MOLLY ROSE FOUNDATION

How effectively do social networks moderate suicide and self-harm content?

An analysis of the Digital Services Act Transparency Database

Contents

Summary

Suicide and self-harm content is a major risk for children's safety and well-being, with increasing evidence of the relationship between exposure to harmful content and self-injury ideation and behaviour among young people.[1]

Until recently, policymakers and civil society have had limited understanding about how tech companies respond to the risks facilitated on and by their services. The EU's Digital Services Act is playing an important role in resetting this information asymmetry. Since September 2023, the largest online platforms have had to meet the DSA's comprehensive transparency requirements, including a duty to publish details of every relevant content moderation decision they make.

This report is the first major analysis of DSA transparency data relating to content moderation decisions relating to suicide and self-harm material. It analyses over 12 million decisions taken by six major platforms between September 2023 and April 2024: Instagram, Facebook, TikTok, Pinterest, Snapchat and X.

Our analysis finds pronounced deficiencies and inconsistencies in the response to suicide and self-harm content across many of the platforms that we analysed. While some online services appear to be investing in proactively identifying and removing harmful content, notably Pinterest, most major platforms appear to be substantially failing to respond to the risk profile of their products.

As a result, children and young people are being inadequately protected from harmful content that remains freely accessible, searchable, and that in many cases can continue to be algorithmically recommended.

Our analysis finds that:

- **Almost all (98 per cent) of content moderation decisions are taken by just two platforms: Pinterest and TikTok**. These platforms deserve substantial credit for moderating content at a scale that is more likely to be commensurate to the volume of harmful content available on their services. Despite the substantive risk profile of Meta's services, **Instagram and Facebook each account for only one per cent of moderation decisions**. **This raises substantive questions about the adequacy of the response of Meta-owned platforms** to suicide and self-harm risks;

- **There are significant inconsistencies and shortcomings in how major platforms respond to suicide and self-harm content on their services**. DSA data suggests that some platforms are failing to prioritise content moderation on the highest risk parts of their services: for example, only 18 per cent of Instagram's content moderation decisions related to image-based posts, with a further 2 per cent related to video-based content. This is despite research that shows

1 See for example Kusi, 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. 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

Instagram's short-form video product, Reels, has the highest risk profile of any of Instagram's major product surfaces;[2]

- **Some platforms appear to act too slowly to remove content, and in some cases, it is arguable whether they take all appropriate measures to restrict access to harmful content**. While almost all of TikTok's moderation decisions relate to content posted on the same day (94.4 per cent), this applies to only two-thirds (67 per cent) of Facebook's decisions. One-sixth of content actioned by Facebook had been available for at least 100 days. Less than one-fifth of Pinterest's moderation decisions relate to content posted on the same day, with 23 per cent relating to content that had been available on the platform for at least a year.

- **There are significant and seemingly irreconcilable differences between the amount of moderation decisions that Meta reports in its DSA filings and that it claims to action in its voluntary transparency reports**. In Q1 2024, we saw only 7.6 per cent of the DSA reports that we might expect from Facebook, and 8.5 per cent from Instagram, if the volume of suicide and self-harm actions claimed in Meta's voluntary reports were proportionately apportioned to its EU user base. This underscores the importance of all regulators using their information disclosure powers to interrogate data supplied by companies, and the important role of external auditing and quality assurance functions to support regulatory outcomes.

- **Regulators should be asking substantive questions about whether Snapchat and X are doing enough to respond to the risks of suicide and self-harm content on their services.** X reported only 0.13 per cent of content moderation decisions that we analysed, while Snapchat submitted just 0.04 per cent of relevant decisions. There also appears to be issues with the consistency and reliability of the data submitted by both platforms to the European Commission, and we encourage them and other regulators to investigate potential discrepancies in company reporting whenever these arise.

2 Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of suicide and self-harm content on Instagram, TikTok and Pinterest.

Context

It is widely understood that mandated transparency requirements are a powerful mechanism open to legislators and regulators to understand how large social media platforms respond to the risks posed by harmful content, and to assess if online platforms are doing enough to protect their users.[3]

The European Union's Digital Services Act has established a comprehensive framework for mandated transparency among the largest online platforms. The DSA's transparency framework is comprised of several relevant articles:

- **Article 33 designates online services with more than 45 million users in the EU as either Very Large Online Platforms (VLOPs) or Very Large Online Search Engines (VLOSEs):** VLOPs and VLOSEs must comply with the most stringent rules set out in the Act, including enhanced transparency and reporting obligations.
- **Article 15 requires that VLOPs and VLOSEs release periodic transparency reports:** these must contain regularly updated data on the number of active users, content moderation processes and actions, and the timeliness of platform interventions.
- **Article 17 mandates VLOPs and VLOSEs to submit a specific and detailed Statement of Reasons (SoR) for each moderation decision that it takes:** each SoR must provide detailed information on the intervention, its legal basis, and the content that was actioned.
- **Article 24 (5) establishes that the European Commission must set up and maintain a database of SoRs:**. in September 2023, the DSA Transparency Database was launched: a standardised, centralised and publicly available database of all SoRs submitted within scope of the DSA.

The Transparency Database is a first-of-its-kind framework and represents an unprecedented resource of self-reported data that enables for the first time the possibility to track, scrutinise and compare real-world platform moderation actions.[4]

When submitting an SoR to the Transparency Database, the DSA framework requires VLOPs and VLOSEs to set out the content type to which each content moderation decision relates. As such, the DSA Database enables researchers to explore the decisions taken by the largest services in relation to a range of online harms, including suicide and self-harm content.

Given the sheer volume of content moderation decisions taken, in just a few months we already have a rich dataset from which to analyse and interpret the content moderation approach of the major platforms. In turn, this allows for an informed and detailed assessment of the adequacy and effectiveness of the moderation strategies deployed by them.

3 Trujillo, A et al (2024) The DSA Transparency Database: Auditing Self-Reported Moderation Actions by Social Media. In ACM, New York: NY

4 Dergacheva, D et al (2023) One Day in Content Moderation: Analysing 24 Hours of Social Media Platforms Content Decisions through the DSA Transparency Database. Lab Platform Governance, Media and Technology

Methodology

This report aims to assess and analyse the content moderation decisions of six major platforms in relation to suicide and self-harm content, and to identify relevant thematic issues relating to the volume and nature of violative content and the processes used to identify and action it.

Using the DSA Transparency Database, we identified and analysed data on Statements of Reasons (SORs) logged between September 2023 and April 2024 for each of the platforms in scope. SoRs were considered relevant if 'self-harm' was logged as the primary reported violation. Note that the DSA's 'self-harm' category includes all content violations relating to suicidality, self-harm, eating disorders and other forms of bodily self-injury, and it is broader than the definitions applied in other comparable legislative regimes.

For the majority of the analysis, we have used data extracted from the DSA Transparency Database Dashboard.[5] More granular analysis was performed on a time-bound sample of records to investigate issues such as length of time taken to detect and action violative content. For each of the services in scope, we used the Transparency Database's Data Download function to isolate Self Harm SORs uploaded in a one-week period, running from 15th to 21st April.

In the case of X and Snapchat, a comparatively low number of SORs were uploaded during this period, and this data should be treated with some caution due to the low sample size.

For our analysis on voluntary reporting, data on the number of monthly active users of Meta's platforms in the EU were sourced from its Article 24(2) disclosures, available on its website. This data was then compared against publicly available global data on the number of active monthly users.[6] Using voluntary reporting for Instagram and Facebook, again obtained from the Meta corporate web site, an estimated figure of action taken on self-harm content posted by EU users was calculated.

As discussed in the analysis, SORs are generated by platform self-reporting mechanisms, and it is readily apparent that there are substantial inconsistencies in how VLOPs interpret their reporting requirements.[7] In the course of our analysis, we identified significant discrepancies in how some large platforms populate their SORs, and readers should interpret the results accordingly.

5 Due to issues with the data export function on the Dashboard not being enabled, data was collected using screenshots of data tables and the Optical Character Recognition (OCR) functionality of ChatGPT 4, with exported data checked for accuracy.
6 Sourced from Statista
7 Trujillo, A et al (2024) The DSA Transparency Database: Auditing Self-Reported Moderation Actions by Social Media. In ACM, New York: NY

Analysis

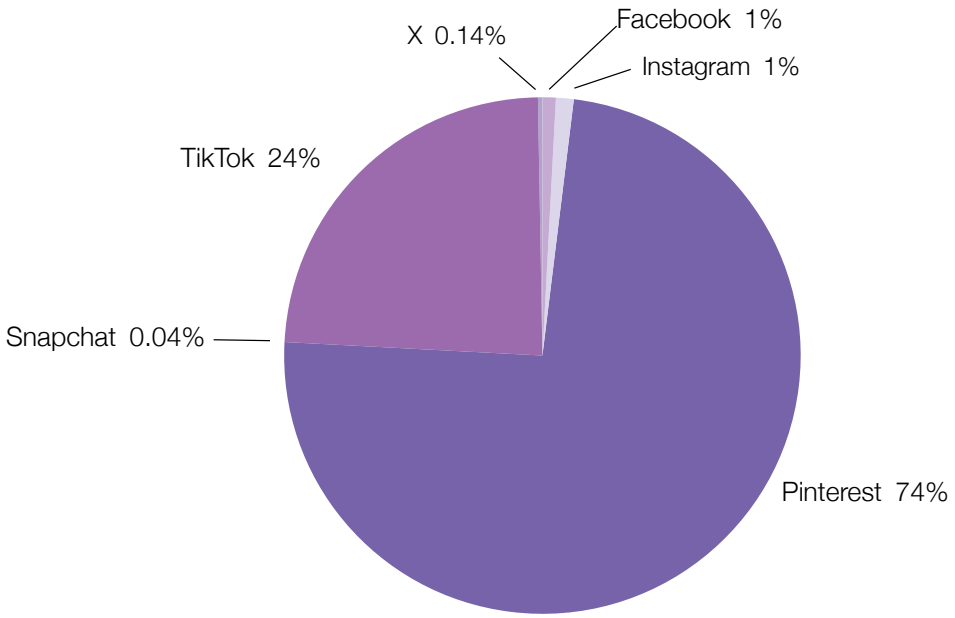
1. How many content moderation decisions do platforms take, and are there differences between major services?

Between September 2023 and April 2024, more than 12 million self-harm SoRs were logged by Pinterest, TikTok, Instagram, Facebook, Snapchat and X (12,194,588 moderation decisions.)

Suicide and self-harm content amounts for a relatively small proportion of all decisions made during this period (2 per cent). The largest volume of reports related to illegal content (33 per cent), unsafe or illegal products (23 per cent), and pornographic or sexualised content (18 per cent.)

Strikingly, almost all of the content moderation decisions relating to suicide or self-harm content were taken by just two social media platforms: Pinterest and TikTok. Pinterest accounts for almost three-quarters of all moderation decisions (74 per cent), while TikTok was responsible for just under one-quarter of decisions made (24 per cent.) See figure 1.

Figure 1: Content moderation decisions by platform



Each of Meta's platforms were responsible for only around 1 per cent of the total number of suicide and self-harm content moderation decisions: Facebook issued 157,891 SORs during this eight-month period (1.28 per cent), while Instagram published only 117,615 (0.96 per cent of all decisions.)

X and Snapchat performed even worse, with X moderating 16,705 items of suicide and self-harm content during this period. Snapchat made only 5,282 relevant decisions.

The differential effectiveness of content moderation strategies across platforms is striking, with the results illustrating a clear lack of investment and commitment from Meta's platforms to adequately target and make progress on violative suicide and self-harm content. There are substantial questions

for the company about why it is identifying and responding to only a small proportion of the harmful content actioned by other sites, including platforms with highly similar functionality.[8]

There are also significant questions about whether Instagram's response is in any way commensurate to the very high-risk profile for suicide and self-harm content on its platform. Instagram's risk profile for harmful content is substantial, with our recent research finding that one in eight posts on the platform, that were shared using well-known suicide and self-harm hashtags, actively promoted suicide and self-harm behaviour.[9] In doing so, these posts were in clear violation of the platform's community standards.

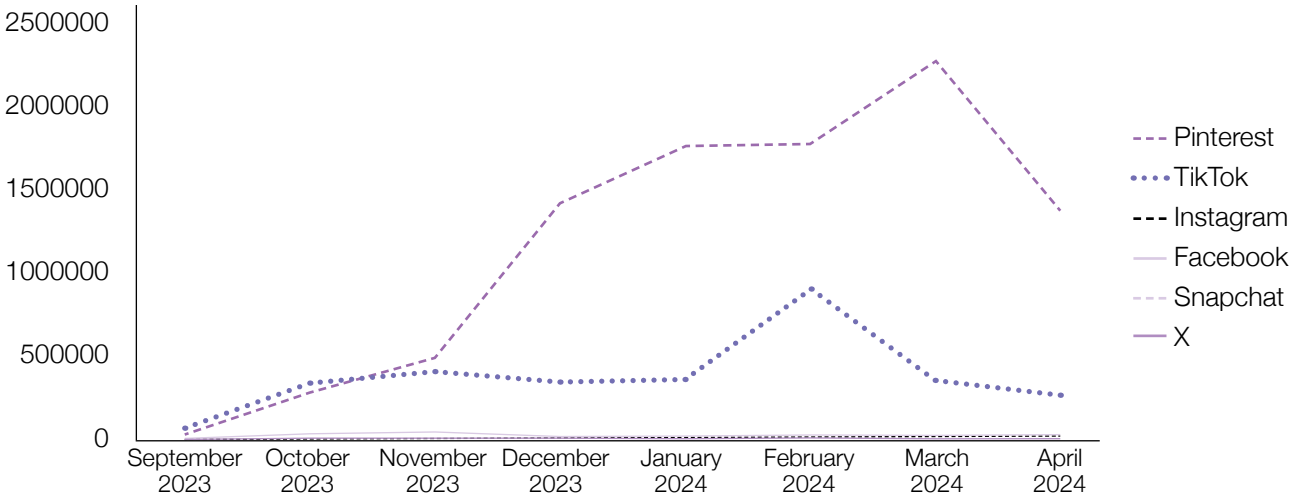
Using a broader definition of harm, we identified that almost half (48 per cent) of analysed posts were likely to cause harm to children and young people, including as a result of cumulative exposure driven by personalized recommendations of content.[10]

Regulators should be asking searching questions of both X and Snapchat. Research has previously found that Twitter/X has substantially failed to address the risks of suicide and self-harm material being distributed on its platform.[11] More recent design features, such as the Communities tab, appear to be readily exploited by users that wish to post and share violative suicide and self-harm material on X, with limited content moderation being observable.

Both Pinterest and TikTok deserve considerable credit for the proactive detection and actioning of suicide and self-harm content at a scale that is more likely to be commensurate to their risk profile of their services. This likely reflects a commercial decision to prioritise content moderation in this area.

As figure 2 shows, Pinterest has been actively increasing the amount of harmful content it has been actioning over recent months, which suggests a determined push to identify and remove relevant harmful content at scale.

Figure 2: Monthly volume of content moderation decisions recorded by platforms



8 This is particularly pertinent when Meta has demonstrated its ability to detect industry-leading volumes of violative content in other threat archetypes. For example, 90 per cent of child sexual abuse reports received by the National Center for Missing and Exploited Children (NCMEC) consistently come from Meta.

9 Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of suicide and self-harm content on Instagram, TikTok and Pinterest.

10 ibid

11 Goldenburg, A et al (2022) Online communities of adolescents and young adults celebrating, glorifying and encouraging self-harm and suicide are growing rapidly on Twitter. Rutgers: Network Contagion Research Institute

2. Do platforms take moderation decisions using automated, manual or hybrid means?

The DSA Transparency Database provides comprehensive data about how large online platforms identify suicide and self-harm content, including the extent to which platforms rely on user reports and/ or have invested in technology to assist in proactive detection and decision-making.

Unsurprisingly almost all content moderation decisions are taken at the platform's own discretion, with over 99 per cent of decisions taken on a voluntary basis. The only platform that claims it makes zero moderation decisions voluntarily is X: the platform claims that all of 16,705 moderation decisions were taken on non-voluntary grounds.

There is a clear relationship between platforms that have invested in wholly or partially automated solutions and the amount of harmful content that is detected. For example, over 93% of Pinterest's content decisions were taken on a partially automated basis.[12] Only a relatively small amount of its decisions (6 per cent) was decided by wholly automated means.

TikTok uses wholly automated means to detect and take action on the overwhelming majority of content reviewed. Six in seven of its content moderation decisions (86 per cent) were taken using wholly automated means, with just over one in eight decisions (13 per cent) taken manually. See figure 3.

Figure 3: Percentage of platform moderations decisions by moderation type

Platform	Not Automated	Not Automated %	Partially Automated	Partially Automated %	Fully Automated	Fully Automated %	Total
Facebook	13,395	*8.5*	144,496	*91.5*	0	*0.0*	**157891**
Instagram	15,792	*13.4*	1018,23	*86.6*	0	*0.0*	**117615**
Pinterest	11,281	*0.1*	845,8022	*93.9*	535,134	*5.9*	**9,004,437**
Snapchat	5,280	*99.9*	0	*0.0*	2	*0.04*	**5282**
Tiktok	395,349	*13.67*	0	*0.0*	2,497,309	*86.3*	**2,892,658**
X	16,705	*100*	0	*0.0*	0	*0.0*	**16705**
TOTAL	**457,802**	***3.8***	**8,704,341**	***71.4***	**3,032,445**	***24.9***	**12,194,588**

There are substantial questions for large platforms that are not using partial or full automation to detect suicide and self-harm content about whether they are able to adequately enforce their relevant community standards.[13] In the case of established or proposed regulatory regimes in the UK and

12 Pinterest sets out more detail about its approach in its DSA Transparency Report, including its approach to hybrid enforcement. This is where a pin is found to be violative and automated systems are then used to identify other examples of this content on the site. Pinterest (2024) Digital Services Act Transparency Report, April 2024

13 MRF believes that platforms should actively use classifiers and other forms of proactive technology to identify and action harmful content at scale, but in order to uphold free expression, such technology should be highly accurate and used in conjunction with appeals processes where an erroneous decision is made.

Canada, there must be significant doubts whether platforms such as Snap and X will be able to adequately meet their regulatory obligations while seemingly failing to proactively detect and act on harmful and potentially illegal forms of content at commensurate scale to the likely level of risk.

Questions should also be asked about the efficacy of Meta's approach to detecting suicide and self-harm content. While both Facebook and Instagram use partially automated means to detect and take decisions on the vast majority of relevant content, it is striking that Meta's approach, presumably the use of AI classifiers, detects considerably lower volumes of content than the technology deployed by Pinterest and TikTok.

Even if we combine the amount of content action by both Facebook and Instagram, Meta's classifers detect and decide on less than one-tenth of the content actioned by TikTok.

Previous research suggests that Meta-owned platforms have a significant risk profile for suicide and self-harm material, and that with the growth of video surfaces as Reels, the prevalence of suicide, self-harm and highly depressive content is increasing.[14] On this basis, it is difficult not to conclude that Meta's commitment to detect and remove suicide and self-harm content falls considerably short of what could be considered a proportionate and technically feasible response.

3. What actions do platforms take on violative suicide and self-harm content?

Our analysis shows that platforms respond to the majority of content moderation decisions by removing the relevant violative content. Across all major platforms relevant content is removed in more than four-fifths of moderation decisions (84 per cent.)

However there appears to be some significant inconsistencies between companies in how they respond to suicide and self-harm content which they opt to continue to host. In only a small handful of decisions (less than 0.2 per cent), platforms opted to take more than one measure to restrict the content, with a high degree of variation in what moderation strategies, age restrictions and other relevant safety-by-design measures platforms adopt.

While in part this reflects the considerable diversity of the social networks in scope, this data also raises substantive questions about what best practice should look like – and it has clear implications in the UK for how Ofcom is proposing to rollout its Protection of Children regulatory scheme.

In the case of Instagram, the platform claims it either removes all relevant content and/ or terminates the users' account. Interestingly, Instagram's DSA disclosures don't reflect a range of other measures which we have observed are in use on the app. For example, we have seen evidence that Instagram uses interstitial sensitivity screens and other safety-by-design measures that are designed to add friction to the user's search and recommendation experience.

In January 2024, Meta announced it would prevent age-inappropriate suicide and self-harm content from the feeds of under 18s.[15] Despite our analysis period running to April 2024, Facebook and Instagram haven't reported age-restricting any relevant content in their SoRs.

14 Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of suicide and self-harm content on Instagram, TikTok and Pinterest.

15 Sato, M (2024) Meta will hide suicided eating disorder content from teens as government pressure mounts. The Verge, 09/01/24

There appear to be significant differences in how platforms restrict access to harmful content, as shown in figure 4. For example, Facebook reports that it algorithmically downranks 12 per cent of suicide and self-harm content. Similarly, TikTok claims that it has age-restricted over 11,000 items of content, an approach not adopted by any other major platform. TikTok's rationale for deeming that content is inappropriate for children but can be shown to potentially vulnerable adults remains unclear.

Figure 4: Actions resulting from content moderation decisions, by platform

Action Taken	Facebook	Instagram	Pinterest	Snapchat	TikTok	X	Grand Total
Age restricted content					11,439		11,439
Demotion of content	18,691						18,691
Disabling access to content				5,010	12		5,022
Labelled content	145						145
Other restriction visibility			1,878,113		7,369		1,885,482
Other restriction visibility + Partial suspension of the provision of the service						16,698	16,698
Other restriction visibility + Total suspension of the provision of the service + Suspension of the account						7	7
Partial suspension of the provision of the service					24,949		24,949
Removal of content	139,044	116,587	7,125,548	249	2,848,887		10,230,315
Suspension of the account				23			23
Termination of the account	11	1,028	776		2		1,817
TOTAL	**157,891**	**117,615**	**9,004,437**	**5,282**	**2,892,658**	**16,705**	**12,194,588**

Strikingly, the large platforms opt to suspend user accounts in seemingly only exceptional circumstances. For example, TikTok terminated only 2 accounts out of 2,892,658 decisions taken. Pinterest terminated 776 accounts after taking decisions on over 9 million items of relevant content.

Partial suspensions were made in only a further 0.2 per cent of cases. This raises substantive questions about whether any of the large platforms are treating suicide and self-harm content with the severity it deserves, particularly content that is posted for malign reasons and/ or that may contribute towards long-term cumulative harm, particularly among children and young people.

Our analysis raises concerns about whether the major platforms are adequately assessing and responding to content that may reasonably be considered illegal. In the UK, suicide and self-harm content may be considered illegal where it encourages or assists suicide or serious self-harm, including by electronic means.

Excluding X, the major social networks determined that only 12 items of content were likely illegal out of more than 12 million decisions made. All of these reports were likely generated by users submitting Article 17 reports, a requirement in the DSA to enable users to report content they reasonably consider to be illegal.

In the vast majority of decisions (99.5 per cent), social networks deemed that Article 17 reports had not reached the criminal threshold but did go on to action the material as a breach of their community standards.

We found significant issues with the data reported by two of the six platforms in scope, Pinterest and X. In Pinterest's case, a substantial sum of decisions was listed as '[an]other restriction visibility', although we note that this may reflect the reporting framework initially adopted by the European Commission.

In respect of X, the company states that all of its decisions resulted in either a suspension or termination of the user's account. It appears that X has notified the Commission that 100 per cent of these decisions relate to illegal content, with no decisions at all relating to breaches of its community standards. X's reporting is a clear outlier that the DSA enforcement unit may wish to investigate further.

4. What types of content are subject to moderation decisions, and what does this tell us about whether platforms are doing enough to tackle risks?

The Transparency Database provides a detailed understanding of the types of media subject to content moderation decisions. Our analysis shows inconsistencies in how platforms moderate content, with content moderation decisions disproportionately skewed towards certain types of media over others.

In the case of both TikTok and Instagram, a strikingly low proportion of moderation decisions related to image and video-based content.

Given that both platforms are primarily video and image-based services, and that recent research has shown the risk profile on both platforms is increasingly concentrated on video-based content,[16] this is deeply concerning. Our analysis of Reels, Instagram's short-form video competitor, identified the highest risk profile of any of Instagram's major product surfaces.[17]

Our findings raise clear questions about whether large platforms are sufficiently investing in the proactive moderation of new and emerging forms of technology, including video and livestreaming functionality. There must be substantive doubts whether their content moderation strategies adequately map onto the risk profiles of their platforms.

In the case of Instagram, only 2 per cent of content moderation decisions related to video content and a further 18 per cent of decisions to images. Only one in ten (10 per cent) of TikTok's decisions related to video content, with just 4 per cent relating to images.

Interestingly, major platforms report that synthetic media already accounts for 1 per cent of content moderation decisions. This suggests that AI-generated suicide and self-harm content is already starting to appear at some scale, although we would caution that some of these results would benefit from further investigation and analysis. For example, Instagram claims that almost four-fifths (79 per cent) of its relevant content moderation decisions related to synthetic content, which either suggests that the platform is actively targeting synthetic content or there are errors in its data reporting. X claims that 100 per cent of content decisions related to synthetic content.

Pinterest's data underscores the importance of a systematic and consistent reporting approach across platforms. Pinterest has so far chosen to report all 9 million of its content moderation decisions as 'other', which seemingly stems from a decision to report its decisions as 'pins', rather than a more granular assessment of the relevant content type.

While we can realistically assume that the majority of content relates to image or video-based posts, Pinterest's approach reinforces the importance of standardised reporting measures being applied across all relevant services, not least to ensure that regulators and civil society can make effective and robust comparisons between them.

16 See for example Niu, S et al (2023) Building credibility, trust, and safety on video sharing platforms. Conference: Chi 23: CHI Conference on Human Factors in Computing Systems

17 Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of suicide and self-harm content on Instagram, TikTok and Pinterest. See also Laura Edelson's research that demonstrates the increased exposure of teen accounts to sexualised content on Reels. Cybersecurity for Democracy: Northeastern University and New York University

We encourage the Commission to take an active supervisory approach with Very Large Online Platforms, including identifying and report back on whether platforms are meeting their transparency requirements as had been envisaged.

Interestingly, TikTok reports over 1,500 decisions relating to audio content. While this is a very small proportion of the platform's overall reports, the risk profile of audio-based content is under-researched and often overlooked. Our analysis suggests that audio clips can be used promote and glorify suicidality and self-harming behaviours, including through the use of song lyrics.[18]

Figure 5: Content moderation decisions by media type

Media	Facebook	Instagram	Pinterest	Snapchat	Tiktok	X	Grand Total
Audio	0	0	0	0	1,507	0	**1,507**
Image	122,229	20,894		151	109,810	0	**253,084**
Other*	11	1,028	9,004,437	99	47,684	0	**9,053,259**
Product	6	0	0	0	0	0	**6**
Synthetic Media	19,773	92,922	0	0	0	1,6705	**129,400**
Text	10,053	0	0	2,141	2,450,582	0	**2,462,776**
Video	5,819	2,771	0	2,891	283,075	0	**294,556**
TOTAL	**157,891**	**117,615**	**9,004,437**	**5,282**	**2,892,658**	**16,705**	**12,194,588**

* From our 7-day sample we were able to see that Pinterest further defines Other as "Pin", Instagram as 'Account', TikTok as 'Profile Information', and Snapchat as 'Account', 'Multi-media' and 'Other chat content'.

Figure 6: percentage of content moderations by media type for each social media platform

	Facebook %	Instagram %	Pinterest %	Snapchat %	TikTok %	X %	Grand Total
Other	0.01	0.87	100	1.87	1.65	0	9053259
Text	6.37	0.00	0	40.53	84.72	0	2462776
Video	3.69	2.36	0	54.73	9.79	0	294556
Image	77.41	17.76	0	2.86	3.80	0	253084
Synthetic Media	12.52	79.01	0	0.00	0.00	100	129400
Audio	0.00	0.00	0	0.00	0.05	0	1507
Product	0.00	0.00	0	0.00	0.00	0	6
TOTAL							**12194588**

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

5. Do platforms identify and act on harmful content quickly enough?

Our analysis finds some significant inconsistencies in the timescales in which social networks identify and act on suicide and self-harm content. The data raises questions about whether some services appropriately triage new and emerging content types effectively.

DSA data suggests that TikTok is by far the most responsive platform to identify and decide on harmful content, with 94.4 per cent of analysed content being actioned on the same day it is posted. See figure 7. While the platform deserves credit for identifying large volumes of harmful content quickly, it is important to remember this is in the context of substantial volumes of suicide and self-harm material still being freely accessible and discoverable on the site.

There appears to be significant differences in how effectively Meta's platforms identify and respond to relevant content. While neither Instagram nor Facebook are detecting harmful content at volumes comparable to some of its competitors, Instagram identifies and decides on content much more rapidly than Facebook. Strikingly, 87 per cent of content decisions taken by Instagram relate to content posted on the same day, compared to only two-thirds (67 per cent) of decisions made by Facebook.

One in six of Facebook's decisions (16 per cent) relate to content posted over 100 days ago. The oldest piece of content actioned by Facebook had been posted in 2009.

Although Pinterest is responsible for more content moderation decisions than any other platform, the speed at which it identifies and decides on harmful content is slow. Less than one-fifth of Pinterest's content decisions (19 per cent) are made on the same day that material is posted, with almost one-quarter (23 per cent) of content having been freely available on the site for at least one year. At least 1,000 items of content had previously been available for over a decade.

While Pinterest deserves credit for investing in a significant push to identify suicide and self-harm content on its platform, and its high removal rates are likely to result in substantially improved safety outcomes in the longer-term, we caution that the platform must be able to address both newly posted content and that which has been hosted on the site for an extended period.

Among platforms that are significantly underreporting suicide and self-harm content, Snapchat claims that 97 per cent of moderation decisions relate to content posted on the same day.

X claims a 100 per cent same-day response, albeit having actioned just 0.002 per cent of the content reviewed by Pinterest.

While the DSA only tracks measures relating to the time taken to identify and moderate content, we recognise that it is also important to assess the reach of violative content – and that in some cases, this metric may be a better indicator of the efficacy of platform approaches.

While the DSA doesn't require reach metrics to be proactively shared, some platforms focus on this measure and proactively share relevant data in their voluntary reports. For example, Pinterest publishes data relating to the reach of violative suicide and self-harm content: in Q4 2023, the platform stated that 75 per cent of relevant violative content was actioned before anyone had viewed it, with only 1 per cent having been viewed by at least 100 accounts.

In practice, it may be desirable to assess the performance of regulated services by both of these metrics. In the case of platforms and/or product services where there is an increased risk profile associated with the immediacy of content, such as X or on Instagram Stories, the time taken to identify and action content is likely to warrant increased focus.

Both metrics are likely to take on increased significance as generative AI reduces the cost and technical barriers associated with producing new suicide and self-harm content. Some malign actors may increasingly look to post substantial amounts of content to seek to overwhelm the trust and safety arrangements of platforms, with resulting risks to the time taken to identify content and am increased risk of algorithmic amplification of harmful content to users.

Figure 7: percentage of same-day content moderation decisions

Platform	Decisions made same day as content posted (%)
TikTok	94.4
Instagram	87
Facebook	67
Pinterest	19
Snapchat *	97
X *	100

* Small sample sizes, reflecting the very low number of decisions taken during the analysis period.

6. Can we compare DSA data with voluntary transparency reports?

While the DSA provides the first set of legally enforceable transparency measures, most of the major social networks simultaneously issue voluntary transparency reports. Voluntary transparency reports have enabled tech firms to set out the scale and effectiveness of their response to online harms but have been widely critiqued as a form of 'transparency theatre'.[19]

Our analysis raises questions about the reliability and integrity of some voluntary transparency reports. While we weren't able to analyze the voluntary transparency reports and SORs submitted by TikTok, Pinterest , Snap and X – either because their reports are issued with significant lag times or provide insufficient information to allow full analysis – comparisons can be made with the voluntary reports submitted by Instagram and Facebook.

Strikingly, our analysis has been unable to reconcile the data submitted by Meta with the claims made in its voluntary transparency reports.

19 Douek, E (2020) The Rise of Content Cartels: Urging Transparency and Accountability in Industry Wide Content Removal Decisions. Knight First Amendment Institute at Columbia University.

According to its legal filings under the DSA, Instagram recorded 39,868 content moderation decisions in Q1 2024. However, this is less than 10 per cent of the decisions we would expect to see logged in the Transparency Database (464,000), if the 5.8 million items of relevant content that the company claims to have actioned in its voluntary reports over this time were evenly apportioned to its EU user base.

Similar discrepancies are apparent when analyzing Facebook's legal and voluntary disclosures. In its DSA filings, Facebook reports that it made 54,148 relevant decisions during Q1 2024. However, this is only 7.6 per cent of the decisions we would expect to see in its disclosures, if the 7.1 million items of content that Facebook says it actioned in its voluntary reports[20] were evenly apportioned to its EU user base.

There are, of course, a number of potential explanations that could explain these potential discrepancies. For example, Meta could have made errors when processing one or both of its data sets, or the company may use different definitions of 'actioning' content when making its legal and voluntary disclosures. It is also important to note that Meta is the only major platform that has externally audited its transparency data, commissioning EY to review its transparency report in 2021.[21]

That said, the discrepancy between Meta's legal and voluntary disclosures is significant – and it cannot be readily reconciled through external analysis. The magnitude of this inconsistency raises substantive questions about the reliability and accuracy of Instagram and Facebook's reporting flows. We therefore urge the European Commission to further investigate these issues and ask Meta for an explanation and reassurance about the quality and integrity of its data.

With other regulatory regimes also beginning to establish their transparency and supervisory arrangements, including Ofcom in the UK and Coimisiún na Meán in Ireland, considerable caution should be applied to ensure that company disclosures are robust and accurate.

20 Meta (2024) Community Standards Enforcement Report, Q1 2024
21 Meta Newsroom (2022) Community Standards Enforcement Report Assessment Results

Discussion

Our findings present several important implications for effective regulation of large online platforms – and underscore the importance of strong transparency measures to support legislators, regulators and civil society in the pursuit of better online safety outcomes.

In this section, we discuss a number of these implications and make several recommendations to support regulators in the UK and EU.

1. Moderation of harmful content is uneven and inconsistent, and should be a primary focus for regulators

The DSA Transparency Database provides a detailed understanding of the approach taken by major online services to enforce their terms of service, and in respect of harmful and/or violative suicide and self-harm content, demonstrates substantial inconsistencies in how platforms respond to their respective risk profiles.

The results suggest that a number of major platforms are making inadequate investment in their overall response to harmful content, with major platforms including Instagram, Facebook, X and Snapchat fundamentally failing to identify and moderate content at sufficient scale. In the case of Meta's platforms, it is challenging to conclude that Instagram and Facebook are responding either adequately or commensurately to the risk profile of their services.

UK and EU regulators should pay close attention to data that suggests major platforms have been insufficiently incentivised to adopt suitably adequate and consistent content moderation approaches when introducing new or higher risk forms of functionality.

For example, Instagram reports that only 2 per cent of its content moderation decisions related to video content: this is despite extensive and growing research that Instagram's short-form video offer, Reels, is exceptionally high-risk for suicide, self-harm and other forms of harmful and/or age-inappropriate material. Reels now accounts for 50 per cent of time spent on the platform.[22]

In the UK context, this data has significant implications for Ofcom's Protection of Children proposals.[23] Ofcom must do more to ensure that its regulatory proposals adequately require companies to invest in the detection of harmful content on the highest-risk part of its services, and the regulator should be prepared to adopt outcome-based measures around content moderation on high-risk product surfaces.

22 In Meta's Q1 Earnings call, CEO Mark Zuckerberg stated: 'On Instagram, Reels and video continue to drive engagement, with Reels alone now making up 50 per cent of the time that's spent within the app.'
23 Ofcom (2024) Protection of Children consulation

Our analysis adds to growing concerns that Instagram, TikTok and other video-based services are failing to invest in adequate forms of proactive detection and content moderation technology, with video-based technology seemingly being rolled out without commensurate investment in technology to identify and mitigate the reasonably foreseeable risks that may result from it.

2. Content moderation must accompany, and in some cases facilitate, a step-change in safety-by-design

Content moderation is a hugely important part of protecting users from harmful or age-inappropriate experiences. It is crucial that regulators understand that moderation is not only a means detect harmful content, but that when it is implemented effectively, it can also actively mitigate and deter harmful material from being posted in the first place.

However, it is also important to stress that content moderation is only one part of a truly effective platform response. Platforms must also be incentivised to focus on and invest in effective safety-by design-measures to mitigate the risks of harmful content being posted, shared or algorithmically recommended.

In the UK, Ofcom sets out the importance of effective moderation processes to underpin a number of its central safety-by-design proposals. For example, the regulator proposes that online services that are medium- or high-risk for primary priority material content, including certain types of self-harm content, must design their recommender systems to filter out relevant content from children's feeds.

In practice, this is enforceable only in cases where platforms are aware of content that is likely to be primary priority or priority content, either because it is undergoing content moderation, it has already been flagged for moderation, or because there is other information that suggests the threshold has been met. In effect, many of Ofcom's safety-by-design approaches are only as effective as the content moderation mechanisms that will be used to operationalise them.

As a direct extension of regulatory design, content moderation and safety-by-design are set to become inextricably and increasingly linked. Strong content moderation arrangements therefore become intrinsic to the delivery of better online safety regulatory outcomes.

3. Transparency frameworks must be seen as a means to an end, and should draw on the broadest possible set of harm indicators

The DSA transparency framework provides a rich and detailed understanding of how platforms make content moderation decisions, and already enable civil society, regulators and legislators to make evidence-based conclusions about the efficacy of major platform approaches.

It will now be important that the transparency regime is proactively used by EU, UK and global regulators, with DSA data on the adequacy, efficacy and consistency of moderation approaches actively informing their supervisory and enforcement approaches. Put simply, transparency mechanisms must be seen and used as a means to drive better regulatory and safety outcomes, not simply as an end in their own right.

As other regulatory regimes develop their own transparency approaches, there is an opportunity for other regulators to compliment and build on the DSA regime. In the UK, Ofcom should look to establish a broad set of transparency metrics, including a focus on impact, risk and process metrics.[24] Ofcom should look to assess a broad range of impact measures, including the reach of violative material.

We also encourage the UK regulator to attach particular weight to measuring the experience of and exposure to online harms amongst distinct sets of service users, including children and other potentially vulnerable groups.[25]

4. Corporate transparency data needs to be subject to audit, with a duty of candour on regulated companies

The DSA enables a decisive shift away from a reliance on voluntary transparency reports, which have been roundly and legitimately critiqued as a form of 'transparency theatre'.[26] In too many instances, platforms have been able to choose and publish self-selected metrics that create a 'false patina of legitimacy'.[27]

Put simply, mandatory transparency frameworks can ensure that transparency is delivered for the public good, not simply as an extension of platform PR strategies.[28] This is particularly important in the context of our research raising substantive questions about the validity of data shared by companies in their voluntary transparency reports.

24 World Economic Forum (2024) Making a Difference: How to Measure Digital Safety Effectively to Reduce Risks Online. White Paper.

25 For example, regulators could mandate companies to commission user experience data that assesses exposure to harm types and that measures its effects, such as the BEEF framework championed by Meta whistle-blower Arturo Bejar.

26 Douek, E (2020) The Rise of Content Cartels: Urging Transparency and Accountability in Industry Wide Content Removal Decisions. Knight First Amendment Institute at Columbia University.

27 ibid

28 Meta's approach to platform transparency has drawn critiques around its usage for PR and reputation management. For example, in June 2023, Meta responded to a Wall Street Journal investigation into child sexual abuse on Instagram by sharing a metric that downplayed the prevalence of child sexual abuse on its service. This directly contradicts the claim made over many years in its voluntary transparency reports that it isn't technically possible to do this.

Since Meta first started publishing voluntary transparency reports, it has claimed it is unable to provide estimates for the prevalence of illegal content on its services. However, as part of its rebuttal to a Wall Street Journal investigation into child sexual abuse on Instagram in 2023, the company contradicted this position and publicly shared a prevalence metric that downplayed the prevalence of child sexual abuse material on the service.

While there may be perfectly valid explanations for the significant discrepancies identified in our report, and clearly there are no regulatory or legal implications associated with companies that make voluntary disclosures that may or may not be accurate, our findings underscore the importance of regulators establishing confidence that any information being applied to them to *them* is highly accurate and robust.

While the DSA makes provision for independent audits,[29] audit and quality assurance powers are noticeably absent from other regimes, including in the UK. With the new UK Government setting out its intention to introduce new legislative measures to strengthen the Online Safety Act as soon as possible,[30] there is an important opportunity to build in independent audit and transparency powers into revised OSA arrangements.

We see a particular role for a new overarching 'duty of candour' on regulated companies:[31] this would introduce a clear and enforceable duty on tech companies to cooperate with regulatory, legal and public investigations. Sanctions could be imposed on large tech companies if they obfuscate, provide knowingly inaccurate information, or otherwise impede or delay the work of official investigations and regulators.

29 The DSA requires that independent auditors must assess compliance of VLOPs and VLOSEs at least annually. The first set out audits should be completed by August 2024 and published no later than three months later.
30 Labour (2024) Change: Labour Party Manifesto 2024
31 Molly Rose Foundation (2024) General Election Manifesto: Five commitments that will transform children's online safety and well-being

MOLLY ROSE FOUNDATION

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

Written by Andy Burrows, Molly Rose Foundation
Data analysis by Alex Waddington, Whetstone Communications

For more information, please contact
a.burrows@mollyrosefoundation.org

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**