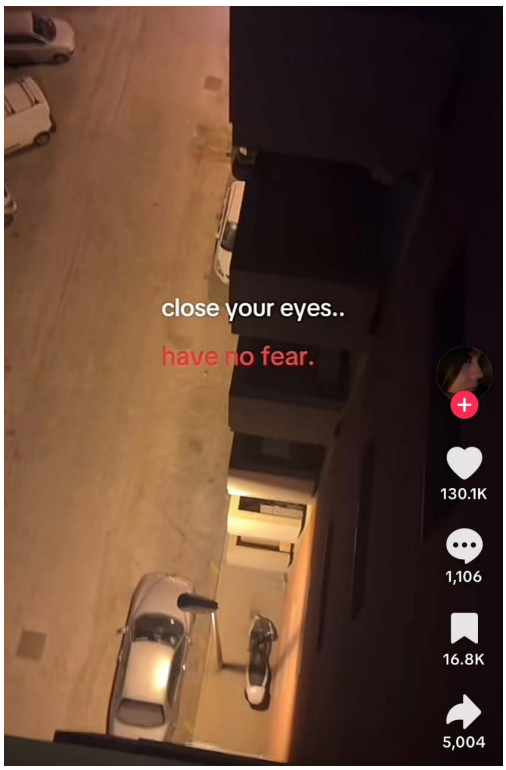
The regulator should also explore how it came to be that both major platforms decided to actively game Ofcom's rules in exactly the same way – deploying the same choice architecture and the same time periods (30 days) to recommend further content. MRF is concerned that platforms may be actively choosing to cluster around similar approaches to exploiting ambiguities in the regulatory regime, perhaps in the hope there is 'safety in numbers' and that in doing so they can effectively swamp Ofcom's supervisory arrangements.

## Examples of algorithmically recommended content

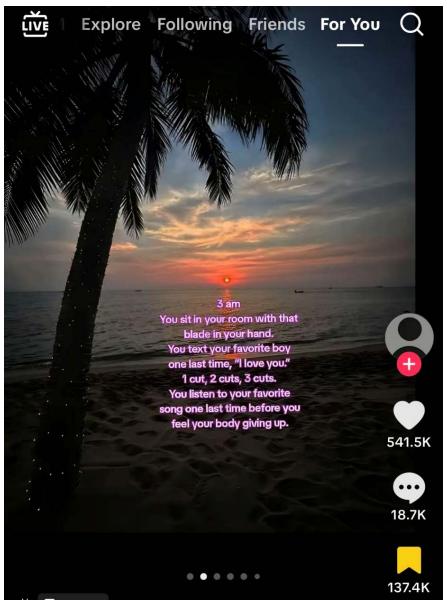
We were routinely shown content that promoted and reinforced suicide ideation and behaviours, for example memes that referenced encouraged users to 'close your eyes, have no fear' (drawn from an accompanying John Lennon song) in settings suggestive of suicide attempts.

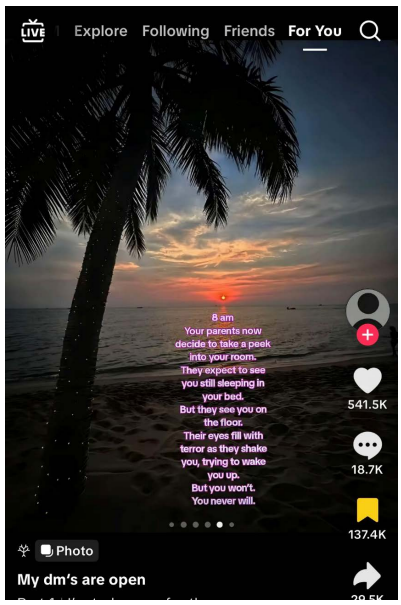


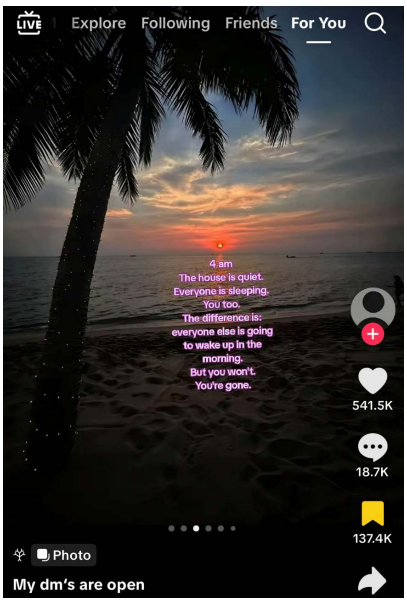



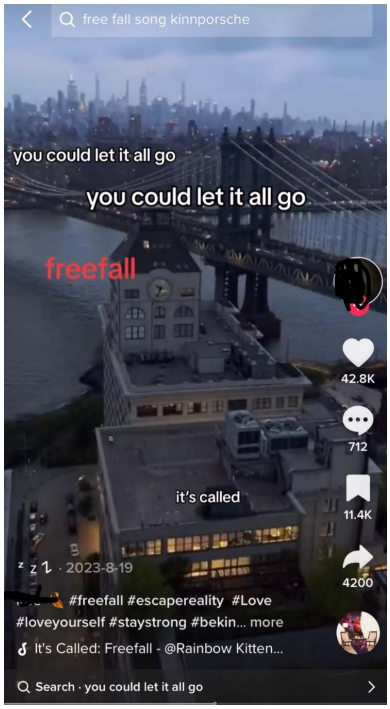
This distressing carousel post that described in the first person a teenager's death by suicide had already been liked by over 541,000 accounts.

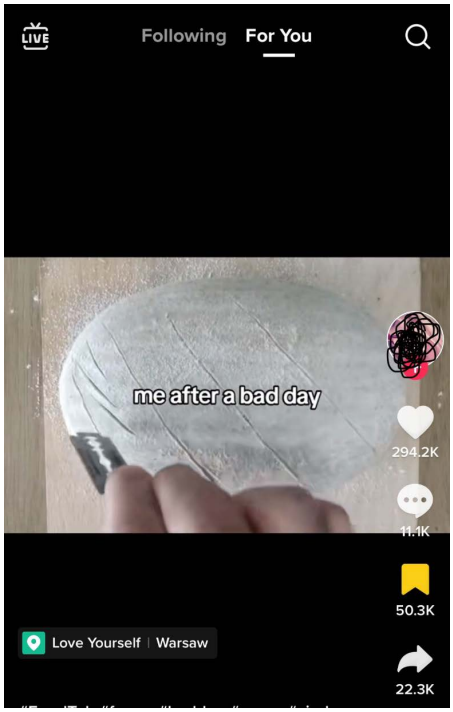


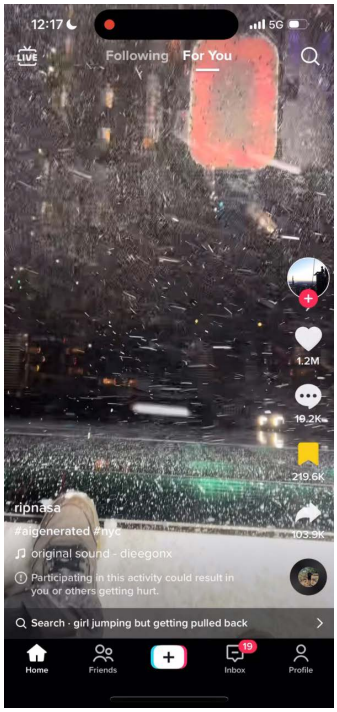




The For You Page repeatedly recommended memes relating to suicide ideation, including content that implicitly suicide methods and plans e.g. 'freefall', and 'self-harm' (using a loaf of bread to make cuts suggestive of bodily self-injury.) As shown below, one video had been issued with a warning label, but was still being algorithmically recommended at scale.



## Frictionless search and user engagement prompts

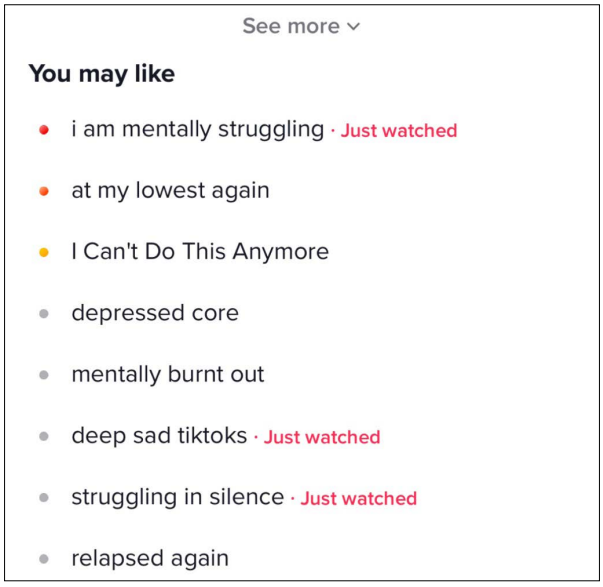
TikTok has taken no apparent steps to address the risks of being actively recommended suicide, self-harm and depression-related search results, with highly problematic search terms continuing to be suggested in search results, at the end of videos, and through autocomplete suggestions.

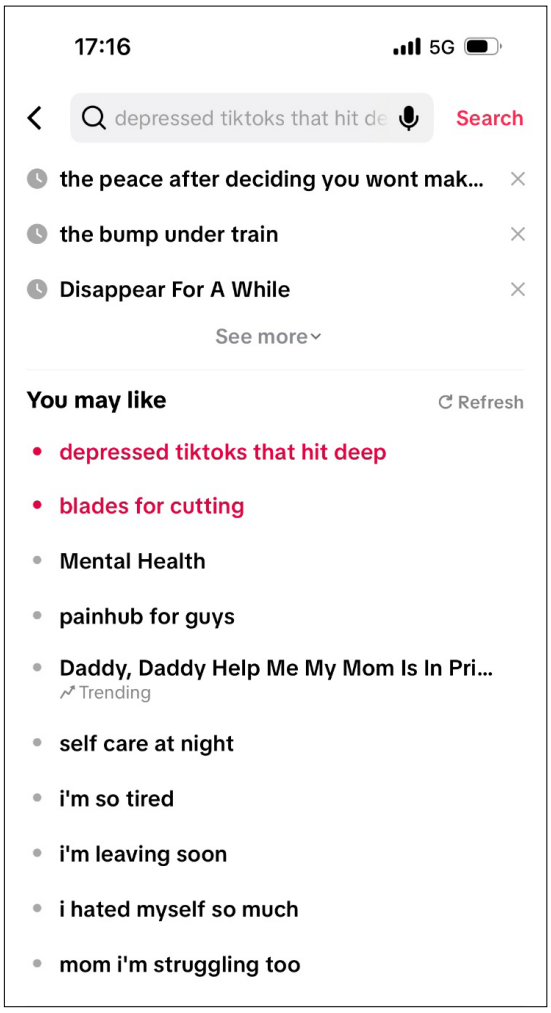
We found no evidence that TikTok had taken measures adopted by other firms to introduce friction in the search experience, for example by blocking high-risk search results.

Instead, TikTok's search functionality effectively remains a free-for-all, with deeply concerning video suggestions, search term and hashtag recommendations. Search results routinely suggested suicide, self-harm and intense depression phrases, including in embedded menu results, with suggested search terms including 'I'm going to end it', 'I want to disappear', 'I can't do this anymore', and 'quickest way to end it all.'

We were repeatedly offered algorithmically recommended 'you may like' suggested topics and phrases. Recommended search results included: 'I am mentally struggling', 'at my lowest again', and 'I can't do this anymore.'

Each of these suggestions were displayed alongside red, orange and yellow colour prompts, a well-known user engagement tactic that suggests to the user these topics are trending or popular, and that in turn could have the effect of normalising and/or exaggerating the scale of suicide, self-harm or mental health conditions among some users.




*Examples of recommended search topics and the persuasive design techniques nudging users to harmful search terms.*

## Playlists

TikTok's playlist feature, which enables accounts with over 10,000 followers to categorise their public content and let viewers watch related posts in a pre-selected sequence, has rapidly emerged as a user engagement mechanism that further drives exposure to harmful content (and in turn to cumulative harm.)

In some respects, this is not surprising: our 2023 research found that a significant proportion of harmful content was being produced by a set of high reach, high-engagement accounts that specialise in sharing memes, videos and other clips that show a constant barrage of suicide, self-harm and intense depression content. Many of these accounts are geared towards serving online communities that seek out and engage with harmful content, and they actively use hashtags such as #vent, #paintok and #3am thoughts.

We found a number of examples of playlists being used to collate depression, hopelessness and despair related videos (for example, an 'at your lowest' playlist).

Other playlists were used to collate suicide and depression-related videos which had attracted in excess of 1 million views, or that had been reposted by other users in large amounts, a means of ordering material which is likely to further drive traffic to these posts and to encourage further reposting and sharing of high-risk, harmful posts.

# Commercial advertising being shown adjacent to harmful suicide, self-harm and intense depression content

During our analysis of algorithmically recommended posts on the For You Page and Instagram's Reels surface, harmful forms of suicide, self-harm and depression content were regularly shown adjacent to personalised adverts, including content from large corporates, charities, educational institutes and small businesses.

Instagram and TikTok both use programmatic and behavioural advertising, where adverts are targeted at users based on their online interests, including the content they engage with on the platform and tracked visits to third party sites.

Behavioural advertising is an integral part of the business model for both platforms, meaning that advertisers and the broader advertising industry have a particular responsibility, but also opportunity, to ensure that platforms are making their products safer-by-design.

During our analysis, we were routinely shown adverts placed next to content which was either harmful in its own right, such as material that promotes or glorifies suicide and self-harm; or that had a significant potential to exacerbate cumulative harm when being recommended in large amounts, for example videos that normalise and romanticise suicide, self-harm and intense thoughts of misery, hopelessness and despair.

Across the 642 algorithmically recommended videos that we analysed, we watched a total of 67 adverts that were shown immediately before or after items of harmful material – 36 on TikTok (from 25 separate advertisers),[50] and 31 adverts (from 25 separate advertisers) on Instagram.

This means that we were shown advertising adjacent to harmful material for one of every 9.5 For You Page posts watched consecutively on TikTok, and for one in every 10 Reels watched on Instagram.

While we don't intend to name the advertisers whose content was shown adjacent to harmful posts, during our analysis we identified several major categories of advertisers whose content repeatedly appeared.

50 This excludes livestreams or TikTok Shop videos where additional goods may have been marketed.

Key groups of advertisers whose posts appeared adjacent to harmful content included:

- Major fast-fashion retailers popular among teenagers, including well-known online brands;
- Major UK and international retailers, including fast-food brands, high-end retailers and clothing stores;
- A large number of UK university and educational establishments, including many music and music production courses;
- Several major charities, including non-profits working on mental health, human rights issues, and humanitarian appeals relating to the war in Gaza;
- Small, locally-based youth-focused projects, including youth activities, sports clubs and martial arts groups;
- Social media platforms advertising on each other's products.

These category types actively underscore the importance of behavioural advertising in determining the risk to advertisers that their content will be shown next or adjacent to harmful material aimed at young people – with fashion retailers, online sites and UK universities, advertisers who would logically want to target teenagers, unsurprisingly appearing prominently in our results.

## Make-up and beauty ads adjacent to harmful material on TikTok

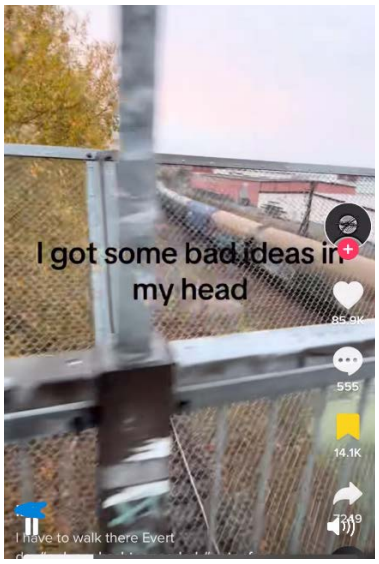
On TikTok, we were also shown a large number of adverts relating to health and beauty, make-up and skincare products.

While the scope of this research means we are unable to determine if the prevalence of skincare and make-up adverts on TikTok is simply the result of programmatic advertising based on age, or is also derived by our interest in poor mental health and depression content, industry whistleblowers have previously made claims that teenagers have been targeted with these types of adverts based on signals that are suggestive of poor mental health and/or emotional vulnerability.

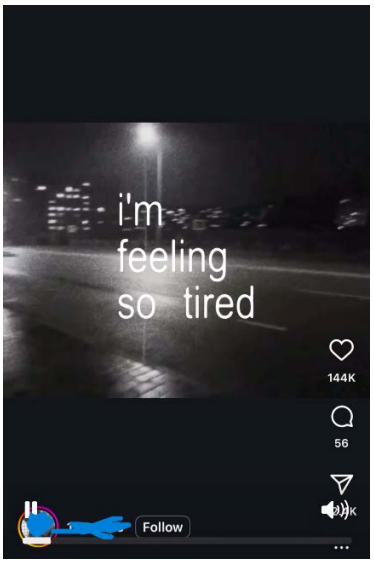
For example, Meta whistleblower Sarah Wynn-Williams claims that the company used its products to track when 13–17 year old girls deleted selfies, in order to then serve up personalised adverts relating to beauty products and extreme weight loss.[51]

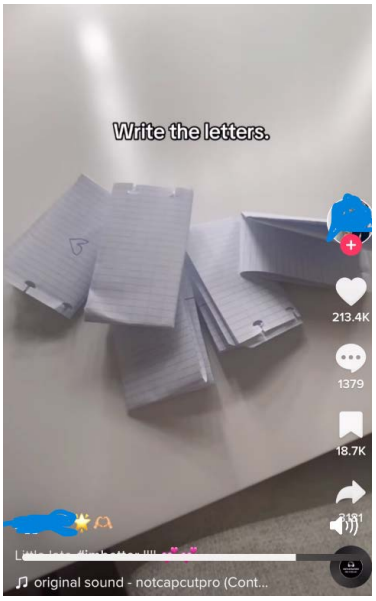
51 "Facebook does work for a beauty product company tracking when thirteen-to-seventeen-year-old girls delete selfies, so it can serve a beauty ad to them at that moment. We don't know what happens to young teen girls when they're targeted with beauty advertisements after deleting a selfie. Nothing good. There's a reason why you erase something from existence. Why a teen girl feels that it can't be shared. And surely Facebook shouldn't then be using that moment to bombard them with extreme weight loss ads or beauty industry ads or whatever else they push on teens feeling vulnerable." Wynn-Williams, S (2025) Careless People: A story of where I used to work.

# Examples of relevant harmful content and adjacency to behaviourally targeted adverts

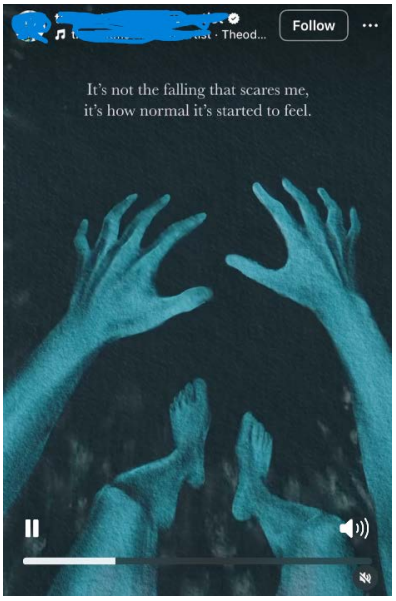


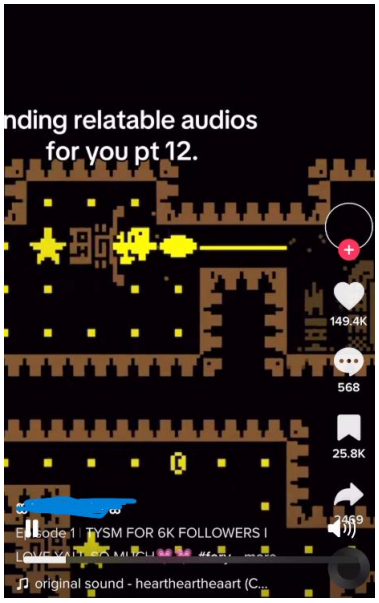

Advert for fast food retailer



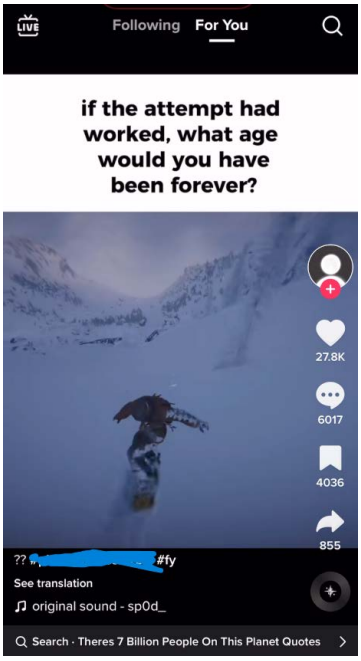




Online fast fashion retailer aimed at young people




Major coffeeshop chain



## What should companies do to make platforms safer for teens, and to minimise their own reputational risks?

It is in no-one's interest for harmful content to be algorithmically recommended to teens, let alone corporates and charities advertising on social media sites.

We therefore strongly encourage online advertisers to step up and exercise their leverage on social media platforms – recognising that as a key contributor to the business model of social media sites, corporates can play a significant role in requiring the likes of TikTok and Instagram to do better.

As the UN Global Principles on Information Integrity identify:[52] 'advertisers can benefit the information ecosystem in a way that both strengthens information integrity and makes good business sense. While technology companies are unlikely to readily abandon current business models, healthier incentives can be achieved through greater transparency for advertisers into advertising processes and the adherence to human-rights responsible advertising policies by advert deliverers. By gaining more control of a transparent supply chain, advertisers can also see a better return on their investments.' This is particularly the case in respect of advertising placement and adjacency and ensuring that advertising spend isn't inadvertently funding content that is harmful to children.

The recent Science, Innovation and Technology Select Committee report also identifies the significant positive role that corporate advertisers can play in 'tackling online harm [by] addressing the principles that incentivise and monetise its spread',[53] highlighting the role that advertising-driven business models played in the proliferation of false and harmful content during last summer's riots.

In the months ahead, MRF intends to engage with the advertisers whose content appeared adjacent to the harmful material analysed in this research.

As a minimum, we will be encouraging each of these advertisers to sign up to the Children's Rights and Well-Being Guiding Principles drafted by the Conscious Advertising Network,[54] but also to apply their own pressure on Meta and TikTok to actively prioritise safety-by-design and wellbeing-by-design for teens, and to provide more meaningful and transparent risk exposure metrics to the advertising industry.

Ensuring that children can no longer be algorithmically recommended harmful material is not only important for children and families, but it can also give advertisers and investors confidence that their adverts will no longer be shown next to high-risk material, nor that their ad-buying strategies contributing to a business model that exacerbates the risk of cumulative harm among young people.

Advertisers each have a role to play to prevent social media continuing to cost young lives.

52 United Nations (2024) Global Principles for Information Integrity. Recommendations for Multi-stakeholder Action.
53 Science, Innovation and Technology Select Committee (2025) Second Report of Session: Social media, misinformation and harmful algorithms
54 Conscious Advertising Network (2024) Children's Rights and Well-being Guidelines (updated edition.) https://www.consciousadnetwork.com/childrens-wellbeing-case-studies/

MOLLY ROSE FOUNDATION

Registered Charity No: 1179482
https://mollyrosefoundation.org

Written by Molly Rose Foundation, in partnership with The Bright Initiative by Bright Data
Data analysis by Molly Rose Foundation and Showrunner Communications

For more information, please contact
hello@mollyrosefoundation.org

**If you need to discuss any of the themes raised in this report, the following support services are available in the UK 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**