Molly Rose Foundation response to Ofcom's consultation on protecting people from illegal harms online

February 2024

Summary

- The Molly Rose Foundation (MRF) welcomes the opportunity to respond to Ofcom's consultation on its regulatory scheme on illegal harms.

- Given our focus on suicide prevention, our response focuses particularly on the risks of exposure to illegal content and behaviour related to suicide and self-harm, and on the systemic design of the regulatory scheme.

- Overall, we have significant concerns about Ofcom's proposed approach and its limited potential to protect users from preventable harm. We strongly encourage Ofcom to re-set its approach to ensure it builds a targeted and effective regulatory approach from the outset, and to prevent a set of design choices being baked into the regime that may significantly and unnecessarily constrain its longer-term impact and effectiveness.

- Ofcom has made a set of strategic decisions about how it intends to operate its regime that seem highly likely to blunt its impact and limit the protections available to vulnerable groups. We are particularly concerned that the regulator's proposed approach to proportionality, evidentiary thresholds and the precautionary principle will result in a regulatory scheme that is at best slow but at worst unable to respond to rapidly changing online harm dynamics.

- The design of the Act means that Ofcom's approach is likely to grant large platforms a 'safe harbour' while setting out regulatory requirements that are substantially less demanding

than what they currently do. Put simply, this approach is unlikely to significantly disrupt or reverse the scale and magnitude of many illegal harms.

- It is difficult to reconcile Ofcom's approach with the expectation of a systemic, outcome-focused and risk-based regime, based on outcomes, which was arguably the clear intention of Parliament when it passed the Online Safety Act.

- We specifically encourage Ofcom to revisit its approach to suicide and self-harm material. As it sBad in tands, the Codes of Practice only contain one measure relating to platform recommender algorithms, despite these being the major driver of suicide and self-harm related illegal harms.

- The regulator sets out a reasonably well-developed albeit still incomplete understanding of the risk profile for suicide and self-harm offences, but then fails to recommend a set of corresponding measures that are commensurate to and appropriately able to tackle the scale and nature of the harms that result.

- We encourage the regulator to adopt a bolder and more ambitious approach, otherwise it risks implementing a scheme that acts as largely a sticking plaster to address the risks posed by suicide and self-harm content, rather than offering strong and systemic protections in the face of inherently preventable harm.

Our response

MRF's response is structured as follows:

- In **section one**, we set out our overarching concerns about Ofcom's proposed approach;
- In **section two**, we focus on the risk factors and drivers of relevant illegal harms, including how platform design features increase the risk of exposure to suicide and self-harm material and the susceptibility of users to its effects;
- In **section three**, we respond to Ofcom's draft Codes of Practice and set out a range of recommended measures for the regulator's consideration;
- In **section four**, we make the case for the Codes of Practice to form part of a broader, more targeted approach to harm reduction, with reference to the evidence and learnings of Meta whistleblower Arturo Bejar.

Section 1: Overarching concerns

- The Molly Rose Foundation (MRF) has significant concerns about Ofcom's proposed approach and its likely effectiveness in reducing exposure to illegal harms and improving the user experience.

- In its consultation materials, Ofcom states that its first Codes 'represent a strong basis on which to build a more comprehensive suite of recommended measures to reduce the risk of harm to users over the long term.' Ofcom also explicitly states that 'our first Codes aim to capture existing good practice within industry […] especially for services whose existing systems are patchy or inadequate.'

- Ofcom has signalled that it sees its proposed approach as a first iteration, and that it expects to expand on this framework in future iterations. While MRF understands the need to adopt an iterative approach, this first set of measures essentially does little more than package together a set of existing best practice approaches (and even then, recommends a set of measures that fall short of what most large platforms currently undertake, in some areas by a substantial margin.)

- If Ofcom's first iteration of the Codes doesn't adequately capture existing platform responses, it is difficult to envisage how these measures can meaningfully disrupt priority and non-priority illegal harms (many of which continue to grow rapidly in their scale and complexity).

- Ofcom appears to be affording itself the luxury of adopting a gradual iterative approach to tackling reasonably foreseeable harms over the long-term'=. While we repeat that the regulator's decision to iterative approach is entirely legitimate, its choice to adopt such a low bar in its initial approach, and the inherent nature of its gradualism, mean that early iterations of this Code will likely amount to nothing more than a sticking plaster approach. A bolder and more ambitious approach commensurate with the scale and nature of illegal harm is urgently required.

<u>Codes of Practice and legislative intentions</u>

- Ofcom's proposals appear to fall considerably short of the reasonable expectations of civil society and those with lived experience of preventable online harm; but more pressingly, suggest a clear disconnect between the likely outcomes of Ofcom's approach and the legislative intentions of Parliament when it passed the Act.

- Under Ofcom's proposed approach, the regulator will consider an online service to be compliant with their illegal safety duty if they implement the measures set out in the relevant Code. This closely mirrors the approach set out in s41(1) of the Act, which sets out

that a provider 'is to be treated as complying with a relevant duty if the provider takes or uses the measures described in a code of practice.'

- It is manifestly not the case that Parliament envisaged platforms being provided with 'safe harbour' status if they meet a set of provisions in Ofcom's Codes that are insufficiently stringent to meet the Act's stated objectives. S41 of the Act clearly requires Ofcom to be confident that its measures will be suitably robust, and Schedule 4 determines that the Code of Practice must be compatible with the pursuit of the online safety objectives, not least that regulated services must have 'effective and proportionate' systems and processes in place, and should be 'designed and operated so as to protect individuals in the United Kingdom who are users of the service from harm.'

- As it stands, Ofcom's set of recommended measures are unlikely to produce a substantial reduction in illegal content and activity; result in platforms developing 'effective and proportionate systems that are capable of protecting users from harm; nor reasonably meet the online safety objectives set out in schedule 4. In some areas, the recommended measures are so underdeveloped that it is entirely reasonably foreseeable that the scale of and exposure to some types of content may actually continue to increase.

- Ofcom's approach therefore delivers an approach which fails to deliver the stated ambitions of the Act, and that in some circumstances, could actually enable companies to scale back their existing safety approaches while still remaining free from the risk of enforcement action.

- It should surely be evident to the regulator that the 'safe harbour' provisions were only intended to apply in circumstances in which Ofcom's recommended measures clearly satisfied the aims of the legislation, and that meaningful improvements in online safety outcomes would result.

- In its consultation response, Ofcom should therefore set out how it considers its gradualist approach to be consistent with the aims of the Act, and how it intends to mitigate the obvious risk of perverse outcomes associated with its initial iterations of the Code and approach.

<u>Proportionality</u>

- In preparing its draft Code, Ofcom has adopted an exceptionally high threshold to determine if a safety approach is proportionate and therefore suitable for inclusion as a recommended measure. Ofcom has opted not to recommend measures that have the potential to prevent harm where it deems there to be insufficient evidence to determine its likely effectiveness or where it perceives uncertainty as to the capacity of regulated providers to adopt them.

- This has led to a draft code that is manifestly insufficient to disrupt the scale and nature of many of the priority harms in scope. For example, there appears to be only one recommended measure that directly targets the algorithmic application of suicide and self-

harm content, despite this being identified in volume three as a primary driver of and high-risk facilitatory mechanism for relevant harms.

- In its approach to proportionality and evidence, the regulator appears to have adopted a standard of proof that is more consistent with that used in a criminal regime ('beyond a reasonable doubt') than for a civil or regulatory regime ('on the balance of probabilities.') This burden of proof seems unnecessarily high.

- While there is clearly a not inconsiderable risk of litigation from regulated companies, our assessment is that Ofcom's overly risk averse approach arguably risks prioritising the interests of industry over service users. Furthermore, Ofcom's approach appears difficult to reconcile with a reasonable reading of the Act, and the parliamentary discussions surrounding it.

- We also have significant concerns that Ofcom's approach risks creating a slow and cumbersome process that is responsive to, rather than appropriately ahead of, the emerging risks and opportunities of new technologies.

- In its first iteration of the Code, Ofcom's approach to proportionality and its application of high evidentiary thresholds has resulted in significant omissions, not least the absence of measures in respect of self-generated images. We note that Ofcom has signaled it envisages adopting relevant measures in future iterations; but given the well-established understanding of the risk profile and the efficacy of relevant platform responses, it is surprising that sufficient measures were not recommended in this first iteration.

- Similarly, Ofcom concluded it has insufficient evidence 'at this stage' to recommend the hashing of terrorist content, despite this being a widely adopted and demonstrably effective approach to detect and remove terrorist material.

- We are concerned that Ofcom's approach will therefore result in the Codes being highly reactive to the emerging risks and opportunities posed by new technologies, including the increasing use of generative AI by regulated user-to-user services. MRF anticipates that generative AI will drive a significant intensification of the risk profile associated with suicide and self-harm, with significant adverse impacts likely in the immediate to medium-term. AI-generated suicide and self-harm content is already being posted to major social media sites, including Instagram, TikTok and Pinterest.[1]

- It seems eminently plausible that it might take several regulatory cycles before Ofcom is able to determine that it is proportionate to respond to the risks posed by generative AI; and is able to identify the efficacy of relevant responses and derive sufficient evidence to recommend these measures in its codes. Ofcom's approach is also likely to result in a significant lag time before it is able to recommend potential safety solutions enabled by these emerging technologies.

[1] Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest. London: Molly Rose Foundation in partnership with The Bright Initiative by Bright Data

- More broadly, we are concerned that Ofcom's approach may result in disincentives for larger regulated companies to invest in innovative new trust and safety approaches, and where new technological solutions are developed, to promptly share evidence of their effectiveness.

- While Ofcom has well-developed supervisory arrangements and information disclosure powers in this regard, if the regulator is required to have to routinely rely on such mechanisms, this will likely extend the timescales associated with being able to identify and recommend relevant measures in future iterations of the Code.

- Ofcom has suggested in discussions with civil society that it these chilling effects can be effectively mitigated by further innovation among third-party providers, and through the continued rapid growth of the safety tech sector. Given the economies of scale and scope associated with the largest regulated companies, this appears to be a highly optimistic assessment.

- However, Ofcom's approach and its emphasis on recommended measures does introduce a risk of potential anticompetitive effects, with larger companies incentivised to focus their acquisition strategies on safety tech providers that could develop new technological solutions (and that could in turn add to their regulatory burden).

- In its consultation response, we therefore encourage Ofcom to share its assessment of the wider market effects of its proposed approach, including how it intends to work with regulators such as the CMA to address potential adverse market impacts.

<u>Precautionary principle</u>

- We are deeply disappointed that Ofcom has opted to develop its codes without appropriate consideration of the precautionary principle. Invoking and applying the precautionary principle carries a general presumption that the burden of proof 'shifts away from the regulator having to demonstrate the potential for harm towards the hazard creator having to demonstrate an acceptable level of safety.[2]

- The precautionary principle creates the 'impetus to take decisions notwithstanding scientific uncertainty about the nature and extent of the risk'.[3] As a well-established regulatory approach, the precautionary principle is a sound basis to develop regulatory measures in markets where there is a clear and pressing need to address harms that result from the functioning of regulated services, but where the evidence base in respect of the mechanics and drivers of harms continues to develop.

- Given the scale and extent of priority harms referenced in the Code, and the well-observed evidentiary challenges associated with demonstrating a causal relationship between social

[2] Interdepartmental Liaison Group on Risk Assessment (2002) The Precautionary Principle: Policy and Application. London: HM Government

[3] ibid

media and many of the illegal harms in scope, Ofcom's approach has the effect of being overly cautious in favour of regulated companies rather than service users.

- It seems clear from both the legislation and the relevant parliamentary debates during its passage that Parliament envisaged that a precautionary approach would be applied, with user safety expected to take precedence in Ofcom's approach. We assert from parliamentary debates that ministers clearly envisaged that the regulator would presume to take measures commensurate with the nature and overall volume of illegal content in the first instance, but with the ability to relax measures should they later proved to be unnecessary or no longer proportionate.

- Our read is that Ofcom could choose to adopt a more substantive precautionary approach within the current statutory framework, not least given the latitude afforded by the provisions set out in sections 10(4) and 236(1.) This interpretation would enable Ofcom to adopt a more bold and ambitious approach that can more effectively respond to the nature and magnitude of the harms in scope. The Act makes no mention of the evidence on which Ofcom must base its recommendations for measures in the codes, other than a requirement that the measures must be technically feasible (Schedule 4(2).

- As it stands, we are concerned that Ofcom's approach risks actively conflating the 'absence of evidence of risk' with 'evidence of the absence of risk'. The regulator risks proceeding with an approach that will constrain its ability to recommend appropriate measures in this and future iterations of the Code, an approach that seems poorly suited to delivering effective long-term harm reduction.

- Ofcom could reasonably interpret its powers under s10(4) to assume an approach more actively informed by the precautionary principle. For example, it could choose to frame its recommendations according to a more outcome-based set of expectations, in which platforms are required to identify and implement suitable and sufficient measures that target specified harms.

- This approach would more effectively contribute towards a harm reduction framework that is geared towards delivering continual improvements in the risk profile (see section 4). For example, Ofcom could specify 'measures' that require platforms to reduce exposure to specified harms over time, and that could be incrementally tightened in each future iteration of the Code.

- In its consultation response, we encourage Ofcom to set out why it has not decided to adopt a more substantive precautionary principle approach in its first iteration of the Codes, and to clarify what if any barriers it perceives exist in the legislative framework that would prevent it from adopting such an approach when developing its scheme.

- In particular, the regulator should explain why it has opted not proceed with a precautionary principle approach when the evidential barriers associated with demonstrating the causal mechanics and relationships of online harms makes this approach manifestly well suited to delivering the regime's objectives, and to delivering immediate progress on harm reduction measures.

Economic application of proportionality

- We encourage Ofcom to provide further information about how it has approached the linked issues of proportionality and the economic basis for choosing to recommend measures or not.

- It very much appears that Ofcom has chosen to take a precautionary approach to imposing measures on industry (opting not to recommend measures where it has doubts about the proportionality of such measures on small and medium-sized firms.) In contrast, it seems there is a requirement for the costs associated with user or societal harms to be demonstrably identified and expressly proven as a precondition for the proportionality of acting on relevant harms to be met.

- This gives reasonable grounds to assume that Ofcom is inadvertently applying its proportionality test in a way that gives the balance of doubt to industry, but not users experiencing or at serious risk of illegal harm.

- In its consultation response, Ofcom should set out further information about how it is balancing the risks to users with the costs of recommending measures to tackle them. This should include a description of the economic model that informs and is actively underpinning its approach.

- In particular, Ofcom should articulate and justify its approach to how it assesses the magnitude of and costs associated with priority harms, including the economic calculations that inform whether it determines that a measure being recommend to tackle a relevant harm is proportionate.

- This should include the regulator's projection of the likely impact of the first iteration of codes on the overall exposure to and impact of the priority harms in scope.

- Clarity on Ofcom's methodology is important, not least as a range of risk management and modelling approaches can arrive at very different outcomes. Ofcom's calculations may be strongly different based on the values it has chosen to adopt.

- For example, we note that Ofcom's consultation materials assess the social and economic cost of a death by suicide as £1.67 million in 2009 prices (£2.23 million in 2023 prices). However, there is increasing evidence that the methodology used to inform this calculation is highly problematic and 'too flawed for it to continue to be used.'[4] There is now an increasing acceptance in academia and among risk economists that the standard UK model for assessing the value of a prevented fatality significantly understates the social and economic costs of an avoidable death, particularly in respect of adolescents and young adults.[5]

[4] Thomas, P (2018) Calculating the value of human life: safety decisions that can be trusted. Policy report. Bristol: University of Bristol. This is because the standard measure, UK VPF, is unrelated to the length of future life and therefore implies the average value of a future day is much greater for an aged person than a young person. This method is consequently poorly suited to quantifying the value of internet-related harms such as the deaths by suicide of young people.

[5] ibid

- The adoption of a J-value method raises the estimated average value of a human life, and the corresponding economic justification for recommending measures that tackle preventable fatalities, more than four-fold. According to the J-value method, the value of a preventable fatality was set at £8.6 million in 2015 prices (£11.3 million in 2024 prices.)[6]

- Suicide related Internet use has been reported in almost one-quarter (24%) of deaths by suicide among young people aged 10 to 19, equivalent to 43 deaths each year.[7] Applying J-values to this data therefore results in an estimated social and economic cost of internet-related deaths by suicide among young people of £486 million per year.

- There is therefore a clear and compelling case for Ofcom to reflect the J-value model when determining the proportionality of recommending relevant measures.

Application and balancing of fundamental rights

- We have significant concerns about how Ofcom is interpreting fundamental human rights in the development of its regulatory scheme, particularly the right to free expression.

- While Sections 22 and 33 of the Act require Ofcom to have regard to freedom of expression when deciding on and implementing its safety measures and policies, Ofcom's approach seems to disproportionately focus on the fundamental rights of speakers, while inadequately considering the chilling impacts of harmful speech (or taking insufficient steps to prevent harmful speech) on the right to free expression and association of other users.[8]

- Ofcom is in effect interpreting section 22 as a measure that constrains the overall ambition of its regulatory scheme. In multiple parts of volume 2, the regulator's approach cites adverse impacts on free expression as grounds not to proceed with recommended measures such as risk scoring, user blocking and some uses of keyword detection.

- In this respect, we are concerned that Ofcom's approach may actually weaken the right to free expression and association for some groups at disproportionate risk of online harms, including women and girls, LGBTQ+ groups, and those with one or more protected characteristics.

- Internal Instagram data commissioned by the whistle-blower Arturo Bejar[9] demonstrates that the failure of the company to adequately prevent teen users being exposed to unwanted

[6] ibid
[7] Rodway, C et al (2022) Online harms? Suicide related online experience: a UK-wide case series study of young people who died by suicide. Psychological Medicine, 53(10), pp1-12
[8] Woods, L (2024) Ofcom's approach to human rights in the illegal harms consultation. London: Online Safety Act Network
[9] A copy of this research, the Bad Experiences and Encounters Framework (BEEF Framework) can be found in appendix one of this response. The research was undertaken among Instagram users in June and July 2021

harms has had an adverse impact on the right to free expression and association. For example, almost three in users aged 13-15 (28%) said that being exposed to self-harm content in the previous week had discouraged them from posting on the site. [10]

- ECHR case law is clear that the failure to provide a safe environment for groups to express themselves – which attracts positive obligations under Article 10 - constitutes an infringement of the free expression rights of victims and those who share their relevant characteristics.[11]

- We also remind the regulator that Article 8 imposes positive obligations in respect of the physical and psychological integrity of an individual from other persons,[12] particularly where that person is a child.[13] It is difficult to conceive how Ofcom's limited set of measures in relation to several priority harms, but of most relevance to us offences relating to suicide and serious self-injury, are consistent with the positive obligations under Article 8 to create a suitable and sufficient legal framework that is both in place and being implemented effectively. [14]

- As the regulator is aware, Ofcom is subject to section 6 of the Human Rights Act, which specifies that it is unlawful for public authority to act in a way which is incompatible with Convention rights. The regulator could therefore be subject to a challenge, including from relevant tech accountability and child protection groups, where there are reasonable grounds to conclude that its obligations under Articles 8 and 10 have not been met.

- Ofcom should therefore be prepared to review its recommendations, taking into account the weight of the rights violations against company revenue, and be able to prepare a final set of recommended measures that it considers are reasonably in accordance with its obligations under ECHR (including the positive obligations expected of it.)

<u>Consultation process</u>

- We would like to express our concern that the size and complexity of Ofcom's consultation has caused significant accessibility and resourcing challenges for civil society groups. Given the implied importance of civil society groups to provide evidence that can inform and extend Ofcom's understanding of the issues, this complexity is likely to have had detrimental impacts on the nature and range of evidence submitted, and in turn on the strength of the final proposals.

---

[10] ibid

[11] Online Safety Act Network (2024) Statement on the Illegal Harms Consultation. London: Online Safety Act Network

[12] European Court of Human Rights (2020) Guide to Article 8: right to respect for private and family life, home and correspondence. Strasbourg: ECHR

[13] KU vs Finland. European Court of Human Rights (2015) Internet case law of the ECHR. Strasbourg: ECHR. This is discussed further in Burrows, A (2020) How to Win the Wild West Web: Six tests for delivering the Online Harms Bill. London: NSPCC

[14] O'Keefe vs Ireland. European Court of Human Rights, Grand Chamber, Application Number 35810/09, 28/01/2014

- We are also concerned that the process has provided few if any meaningful mechanisms for those with lived experience of online harms to submit their views or feedback. We would remind the regulator that the experience of those affected by online harms should be central to its approach.

- There is a manifest risk that some people and groups with lived experience of harm have felt unable to participate in a process that appears poorly designed for them. This not only runs contrary to the logic of securing good consultation outcomes, it carries a risk that the process is considered exclusionary or even re-traumatising for some groups or individuals with lived experience of online facilitated illegal behaviour.

Section 2: risk factors and understanding of drivers of harm

- This section of our response focuses on Ofcom's understanding of the drivers and dynamics of illegal content, with particular focus on the risk profiles and register of risks set out in volume 2 of the consultation.

- In its risk profiles, Ofcom sets out a range of ways in which platform design choices and product features may facilitate exposure to illegal content, including suicide and self-harm material that incites, instructs or otherwise encourages users to engage in serious self-harm or suicidal behaviours.

- We strongly welcome Ofcom's proposed approach that recognises there are a range of drivers that facilitate exposure to illegal suicide and self-harm material, but that may also increase the susceptibility and vulnerability of certain groups of users to illegal harm, for example those with pre-existing mental health conditions.

- In its finalised register of risks, we encourage Ofcom to expand on this approach and explicitly set out that regulated companies must consider the *broadest* possible set of ways in which users may be exposed to or otherwise become more susceptible to the risks in respect of illegal self-harm and suicide content on their sites.

- This approach is wholly consistent with Ofcom's statutory duties, and we consider it actively necessary to support companies in meeting their requirements to risk assess and prevent their platforms being used for the commission or facilitation of priority offences, as set out in sections 9(5)(c) and 10(b).

- This proposed approach envisages that it is appropriate and necessary to tackle the risks posed by illegal suicide and self-harm material upstream. By focusing on the risk profile before the criminal threshold is necessarily reached, this approach is also broadly analogous with Ofcom's approach to sexual grooming.

- In accordance with the approach set out above, Ofcom's register of risks should explicitly set out the risk factors associated with how the design and operation of online services may reasonably facilitate or enable each of the following:
    - the discovery of or exposure to content that incites, instructs or encourages serious acts of self-harm or suicide;
    - behaviours that may reasonably incite, instruct or encourage serious acts of self-harm or suicide;
    - the exposure of users to content, whether illegal or otherwise, that may reasonably increase the vulnerability of and susceptibility of users to illegal content relating to suicide and self-harm (and its effects);
    - the enabling of service users to identify or communicate with each other in a way that results in increased exposure to or susceptibility in respect of illegal suicide and self-harm content.

- In the rest of this section, we provide additional evidence that should support Ofcom in the development of an expanded register of risks.

- We also note that the evidence base in respect of suicide and self-harm content is still actively developing, and that the available evidence is less developed compared to other harm archetypes, for example CSEA. In this context, we wish to remind the regulator that it should resist assuming the absence of evidence or harm with the absence of harm altogether. (and re-assert the importance of the precautionary principle.) We encourage Ofcom to set out likely risk dynamics where it is reasonable to assume that harm may take place.

Evidence on the scale, nature and impacts of suicide and self-harm content

<u>The scale of and exposure to suicide and self-harm online content</u>

- Internal industry data supports Ofcom's findings that adolescents and young adults are more likely to be exposed to self-harm or suicide content than the population as a whole. For example, an internal survey of 13-15 year olds using Instagram, commissioned and subsequently leaked by the Meta whistleblower Arturo Bejar, found that 6.7% of the platforms users had seen someone harm themselves, or threaten to do so, in the previous seven days.[15]

- A substantial minority of teen users are being exposed to potentially harmful suicide or self-harm content on a frequent or even daily basis. According to the internal Instagram Bad Experiences and Encounters survey, more than two-thirds of those who had seen suicide or self-harm material had seen multiple items of content in the previous week. One in nine young teens aged 13-15 (11.1%) had seen at least ten items of self-harm content during that period.[16]

- Research conducted by the Molly Rose Foundation has found that substantial amounts of harmful suicide and self-harm content remain readily accessible and discoverable on major social networks.[17] Almost half of the most engaged posts on TikTok (49%) and Instagram (48%), and that were posted using well-known suicide and self-harm hashtags, contained material that promoted or glorified suicide and self-harm, referenced suicide ideation, or otherwise contained intense themes of misery, hopelessness and depression.

- Among the harmful posts we analysed on Instagram, two-thirds contained material that promoted or glorified suicide and self-harm (in clear violation of Instagram's community

[15] The Bad Experiences and Encounters Framework research can be found in appendix one
[16] Ibid
[17] Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest. London: Molly Rose Foundation in partnership with The Bright Initiative by Bright Data

standards.) While our research did not expressly set out to determine whether harmful content met or exceeded the criminal threshold, we assessed that a substantial minority of of these posts likely did.

- Our research identified a differential risk of exposure to suicide and self-harm content across different product surfaces. For example, an exceptionally high volume of harmful content was algorithmically recommended on Instagram's short form video product, Reels.

- In our analysis of Reels, 99% of the short form videos we were algorithmically shown, through watching a set of posts recommended by the app's autoplay function, contained at least one type of harmful material, with more than half of posts referencing suicide ideation (often through graphic and slickly produced memes.) We consider it likely that this is the result of a commercial decision to grow the product's user base, at the potential expense of user safety, and in a race for market share.

- The differential exposure to suicide and self-harm was also reported in the internal Instagram survey, with young teens most likely to be exposed to self-harm content on platform surfaces that rely on algorithmic recommender systems. Among teens who had seen self-harm in the previous seven days, almost one-third (31.9%) had seen it on their feed or Instagram Stories, while 25% had seen it on the Explore tab.

Exposure and potential impacts

- Suicide is the third leading cause of death among 15 to 19-year olds,[18] and the most recent annual figures indicate that 524 people aged 24 under died by suicide in the UK.[19]

- Findings from multiple studies have raised concerns about the harmful effects of exposure to self-harm and suicide related online content; the impact of engaging with material that promotes, glorifies or incites serious acts of self-injury; and the behaviour of malign actors who identify and target other users to encourage, incite or otherwise facilitate suicidal and/or self-injury acts.

- There is emerging evidence of the relationship between exposure to harmful online content and resulting suicide and self-harm risks, with recent research concluding that suicide-related online experience is a 'common but likely underestimated antecedent' to suicide in young people.[20] Suicide-related internet use has been reported in 24% of deaths by suicide among young people aged 10 to 19, equivalent to 43 deaths per year. [21]

[18] Department of Health and Social Care (2023) Suicide Prevention in England: Five Year Cross Sector Strategy
[19] Office for National Statistics (2023) Quarterly Suicide Death Registrations in England: 2001 to 2021, and Q1 to Q4 2022 provisional data. Newport, Office for National Statistics
[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] Rodway, C et al (2022) Online harms? Suicide related online experience: a UK-wide case series study of young people who died by suicide. Psychological Medicine, 53(10), pp1-12

- Suicide and self-harm related internet use has been reported in 26% of child hospitalizations relating to self-harm.[22] We agree with Ofcom that it is practically difficult to determine between suicide and self-harm online content that is often highly interconnected and related in its nature; and in any event, self-harm is identified as a major risk factor for suicide in adolescents and young people.

- Self-harm rates among children and young people are also rising. Between 2011/12 and 2021/22, hospital admissions for self-harm content among 10 to 14-year-olds in England more than doubled (a 124% increase).[23] There were 42,793 admissions among young people aged 10-24.[24] In 2014, one in five female 16- to 24-year-olds reported non-suicidal self-harm, a threefold increase since 2000.[25]

- There are an estimated 200,000 hospital presentations for self-harm year in England, although the occurrence of self-harm in the community is likely to be considerably higher.[26]

- There is a clear relationship between suicide-related internet use and rates of suicide in groups with certain protected characteristics. Research shows that suicide related Internet use is recorded more frequently in the death by suicide girls, and in cases affecting adolescents who are identified as LGBTQ+. [27]

- Suicide and self-harm related Internet use results in significant social and economic costs. While further economic modelling is required, the total costs of self-harm hospital admissions to the NHS in England is at least an estimated £213 million per year (2024 prices.)[28] Among people aged 10-19, we estimate that in England alone over 8,100 annual admissions are associated with harmful internet material each year. [29]

<u>Mechanics and drivers of online suicide and self-harm risks</u>

- Findings from multiple studies have raised concerns about the harmful effects of self-harm and suicide related online content. While further research is needed to determine the strength of a causal relationship, and suicide and self-harm content has been found to have

[22] Padmanathan, P (2018) Suicide and Self-Harm Related Internet Use: a Cross-Sectional Study and Clinician Focus Groups. Crisis, 39(6), pp469-478
[23] Nuffield Trust (2023) Hospital admissions as a result of self-harm in children and young people.
[24] Office for Health Improvement and Disparities (2024) Public Health Profiles: Self-Harm.
[25] 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
[26] Department of Health and Social Care (2023) Suicide Prevention in England: Five Year Cross Sector Strategy
[27] ibid
[28] Based on a total cost of £167 million based on hospitalisations in England, calculated by Tsiachristas, A et al (2020 Incidents and general hospital costs of self-harm across England: estimates based on the multicentre study of self-harm
[29] This figure is calculated by using NHS England data for the total number of hospital admissions for self-harm in 2021/22 among people aged 10-19 and applying Padmanathan et al's analysis of how many child hospitalisations display suicide and self-harm internet-related (26% of all admissions)

both harmful and protective effects, a recent systematic review concludes that harmful effects predominate.[30]

- Potentially harmful impact of self-harm and suicide content may include:

    - increases in the frequency and/or severity of self-harm behaviour and suicide ideation. Arendt et al (2019) found that one-third of participants in their study carried out the same or similar types of self-harm after observing it on the site they studied, Instagram;[31]

    - 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;[32]

    - engaging with self-harm content may result in emotional, cognitive and physiological impacts, which may trigger or exacerbate self-harm behaviours and suicidal thoughts;[33]

    - engaging with harmful content may result in the development of a 'self-harm' or 'suicide' identity, in some cases resulting in habituation to seeking harmful stimuli and the cementation of suicide ideation or self-harm behaviours;[34]

    - the risks of a 'contagion' effect, in which behaviours or ideation develop and following exposure to harmful content, including as a result of poor platform design choices and practices that push out suicide and self-harm content to children; [35]

    - an adverse 'assortative relating' effect, in which young people experiencing suicide ideation or thoughts of self-harm are more likely to identify and build relationships with other users experiencing similar actions and thoughts[36]. Although this technology-facilitated effect may provide adolescents with much needed immediate connection, validation, help and support[37], it also presents significant risks (including the potential for unintended consequences.) For example, self-harm may become

[30] 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
[31] 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
[32] ibid
[33] 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
[34] ibid
[35] Seong, E et al (2021) Relationship of Social and Behavioural Characteristics to Suicidality in Community Adolescents with Self-Harm: Considering Contagion and Connection on Social Media. Front Psychol. 12: 691438
[36] 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
[37] 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.

portrayed as unacceptable or normalised coping mechanism, and social support may inadvertently preclude off-line or expert oriented forms of help seeking (establishing a sense that those who do not self-harm 'would not understand'.)[38]

- Studies point to a higher risk of adverse impacts associated with suicide and self-harm content in adolescent girls[39] and those already experiencing poor mental health, including health conditions such as depression, anxiety and poor body image.[40]

- A recent systematic review found that adolescents with clinical level mental health problems may be particularly vulnerable to digitally mediated harm.[41] Young people diagnosed with depression reported more problematic internet use, as well as difficulties in regulating their digital engagement compared to their nonclinical peers.[42]

- Cross-sectional studies have shown higher rates of social anxiety, depression, or suicidal ideation in people who report suicide and self-harm related Internet use compared with those who do not.[43]

- In summer 2023, the US Surgeon General issued a landmark advisory on the growing concerns about the effects of social media on young people's health and well-being. Advisories are usually reserved for urgent and significant public health challenges that require immediate awareness and action. It concluded that; 'at this time, we do not yet have enough evidence to determine if social media is sufficiently safe for children and adolescents to use.'[44]

- The Meta whistleblower Frances Haugen released a series of internal research reports that suggested Instagram was where it contributed to poor mental health and well-being outcomes for a significant minority of its teenage users.[45] For example, she disclosed an internal survey found that 13.5% of UK teenage girls who had experienced suicidal thoughts said that Instagram had exacerbated or worsened their suicide ideation.

- In a study of 1,282 teenage Instagram users, one in five respondents had thought about suicide or self-harm, with strongly observed risks in respect of social comparison, social

---

[38] ibid

[39] Nesi, J et al (2021) online self-injury activities among psychiatrically hospitalised adolescents: prevalence, functions and perceived consequences. Research on Child and Adolescent Psychopathology, 49, pp519-531

[40] For example, Meszaros et al (2020) found problematic Internet use was significantly positively correlated with symptoms relating to self injury affective disorders and anxiety. Meszaros, G et al (2020) Non-suicidal Self Injury: Its associations with pathological Internet use and psychopathology among adolescents. Frontiers in Psychiatry. 11, P814

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

[42] See for example Ucar, H et al (2020) Risky) Cyber Behaviours in Adolescents with Depression: a case-control study. Journal of Affective Disorders, 270, pp51-58

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

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

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

pressure and negative interactions with other users. Teenagers experiencing poor mental health, or that reported being generally unsatisfied with their lives, were much more likely to see mental health related content, and to self-report this made them feel worse.

Evidence of functionality-driven risk factors

Algorithmic recommender systems

- Algorithmic recommendation systems are arguably the greatest single driver of suicide and self-harm content on social networks and can result in users being exposed to large volumes of harmful material. This may in turn increase the vulnerability and susceptibility of users to illegal suicide and self-content and behaviours.

- Platform algorithms continue to push out substantial amounts of content relating to suicide, self-harm and intense feelings of misery and hopelessness. In our recent research, almost half of the most engaged posts on Instagram (48%) and TikTok (49%), and that were posted using well-known suicide and self-harm hashtags, contained material that was likely to be harmful.[46]

- While a significant proportion of harmful posts on Instagram contained material that promoted or glorified suicide and self-harm (in clear violation of the platform's community standards), much of the risk stems from the cumulative impact of viewing large amounts of emotionally disturbing or triggering material.

- Our research suggests that content that may not necessarily pose a risk when viewed in isolation may contribute towards a substantial risk if it is algorithmically recommended in feeds, search results or through autoplay functions, including over a long-term, cumulative basis. This is consistent with much of the material algorithmically recommended to Molly, who viewed 2,000 posts relating to suicide and self-harm in the six months before she died.

- Algorithmically-recommended suicide and self-harm content is capable of achieving extraordinary levels of reach. On TikTok, more than half of the harmful posts we analysed 54%) received over one million views, and almost two-thirds of posts (64%) were viewed more than 250,000 times.

- Harmful content also generates a substantial amount of likes. More than half of harmful posts (51%) were liked by at least 250,000 accounts, with one in eight (12 per cent) liked by at least one million users.

[46] Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest. London: Molly Rose Foundation in partnership with The Bright Initiative by Bright Data

- Our research found that algorithms appeared to fuel clear assortative relating effects, with the algorithmic recommendation of both content and accounts actively enabling users with an interest in suicide, self-harm and depression content to effortlessly identify and build relationships with each other.

- While algorithmically facilitated assortative relating can result in some protective effects, there is an obvious risk that this also enables users with malign intentions to identify and target users who are experiencing mental distress. Users may be identified by malign actors for the purposes of inciting or encouraging suicide or serious self-harm, to signpost them to high-risk third party sites such as suicide fora or closed groups, and for other illegal acts such as sexual grooming.

- The Canadian Centre for Child Protection has found evidence that child sexual abusers use Discord servers and messaging channels focusing on poor mental health, self-harm and suicidality for the purposes of identifying and grooming potentially vulnerable children.[47]

- Instagram's internal data, leaked by the whistleblower Arturo Bejar, shows that the vast majority of exposure to self-harm content appears to be driven by algorithmic effects. Over 90% of self-harm content seen by 13-15 year olds was posted by people that the teens either didn't know at all (71.3%) or only knew through the platform (19.2%). [48]

- Algorithmic design, and the ability of recommended systems to personalise, curate and suggest even more extreme content to vulnerable users, has been shown to be a particular driver of adverse mental health and well-being impacts, and plays an important role in driving the poorly regulated Internet use of some adolescents.[49] Put simply, algorithmic design may not only expose vulnerable adolescents to harmful content but incentivises them to engage with it more intensively and for longer.

- In our research, we found evidence of some young people using the amount of suicide, self-harm and highly depressive content they will be algorithmically recommended as a barometer for their mental health and well-being at a particular time. 'Guess I'm not good in my head again', one teenager remarked on the amount of harmful content he was being recommended.

<u>Comments and discussion spaces</u>

- Functionality that enables users to post comments and start discussions relating to suicide and self-harm can be a major area of risk for potentially vulnerable users, with risk profiles associated with both large social media sites and high-risk suicide discussion fora.

- Suicide, self-harm and highly depressive content generates exceptionally high levels of engagement, meaning that some social media posts effectively serve as a de facto discussion

[47] Shared in discussions with the Canadian Centre for Child Protection in February 2024
[48] The research is presented in appendix one

forum for users who are experiencing suicide ideation, thoughts of self-harm or other types of emotional distress.

- For example, as part of MRF's recent research into the nature and prevalence of suicide and self-harm risks on TikTok, we found that more than one-third of posts had received over 2,500 replies (some of which were effectively morphed into discussion threads.) 20 per cent had received 5,000 comments or more.[50]

- We found limited evidence that comments were being effectively moderated, with multiple examples of comments that encouraged or promoted the user to escalate their self-harm behaviours or consider taking their own life.

- There are also broader unintended consequences associated with large volumes of largely unmoderated comments. For example, research suggests that large volumes of comments may risk normalising self-harm is an acceptable coping strategy, trigger emotional dysregulation effects or may encourage adolescents to understand that suicide ideation and self-harm behaviours are more common than they are in reality.

- There is also the risk that online social support may intentionally or inadvertently preclude seeking clinical expert help. In a substantial number of posts, we found that users expressed a sentiment that only those who experienced suicidal or self-harm ideation 'could truly understand or offer them support.' [51]

- Social media platforms appear to play a key role in signposting users experiencing suicide ideation towards pro suicide discussion fora, where significant disturbing volumes of illegal activity can be readily observed. The National Crime Agency (NCA) estimates that up to 90 deaths in the UK may be linked to Kenneth Law, the Canadian national who allegedly sold 'suicide kits' to individuals. A substantial proportion of these contacts were allegedly made through a well-known pro suicide group already known to the regulator. [52]

- Research into this suicide forum has found that 30% topics found in discussions deal with suicide methods, including questions on methods on how to acquire them.[53]

- X has introduced 'community spaces', dedicated channels on specific topics or interests that are algorithmically recommended in user feeds. These include a number of community spaces related to suicide and self-harm, including spaces dedicated to well-known suicide and self-harm hashtags that Twitter/X had previously claimed they would no longer algorithmically recommend to its users.

---

[50] Stoilova, M et al (2021) Adolescents' health vulnerabilities and the experience and impact of digital technologies: multi-method pilot study. Reported in Kostryke-Allchorne, K (2023) Review: Digital experiences and the impact on the lives of adolescents with pre-existing anxiety, depression, eating non-suicidal self-injury conditions - a systematic review. Child and Adolescent Mental Health, 28(1), pp22-32

[51] 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.

[52] Comments from the National Crime Agency to The Times. Beal, J (2023) 'Suicide Poison' Chef Linked to Eighty-Eight Deaths in the UK. The Times, 25/08/23

[53] Sartori, E (2022) Analysing Sanction Suicide: a case study on pro-choice suicide sites. Padova: Universita degli Studi di Padova

- As recently as February 2024, spaces related to suicide and self-harm hashtags featured an extensive range of problematic content, including posts promoting suicide and self-harm, graphic images that violate the company's policies, and evidence that the groups were being used to encourage or incite others to commit harm.[54] Membership of these spaces, and active engagement with or posting content in them, readily enables users experiencing suicide ideation or self-harm to identify and connect with each other. Furthermore, these spaces enable vulnerable adolescents and young adults to be identified by those wishing to cause them harm.

- Previous research found a significant spike in the growth of content related to such hashtags, with the number of users featuring the hashtag #shtwt in their profile bios doubling, and the number of relevant tweets increasing seven-fold, over a nine-month period between October 2021 and July 2022.[55] MRF analysis suggests these algorithmically recommended channels are recording a substantial amount of daily posts.

<u>Saving and sharing functionality</u>

- We have significant concerns about the ways in which users can save, store, engage or share suicide and self-harm related material on social networks, often through a single click.

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

- While further research is needed to understand this set of consumption patterns, there is a reasonably foreseeable risk this could facilitate 'binge watching' of harmful content, and could result in emotional dysregulation, triggering thoughts, and even the onset thoughts of self-harm and suicide ideation.

- We encourage the regulator to require companies to identify and act on clearly observable suicide and self-harm risk pathways, in which platform algorithms recommend large volumes of harmful material to potentially vulnerable users; their mental health declines as a result of being exposed to large amounts of cumulatively harmful content; and users can amass substantial volumes of albums or collections of harmful content to 'binge watch' on-demand, for example when feeling emotionally vulnerable or triggered.

<u>Search and discoverability features</u>

[54] Analysis undertaken by the Molly Rose Foundation
[55] 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

- A range of search and discoverability features on social networks increase the risk that users could be readily exposed to harmful suicide and self-harm content, including material that may be directly illegal or that could leave users more susceptible to the adverse effects of illegal content and behaviour.

- Our recent research observed a number of high-risk design features on both TikTok and Pinterest, including in video search recommendations that are highly problematic recommended hashtags and search terms ('people also search for 'quickest way to end it.')[56]

- TikTok returns a list of recommended search terms, many of which were highly problematic ('others searched for 'I feel like I'm drowning mentally' and 'I don't think I'll be here much longer.') Both TikTok and Pinterest also generate autocomplete suggestions for search terms, with a search for 'want to….' prompting options including 'want to end it', 'want to give up', and want to go missing.'

- Pinterest uses a range of particularly pervasive user engagement prompts, with algorithmically recommended suggestions of harmful content displayed on the app's home page and its 'updates' feed. Perhaps most perniciously, we observed that Pinterest sent us daily emails recommending a selection of harmful suicide and self-posts that 'we might like.' A substantial proportion of these posts contained material that promoted or glorified suicide or bodily injury, meaning they were in breach of the platform's guidelines.

- We are also aware that user engagement features are directing users towards prohibited suicide and self-harm content on X, including material that may produce normalising or desensitisation effects. For example, in February 2024, users were emailed links to a video of man taking his own life following a domestic dispute, which the platform continued to algorithmically recommend for several weeks despite it being reported by a large number of users.[57]

- Twitter/X also uses recommender algorithms to recommend community spaces, such as the community spaces being used to actively promote, encourage and instruct self-harm acts discussed earlier in this section.[58]

- Hashtags continue to play a substantive role in enabling users to readily access and discover potentially harmful suicide and self-harm content.[59] Following the initial reporting of Molly's story in 2019, Instagram pledged to introduce sensitivity screens and block access to problematic hashtags. However, our research found virtually no sensitivity screens in place

[56] Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest. London: Molly Rose Foundation in partnership with The Bright Initiative by Bright Data

[57] The Molly Rose Foundation was made aware of this video and repeated unsuccessful reports by users who were concerned they received these email recommendations and by X's failure to remove such violative content. Further information can be supplied on request,

[58] 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

[59] Picardo, P et al (2020) Suicide and self-harm that on Instagram: a systematic literature review. PLos One, 15(9)

(in less than 1 per cent of harmful posts) , and the inconsistent and haphazard application of measures that could introduce friction into the search experience.

Biographical features

- Our recent research found multiple ways in which accounts distributing harmful suicide and self-harm content able to exploit platform design features around biographical features and anonymity.

- Multiple research projects have demonstrated the potential protective effects associated with anonymity. This includes teens and adults who are experiencing mental health problems, suicide ideation and thoughts of self-harm to express their feelings, vent, and receive peer on peer support.[60] However, we also found evidence of multiple accounts that used their anonymous status to identify as teens experiencing suicide and self-harm, when it appeared their primary motivation was to spread harmful and potentially illegal content.

- Many of these accounts demonstrate signals that raise questions about their authenticity, for example the use of similar or identical terms and phrases in their bios that suggested some relatively sophisticated understanding of strategies to game content moderation practices.

- We found numerous examples of high engagement accounts that were able to fraudulently identify themselves in their Instagram bios using description such as 'mental health resources', 'public figures' and 'crisis prevention centres'. High engagement accounts typically post a large volume of memes, videos and text-based posts to quickly gain followers and maximise user engagement, and our research shows were responsible for a significant amount of the most-engaged with harmful suicide and self-harm content.

- The use of such labels clearly imbues a false sense of legitimacy and demonstrates how platform design choices can be readily gamed by users. Even in instances where the accounts appeared genuinely committed to offering peer support, there are obvious risks if high engagement accounts can overstate or misrepresent their status to potentially vulnerable followers.

- These risks appear particularly significant in the context of threat actors who may be looking to identify vulnerable users for the purpose of committing criminal offences, including the incitement of acts of suicide and serious self-injury, sexual grooming, and other forms of coercive behaviour.

Ephemeral stories and broadcast channels

[60] 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

- Multiple high engagement accounts have made full use of recent design choices, including Instagram's Stories and broadcast channel features, to rapidly build their follow base and engagement levels.

- The Stories feature demonstrates a noticeably higher risk profile than most other platform surfaces, with the internal survey leaked by Arturo Bejar suggesting that teens were more likely to be exposed to self-harm content on their feeds or in Stories than on any other part of the platform.

- High engagement accounts have been early adopters of broadcast channels, a new design feature in which followers can subscribe to receive posts and messages that appear alongside their DMs.

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

- These features enable teens and young adults who may be experiencing intense depression, suicide ideation or thoughts of self-harm to be readily identified by other users, and in turn to be potentially targeted and contacted by threat actors looking to target them for the purposes of illegal acts, for example the incitement of acts of suicide and self-harm and/or sexual grooming.

- There is also the potential for significant unintended consequences, including the risks associated with descriptive normalisation (the perception the behaviour is more common than it actually is)[61] and social learning effects (where there is a risk that mood behaviours may be modelled or imitated based on exposure to the shared characteristics of the group.)[62]

- It is reasonably foreseeable that these unintended consequences may make some users more vulnerable to the risks associated with harmful suicide and self-harm content, and subsequently more vulnerable and susceptible to illegal material or behaviours.

<u>DMs and private messaging</u>

- DMs appear to play a significant role in the way that users engage with suicide and self-harm material. For example, Instagram's internal research shows that 14.9% of teens who had seen suicide or self-harm content in the last seven days had received this through a private message.[63]

[61] ibid
[62] 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
[63] See appendix one.

- Many account bios actively encourage DMs as a means for user-to-user communication, with many high engagement accounts adopting similar and/or identical bios that encourage users experiencing emotional distress to message them. The potential for this tactic to be exploited by threat actors, including those wishing to target vulnerable users for illegal acts, is clear.

- Although further research is needed into the impacts and mechanics of suicide and self-harm related risks in private messaging, the reasonable assumption is that the majority of illegal content and behaviour is likely to take place in private messages, closed messaging groups and other private spaces.

- While there is presently limited evidence about the ways in which end-to-end encryption is used to enable relevant offences, we have significant concerns that the forthcoming rollout of encryption on Meta's platforms is likely to result in a worrying increase in the risk profile associated with relevant suicide and self-harm offences.

- The Canadian Centre for Child Protection (C3P) has identified that serious acts of self-harm takes place on private messaging services as a result of sexual grooming and coercive control behaviours, with young people being coerced into acts of self-harm to abuse, degrade and control them. [64]

- C3P states that pathways can start on social media platforms or on Discord channels and groups where users discuss themes such as poor mental health, low self-worth and loneliness.

- Police in British Columbia have recently warned about violent online groups that deliberately target vulnerable minors aged 8-17 and pressure them into recording online streaming self-harm and producing child sexual abuse material[65]. LGBTQ+ youth, ethnic minorities and adolescents with problems are disproportionately targeted, with clear interlinkages apparent between illegal suicide and self-harm offences and child sexual abuse. This disturbing new trend carries clear parallels to more established child sexual abuse threat vectors, including livestreamed monetised abuse.

- The FBI has also issued an advisory about the growth of self-injury grooming groups, stating that the intention of such groups is to actively force minors 'to kill themselves on online streams for their own entertainment or for their own sense of fame.' [66]

- These groups use extortion and blackmail tactics, such as threatening to swat or dox users or share self-generated images of them, unless they agree to livestream self-harm activities including cutting, stabbing or 'fansigning' i.e. writing or cutting specific numbers, letters, symbols or names onto your body. The FBI states that social media, dating apps and other online sites have been used by these groups.[67]

[64] Discussions with the Canadian Centre for Child Protection held in February 2024
[65] Roumelotis, I et al (2024) Violent online groups are pressuring you into harming themselves, authorities warn. CBC News, 09/02/24
[66] FBI (2023) Public Service Announcement: the violence online groups extort miners to self-harm and produce child sexual abuse material. Posted 12/09/23
[67] ibid

- We would also remind the regulator that the lack of evidence associated with the risks associated with suicide and self-harm content in private messaging is strongly associated with evidentiary challenges in retrieving such content.

- Through our work with other bereaved families, including the Bereaved Parents for Online Safety group, we are aware of other parents that have reasonable grounds to believe that children may have been exposed to suicide and self-harm related material in private messages. However, none of these families have been able to retrieve relevant data from companies, even where warrants were issued.

- In Molly's case, while Instagram and Pinterest eventually provided data on what Molly had seen, recommended or received through direct shares, neither platform provided text of her private messages to our legal team. We know that Molly had blocked a number of users in the months before her death, but without access to the relevant messages are unlikely to ever the circumstances behind this.

Risk factors driven by business models and commercial profiles

- We are surprised that Ofcom was unable to find evidence that supports the relationship between the business models of regulated companies and the risk of being exposed to illegal suicide and self-harm content on their services.

- There is increasing evidence that commercial and revenue drivers have actively informed the approach of social media companies to child safety and wellbeing issues, including suicide and self-harm content.

- As Ofcom itself recognises, algorithmically recommended systems are a major driver of exposure to illegal suicide and self-harm material and underpin the business models of most of the major social networks in scope.

- Our research into the nature and prevalence of suicide and self-harm material on Instagram found significantly greater volumes of harmful material on Reels than on any other part of the platform. 99% of the videos we were algorithmically recommended were identified as harmful, with much of the content being suicide and self-harm related memes. [68]

- In our assessment, the significantly increased risk of exposure to harmful content on Reels can only be explained by Instagram's tolerances being lower on than on any other part of platform. The Reels surface has been identified as a major growth area by Instagram's parent company, Meta, with the company emphasising increased time spent on Reels as a key metric in its corporate and earnings reports.[69]

[68] Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, Tiktok and Pinterest. London: Molly Rose Foundation in partnership with The Bright Initiative by Bright Data

[69] For example, in Meta's Q1 2023 quarterly earnings call, Mark Zuckerberg proactively shared that time spent on Instagram has grown more than 24% since the company launched reels thanks to its AI powered content

- Recent disclosures made by the US Senate Judiciary Committee have highlighted that over several years Meta has consistently opted not to invest in critical aspects of child safety and well-being, including suicide and self-harm content, with two separate business cases to create youth safety teams rejected by senior executives in both 2019 and 2021.[70]

- In April 2019, the Head of Instagram Adam Mosseri rejected the first business case on the grounds there were pressures on headcount across the business. At the time, Instagram had zero Product resource and only 0.2 FTE of a research post working on child well-being issues, including suicide and self-harm content.[71]

- In the business case, Mosseri was warned that Instagram was failing to optimise its product for youth; the platform was missing borderline and adjacent suicide and self-harm content; and that so-called SSI content was not customized consistently for youth across the company.

- Weeks earlier, senior executives including Adam Mosseri and Mark Zuckerberg had been separately warned there was a 'palpable risk' of a repeat of further child deaths, in the immediate period after Molly's story was first shared.[72] Internal company memos attributed this to a lack of product investment and Instagram's recommender systems continuing to push out harmful suicide and self-harm content to its users. [73]

- In 2021, the second business case was submitted to Mark Zuckerberg by Nick Clegg, who leant heavily on the risks to Meta's investment in the metaverse to try and secure his approval for a new child well-being team.

- Clegg told Zuckerberg that Meta 'was not on track to succeed for our core well-being topics', which included problematic use and suicide and self-injury. He added that 'if not addressed, [increased regulatory risk and external criticism] will follow us into the metaverse.'[74]

- Prior to Mark Zuckerberg declining both business cases, Nick Clegg has been warned this was the likely result by Chief Product Officer Chris Cox, who highlighted the cost implications when he wrote that 'there is a very low likelihood that Mark will approve given how overconstrained we are.'[75]

- There is further evidence that Meta has not been willing to proceed with child well-being improvements where these affect the company's bottom line, however modestly. Recent legal disclosures have demonstrated that Instagram declined to rollout Project Daisy, a pilot

---

recommendations. Chief Financial Officer Susan Li commented 'we are very pleased with what we've seen Reels drive in terms of incremental engagement on the platform so far.'

[70] Internal Meta emails were released by the Senate Judiciary Committee following its hearing with social media and gaming CEOs in January 2024. Copies available on request.

[71] ibid

[72] Unsealed filings from the 41 State Attorneys Generals case against Meta, November 2023

[73] ibid

[74] Sourced from the internal Meta emails released by the Senate Judiciary Committee

[75] Ibid

that tested the removal of like counts from teen feeds, because the changes led to a small decline in user engagement and a 1% fall in ad revenue.[76]

- Project Daisy had demonstrated substantial improvements in the emotional well-being of teenage users. The pilot responded to the pronounced risks of negative social comparison highlighted by Instagram's internal research (leaked by Frances Haugen),[77] and extensive academic research warning that negative social comparison was a particular risk factor for poor mental health among adolescent girls.[78]

[76] Unsealed documents from the New Mexico vs Meta lawsuit being taught by the state's Attorney General, January 2024

[77] comments are available on the Wall Street Journal website, as part of their Facebook Files investigation

[78] For example, see de Vries, D et al (2015) Facebook and self perception: individual susceptibility to negative social comparison on Facebook. Personality and Individual Differences, 86, pp217-221

Section 3: illegal content Codes of Practice

- This section of our response focuses on the measures recommended by Ofcom to mitigate the risk of illegal harms, captured in its illegal content Codes of Practice.

- Ofcom has set out that it believes these first Codes 'represent a strong basis on which to build a more comprehensive suite of recommended measures to reduce the risk of harm to users over the long term'. The regulator also explicitly states that its aim is to 'capture best practice within industry and set clear expectations on raising standards of user protection, especially for services whose existing systems are patchy or inadequate.'

- In our assessment, the Codes fundamentally fail to deliver the systemic, risk-based regime, focused on outcomes rather than prescriptive measures, that Parliament intended when passing the Act. The Codes inadequately respond to the harm dynamics set out in Ofcom's own risk profiles and register of risks, and in their current iteration, do not appear to meet the online safety objectives specified in schedule 4.

- We remind the regulator that section 4(b) of the online safety objectives requires online platforms to design and operate their services in a way that protects users from harm, with particular regard to the algorithms, functionalities and other design features of their service. Ofcom is required to ensure that the measures set out in its codes of practice are compatible with and can actively contribute to ensuring the online safety objectives are being met.

- While we recognise Ofcom's intention to adopt an iterative approach, it is difficult to reasonably conclude that its draft Codes enable the online safety objectives to be met. For example, Ofcom proposes only one downstream measure relating to platform algorithms, despite this being the primary driver of most relevant suicide and self-harm online material.

- As a result, it is difficult to envisage how the draft Codes will deliver any meaningful reduction in the magnitude of and exposure to illegal suicide and self-harm content. In this respect, neither it is evident that the measures will provide for a higher standard of protection for children than adults.

- Given the yawning disconnect between the drivers of harm set out in volume 2, and the measures being recommended to address them in volume 4, we are deeply concerned that Ofcom's approach will prove to be fundamentally ineffective in respect of relevant suicide and self-harm offences; and that many potentially vulnerable online users may continue to be exposed to substantial but largely or wholly unmitigated risks.

<u>Implementation choices</u>

- Ofcom has made a number of strategic decisions about how it develops its Codes of Practice that are deeply problematic and risk significantly constraining their effectiveness.

Safety-by-design

- As set out above, there is a palpable disconnect between the evidence of harm presented in the risk profiles and register of risks and the mitigation for those harms proposed in the codes of practice. While Ofcom correctly identifies the significant role of systemic design choices and functionalities, and indeed is required to do so to comply with its requirements set out in schedule 4(3), its codes of practice focus predominantly on ex-post measures, such as content moderation and takedown, rather than effective and proactive 'safety-by-design' measures that could effectively mitigate the risk that harm is allowed to perpetuate or be amplified in the first place.

- Ofcom does not appear to have considered whether functionalities that are demonstrably harmful, but where it deems insufficient mitigations currently exist, should simply not be allowed to operate until and unless suitable and sufficient risk mitigations are in place. [79]

- Ofcom's weak 'safety-by-design' approach, and its lack of emphasis on upstream preventative approaches, inherently results in a regime that focuses on a prescriptive, tick box set of recommended measures. This approach is poorly suited to delivering harm reduction outcomes and even to prevent some harms from continuing to escalate.

- This approach also fails to incentivise companies to develop suitably innovative and bespoke approaches to the diverse but often highly specific ways in which functionalities and design choices may contribute towards risk profiles across the broad range of services in scope.

Small but high-risk platforms

- Ofcom proposes a differentiated set of requirements for large services (used by at least 7 million monthly users) and small companies (everything else.) This approach seems poorly targeted towards the risks associated with suicide and self-harm content and may continue to expose users to unacceptably high levels of preventable harm.

- Many of Ofcom's recommended measures – including board or governance oversight of risk management – only apply to 'large' companies, and the threshold has been placed so high that many medium-sized but potentially medium or high-risk services may fall out of scope, for example Discord, Telegram, Twitch and Roblox.

- Despite requests, Ofcom has not provided any snapshot data to indicatively suggest which platforms may become designated as large platforms. This has impeded our ability to assess the likely impact of the regime in a fully informed and meaningful way.

- We are particularly concerned by the dearth of measures relating to small but very high-risk sites, including suicide fora that have been associated with the facilitation, incitement and encouragement of acts of suicide and serious bodily harm. Dozens of UK deaths by suicide

[79] Online Safety Act Network (2024) OSA Network statement on illegal harms consultation

have alleged links to pro-suicide sites.[80] As it stands, the regulator has not recommended any downstream measures for small platforms specifically in relation to suicide and self-harm.

- Ofcom's approach seems overly geared towards the economic costs of mitigating harms, rather than the costs of their impacts. Volume 4 specifically states that small companies are exempt from following many of the measures to avoid incurring costs or stifling innovation.

- Similarly, in its recent blog on the size and risk of platforms and how this informs its approach, Ofcom states that 'where we do not yet know whether it is proportionate to extend a measure to smaller services, we have not done so.''[81]

- We wish to endorse the OSA Network's position that it would be helpful to understand the legal basis upon which Ofcom has determined it is acceptable to use size as a factor in determining safety standards. This is particularly pertinent given the changes made late in the Parliamentary passage of the Bill to allow category 1 designation to apply to companies on the basis of either size or risk, which in turn brough small, high-harm services into scope of additional duties).[82]

<u>Conjoined and interrelated harm profiles</u>

- Ofcom's approach fails to give due regard to a range of factors which may influence the risk profile, including how factors may combine or conjoin together to exacerbate risks.

- We have significant concerns that the regulator has inadequately reflected the cross-platform nature of harm in both its risk profiles and recommended measures. The Government has previously set out that the Act requires platforms to take steps to address the cross-platform nature of harms when meeting its illegal safety duty.[83]

- However, as it stands Ofcom inadequately reflects cross-platform harms in its risk profiles and contains no relevant requirements in its Codes. It remains unclear how the regulator intends to enforce this aspect of the illegal content safety duty when there are no relevant measures in its Codes, but companies benefit from a safe harbour as long as they comply with its measures.

- We are also concerned that the regulator has largely treated each of the priority offences as largely siloed harm archetypes, when in practice many of them are extensively interrelated and conjoined. For example, there are often extensive linkages between child sexual abuse,

[80] Comments from the National Crime Agency
[81] Ofcom (2024) Why size and risk matter in our approach to online safety. Ofcom blog posted 30/01/24
[82] Online Safety Act Network (2024) OSA Network statement on illegal harms consultation
[83] Department for Science, Innovation and Technology (2023) Overview of expected impact of changes to the Online Safety Bill

suicide and self-harm offences, with groups inciting and blackmailing children into livestreamed or recorded acts of sexual abuse or serious self-harm.[84]

- We are concerned that platforms may not be sufficiently incentivised by the regime to identify and act on relevant linkages between harm archetypes, and that there may be clear disincentives for companies to identify complex interlinkages between illegal acts when meeting their relevant requirements, including in respect of new and emerging threat vectors.

- While we appreciate the relative complexity of these matters, we consider it vital that the regulator appropriately captures this complexity if it is to appropriately tackle and respond to the nature of many of the most egregious harms in its scope.

- We wish to reiterate our concern that the regulator must not conflate the absence of evidence with the evidence of risk. If the regulator requires further evidence, it should actively engage with law enforcement, child protection agencies and/or commission further rapid research to ensure it has a sufficient understanding of the relevant risk dynamics, and so it can in turn can develop a set of proposals that are commensurate to and appropriately correspond with the level and nature of risks.

Recommended measures and key omissions

Algorithms and recommender systems

- As volume 2 sets out, and our response builds upon, recommender systems are one of the most substantive drivers of exposure to illegal suicide and self-harm material. It is therefore surprising that the regulator has proposed only one measure that is designed to directly reduce the potential risks posed by algorithmic recommendation and curation.

- Specifically, the regulator proposes that when some regulated services undertake on-platform tests on their recommender systems, these should be expected to collect safety metrics that will allow them to assess whether the changes are likely to increase the exposure of users to illegal content. Ofcom only proposes to apply this measure to regulated services that already perform on-platform tests, and to services that have a medium or high-risk in respect of at least two illegal harms.

- Ofcom's proposed approach reflects the regulator's decision to emphasise economic proportionality over the tackling of harms experienced by users. Other regulatory regimes, such as the ICO's Children's Code, correctly recognise that if a platform wishes to use recommender systems to promote content to its users, the platform should only be able to do so once it demonstrates it can do so safely.

[84] Federal Bureau of Investigation (2023) Service Announcement: Violence Online Groups Extort Miners to Self-Harm and Produce Child Sexual Abuse Material. Issued 12/09/23

- Ofcom's approach sets out that is only proportionate for a platform to perform user testing where it is proportionate to do so. In this case, the regulator has applied measures where it has assessed that 'for services in scope of the measure these costs are likely to be a relatively small addition to their existing costs.'

- We strongly encourage Ofcom to adopt a substantially more ambitious approach that recognises that recommender systems should be designed and operated in a way that ensures the reasonably foreseeable safety and well-being of users, particularly children, young adults, and other groups who may be disproportionately impacted by, or vulnerable to, exposure to illegal harms.

- Given the extent to which algorithmic recommendation, personalisation and content curation drives exposure to illegal content, this approach should be seen as a necessary precondition for recommender systems to be used.

- Platforms should be expected to demonstrate they have considered how their algorithms may actively result in the exposure of illegal content and/or may make users more vulnerable or susceptible to the effects of illegal content and behaviours. Furthermore, if a large service performs on-platform tests that demonstrate a substantially lower risk of exposure to harm than that subsequently recorded by the platform's users, for example through mandated user or external surveys, the regulator should be prepared to actively investigate and take steps to ensure the suitability and sufficiency of the platform's tests and risk assessment protocols.

- As MRF's research has shown, recommender systems may frequently result in exposure to harm by contributing to well-established harm pathways, for example through the ways in which TikTok, Instagram and Pinterest recommend large volumes of harmful material, enable users to save it using a 'one-click' option, and then enable users to have access to large libraries of harmful and potentially illegal content, often for the purposes of 'binge watching.' [85]

- The regulator should therefore require companies to consider and act on the risks posed by recommender systems in the broadest possible sense, rather than in isolation.

- In its risk profiles and recommended measures, the regulator should emphasise the importance of disrupting harm pathways where a range of design features may combine to exacerbate the risk of exposure to illegal content, including suicide and self-harm content and behaviours.

- In adopting this approach, and in reflection of the centrality of recommender systems to harm pathways contained wholly on social networks or that may extend across third party platforms or multiple sites, it is both appropriate and proportionate for the regulator to recommend a more stringent and comprehensive set of relevant measures.

[85] Molly Rose Foundation (2023) Preventable yet pervasive: the prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, Tiktok and Pinterest. London: Molly Rose Foundation in partnership with The Bright Initiative by Bright Data

Recommendation of accounts based on common interests

- While we welcome the regulator's recommended measures around default settings and support for child users, we recommend that these measures should apply across all platforms that present a medium or high-risk of exposure to priority illegal content (rather than simply services that present a medium or high risk of grooming.)

- This recommendation reflects the clear similarities in the risk profile between sexual grooming and attempts by threat actors to identify and target users experiencing poor mental health, suicide ideation and thoughts of harm, specifically for the purposes of inciting suicide or serious self-harm behaviours.

- In both cases, algorithmic recommendation of other users readily enables threat actors to identify and make contact with large number of potentially vulnerable users based around shared characteristics or common interests (such as a displayed interest in suicide and self-harm material.)

- Given the risks associated with suicide and self-harm material for young adults, we also recommend that the regulator prevents platforms from algorithmically recommending other accounts based on shared interests, such as suicide and self-harm material, where these can reasonably be considered to increase the risk profile associated with illegal content.

- This proportionate, safety-by-design approach reflects the emerging evidence base that finds that while the formation of online suicide and self-harm communities may result in both protective and harmful effects, harmful effects predominate.

- Recommendation systems actively drive assortative relating effects, which may result in users assessing or believing that suicide or self-harm behaviours are more common than they actually are; that self-harming behaviours are a normalised or optimal coping response; and that could leave users more exposed to the risks of being exposed to illegal content and its effects.[86]

- These measures inherently aim to add friction to the search and discoverability mechanics associated with recommender systems, but do not prevent other users from being able to identify or engage with relevant content through other means.

- As such, these measures would have minimal adverse free expression impacts; and by reducing the potential exposure to illegal and potentially distressing material, may actually result in positive impacts associated with increased freedom of association.

[86] 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

Content moderation

- We support the recommended measure that platforms must have systems or processes in place that are designed to swiftly take down illegal content of which it is aware. However, we note that the Codes of Practice don't require suitable and sufficient proactive processes to detect certain forms of priority content in the first place, including suicide and self-harm material.

- We are furthermore concerned that the guidance emphasises platforms making judgements around individual items of content – rather than, as the Act states – to operate a proportionate system designed that should have that effect. Much of the success of the regime will be determined by the efficacy of platform systems to identify and correctly remove material that can reasonably be considered illegal.

- Ofcom's guidance perhaps inevitably focuses on the risks to free expression associated with excessive takedowns. However, we are concerned that the regulator's guidance, and specifically the adoption of a burden of proof that is closer to a criminal than a civil or regulatory regime, presents an equal or greater risk that platforms will used to interpret the guidance in a way that causes it to adopt a very high bar before it considers content illegal, and removes it accordingly.

- We remind the regulator that the Canadian Centre for Child Protection found that some user-to-user services were refusing to comply with child sexual abuse takedown notices because of how they operationalised their content moderation policies. This resulted in some online services refusing to move and reports images of children estimated to be only 10 years old.[87]

Content moderation policies and possesses that reflect and respond to harm dynamics

- While we welcome the requirement for platforms to prepare internal content policies setting out how their policies should be operationalised and enforced, we note that the relevant measure focuses on the identification and removal of illegal content, rather than adopting a broader 'safety-by-design' led approach that reflects the harm dynamics set out in Ofcom's risk profiles.

- As the risk profiles set out, a range of design features including algorithms, platform engagement features and other design features can, in isolation or combination, increase the potential risk of exposure to priority illegal content, including through recommending content that may not in and of itself be illegal.

- If the regime is to adopt an upstream, targeted harm reduction approach, the regulator should therefore require companies to set out how they operationalise their policies to restrict the spread of such content, including how and under what circumstances this material may be de-ranked, down-weighted or deemed unsuitable for algorithmic recommendation for some or all of its users.

[87] Canadian Centre for Child Protection (2019) How We Are Failing Children: Changing the Paradigm.

- We welcome Ofcom's proposal that regulated services should prepare and apply a policy in respect of the prioritisation of content for review. However, we also recommend that the guidance instructs companies to have regard not only to the virality of content (as expressed by the number of views), but also the amount of times that content is being saved, shared or commented on.

- In respect of suicide and self-harm material, posts being saved at shared at high frequency can significantly extend the risk profile associated with illegal material, as well as being an effective indicator of potential harm.

Risk scoring

- We are disappointed that Ofcom has opted not to include cumulative risk scoring systems as a recommended measure to tackle priority offences, including to prevent exposure to and harms resulting from exposure to illegal suicide and self-harm content.

- The Meta whistleblower Arturo Bejar has disclosed to MRF that Meta developed and successfully deployed risk scoring technology to identify and provide support to users who were identified as an immediate suicide risk. Internal assessments found that the tool was highly successful and had saved hundreds of lives, although when he returned to the company in 2019, Meta had discontinued its use.[88]

- Given Meta was able to successfully develop and deploy risk scoring systems as early as 2016, it is manifestly proportionate for the regulator to recommend that large companies introduce risk scoring solutions.

- As a minimum, we consider that risk scoring should be used to inform and support platform content moderation practices; and to identify and provide support to users who may be particularly susceptible to suicide or self-harm behaviours.

- For example, risk scoring could be used to identify users who are at particular risk because they have been exposed to illegal content or to cumulative amounts of relevant harmful material.

- While we acknowledge there is a significant complexity involved in these systems, and there could be potential adverse impacts on user privacy or freedom of expression, these risks should be seen in the context of the targeted use cases being proposed.

- We also encourage Ofcom to assess the substantial economic and social costs associated with deaths by suicide where this is an online element and re-assess the relative risks and benefits accordingly.

[88] Discussions between the Molly Rose Foundation and Arturo Bejar

Keyword detection

- As the regulator sets out, keyword detection already plays a vital role in text-based content moderation, including the detection, monitoring and filtering of violative and illegal content. There are two main types of standard keyword detection: direct matching, which requires words to exactly match those on the keyword list; and fuzzy matching which allows for words to be identified where there is a partial match.

- It is disappointing that Ofcom has opted not to recommend the use of keyword measures, outside of fraud offences, citing limited evidence about the accuracy of such technologies.

- In citing concerns about the potential for keyword detection to generate a high volume of false positives, Ofcom appears to be attaching greater weight to the potential impacts on free expression than to the merits of recommending measures that can meaningfully reduce exposure to harm.

- We note that several platforms highlight the importance of keyword detection as part of their trust and safety approaches to suicide and self-harm content. For example, Pinterest maintains what it describes as a 'voluminous Sensitive Terms List', with over 50,000 terms on the list. Pinterest users keyword detection to identify violative content, block search terms and prevent relevant auto complete search results from being generated.[89]

- We note that Pinterest identifies as a midsize platform, but that it has still been able to develop and invest in extensive keyword detection mechanisms. On this basis it would be difficult for the regulator to determine that keyword detection mechanisms are not a proportionate measure to suggest.

- Furthermore, we invite the regulator to comment on why seemingly attaches greater weight to the accuracy of this measure in respect of free expression, rather than its efficacy in detecting harm.

Live streaming

- We are surprised that the regulator has opted not to recommend measures in relation to the risks posed by live streaming, despite the fact this is highlighted as one of the primary high-risk functionalities in volume 3.

- Live streaming poses particular risk in respect of suicide and self-harm: this includes organised criminal groups coercing and extorting children to commit acts of self-harm on live streams; live streamed deaths by suicide; and videos of live-streamed deaths being posted and shared on third party social media sites.

[89] Letter from Pinterest to Senior Coroner Andrew Walker responding to the Prevention of Future Deaths report issued following the inquest into Molly's death

- Following a live streamed death by suicide on Facebook Live in August 2020, the video was extensively posted online across multiple social media sites, including Facebook, YouTube, TikTok and Instagram. This was a particularly pronounced problem on TikTok, where the video was posted repeatedly by users who deployed adversarial posting tactics to evade content moderation.

- TikTok's algorithms actively recommended the suicide video to other users, including children. One mother of a 14 year old girl told the BBC the daughter had slept with the light on, felt scared to leave the house and had missed the day of school as a result of being accidentally exposed to the video.[90]

- The problem posed by such content, including the potential for it to be spread with considerable velocity and virality by malign actors on third-party sites, could result in particularly problematic effects for those already experiencing suicide ideation, emotional distress and other forms of poor mental health. It may also reasonably make these groups more susceptible to the adverse impacts of being exposed to further harmful and/or illegal content.

- Following this incident, TikTok wrote to other social media platforms recommending that industry collaborate on the development of rapid response mechanisms to identify and remove videos of live streamed suicide content, including hash-matching technologies and other similar techniques deployed following the Christchurch terrorist attack. No further progress was made in this regard, and it is therefore unclear what if any substantive steps have been taken to minimise the risk of further repeats.

Search platforms

- Search platforms can be a substantial means through which users can be exposed to illegal suicide and self-harm content, including children.

- For example, major search engines continue to readily display links to a highly problematic pro-suicide forum. The first page of Google results also contains links to Reddit and Wikipedia pages relating to the site, as well as a platform that recommends 'top alternatives and competitors', including platforms referencing a 'loss of hope' and messaging boards used by incels.[91]

- We note the regulator's recent research that found substantial amounts of harmful suicide and self-harm material were available through major search engines, with more than one in five (22%) of results linking, in a single click, to content which celebrates, glorifies or offers instruction about non-suicidal self-injury, suicide eating disorders.[92]

[90] Wakefield, J (2020) TikTok tries to move widely shared suicide clip. BBC News, 08/09/24
[91] Analysis undertaken by the Molly Rose Foundation in February 2024
[92] Jussim, L et al (2024) One Click Away: a study on the prevalence of nonsuicidal self injury, suicide and eating disorder content accessible by search engines. Rutgers: Network Contagion Research Institute (commissioned by Ofcom)

- While we will not repeat our detailed discussion about the ways in which platform design choices and functionalities can increase the risk of exposure to harmful content, we encourage the regulator to cross-reference our concerns about social media to search engines where appropriate.

- We are particularly concerned that search engine algorithms not only recommend websites containing harmful content, but also prioritise them in search results. Ofcom's research found that users were more likely to discover links to harmful content in the top-five page 1 results than in search results overall.

- We also note that users were six times more likely to find harmful content about self-injury when entering deliberately obscure search terms, a common practice among online communities.

- However, we disagree with the regulator that the specific and evolving nature of these terms (so-called 'algospeak') pose significant detection challenges for services. It is entirely reasonable to expect that platforms should have appropriate ongoing detection and monitoring processes to track emerging changes in user behaviour and search terms, and to take relevant corresponding measures accordingly. [93]

- We also encourage the regulator to explicitly consider so-called 'data voids' as an explicit risk factor in its illegal content scheme, and to require platforms to take additional measures in response to the resulting risks.

- 'Data voids' refer to situations where the search demand for certain keywords is not met with reliable safe information, due to the search engine's algorithms not being adequately updated. Searches using cryptic language may therefore lead to more harmful content, as algorithms aim to provide relevant results, but lack safe and accurate information to fill these gaps.

Governance measures

Senior manager accountability

- We strongly support Ofcom's proposal to recommend a set of measures in respect of ensuring senior manager visibility of and accountability for risks, and related requirements to require platforms to establish clear lines of accountability for compliance with the Codes of Practice.

[93] In fact, the investment in methods to track these linguistic changes among some sites are the primary driver for the development of 'algospeak' in the first place. See for example Steen, E et al (2023) You can (not) say what you want: using algo speak to contest and evade algorithmic content moderation on TikTok. Social Media and Society, 9(3)

- MRF has argued consistently that senior manager liability and a shift in the organisational culture of large firms must be viewed as a prerequisite for securing good regulatory outcomes. To ensure the success of its regulatory scheme, Ofcom should therefore be actively targeting a culture of accountability, responsibility and safety-by-design across all layers of the companies it regulates.

- We strongly welcome Ofcom's proposal that all companies should have a named person who is responsible for how the relevant platform complies with its regulatory requirements, and that the person should be accountable to the most senior applicable governance forum, ordinarily the company's Board.

- We also welcome Ofcom's recommendation that all senior members of staff should have written statements of responsibilities, and that these will broadly mirror the arrangements in the financial services sector. This measure is important to ensure that all key responsibilities for online safety decision-making are appropriately assigned, and to ensure there is clarity and ownership around all aspects of responsibility and risks.

- We welcome Ofcom's recognition that senior manager accountability is a cornerstone of other regulatory regimes, including the Senior Managers and Certification Regime in the financial services sector. Findings from a 2020 review by the Prudential Regulation Authority reported positive behavioural change and improvement in risk management practices among companies that are subject to the requirements.[94]

- The testimony of whistleblowers such as Frances Haugen underscores the importance of clear risk ownership and assigned responsibilities for product safety. In discussions with civil society, Haugen set out how ambiguous reporting lines and a lack of clarity about who owned trust and safety risks often led to poor safety outcomes and disincentives for safety to be proactively designed into Meta's services.[95]

- Recent Senate disclosures underscore the current lack of clear safety responsibilities in companies, and in relation to Meta specifically, significant internal pushback against proposals to create a dedicated youth and safety wellbeing resource. Head of Instagram Adam Mosseri and Chief Product Officer Chris Cox were strongly advised by an unnamed senior member of staff not to introduce a horizontal lead for youth safety because 'there is [already] a lot of layers being built up on teams doing the work, plus having too much central oversight demotivates local product and research teams.' [96]

- While we welcome Ofcom's proposed measures, we are disappointed that the regulator has opted not to proceed with broader and deeper accountability measures at this stage. Experience from the financial services regime underscores the importance of Senior Manager Liability at all relevant levels of the business, with clear personal incentives to

[94] Bank Of England Prudential Regulation Authority (2020) Evaluation of the Senior Managers and Certification Regime

[95] NSPCC hosted a roundtable session with Frances Haugen for civil society groups during her 2021 visit to the UK.

[96] Internal company emails were unsealed by the Senate Judiciary Committee on 31/01/24 following their committee hearing 'Big Tech and the online child sexual exploitation crisis.' MRF can provide copies to Ofcom.

comply with regulatory requirements and to take steps to address or report areas of concern and non-compliance.

- We are also unclear how the regulator intends to enforce these measures, and how it envisages them having suitable and sufficient bite to drive meaningful changes in culture, accountability and risk ownership. In particular, we note that the regulator has not set out proposals to hold named persons personally responsible for failures to comply with their relevant regulatory requirements.

- In practice, it seems highly unlikely that these measures will sufficiently incentivise either named persons or corporate entities to do any more than pay lip service to the compliance responsibilities being expected of them.

- Corporate incentives to emphasise revenue maximisation will likely continue to outweigh the personal or corporate repercussions of failing to adhere to regulatory requirements, and it seems highly likely that named persons will therefore view these responsibilities as little more than a 'box ticking' exercise.

- The recent Senate disclosures underscore the insufficient weight that Meta attaches to safety and wellbeing outcomes, with Adam Mosseri and Chris Cox offered 'private food for thought' by a senior member of staff that horizontal reporting lines 'are hard at FB and operationally would only make sense to do for big things e.g. creators, vs smaller efforts like well-being.'[97]

- In its response, we encourage Ofcom to set out whether it realistically considers that its proposals will be capable of securing meaningful changes to the organisational culture set out above.

- If it is unable to conclude that appropriate changes to accountability and risk ownership can be achieved, the regulator should be prepared to set out additional measures in its first iteration of the Codes.

<u>Quality assurance and risk monitoring</u>

- We welcome the requirement for large companies to have an internal monitoring and assurance function that can provide independent assurance to an overall governance body that its risk mitigation processes are suitably robust.

- Effective quality assurance mechanisms are essential to ensure that the risk assessment and horizon scanning activities being undertaken by large companies are effective. We consider this to be particularly important to ensure that new harms are proactively tracked and identified, particularly when there are perverse incentives for companies not to do this well.

[97] Ibid

- However, we are concerned that the regulator is not proposing external audit or oversight requirements. Some companies, such as Meta, have already undertaken external quality assurance audits of their transparency metrics.[98] More broadly, external quality assurance is an established part of other regulatory regimes.

- Furthermore, we are not convinced that these measures are being introduced in a way that can meaningfully overcome the pronounced corporate incentives to treat this and other relevant measures as simply a 'box ticking' exercise. We strongly encourage the regulator to set out that quality assurance and risk monitoring functions must therefore be of a suitable and sufficient quality, not simply exist.

- We also encourage the regulator to reflect on the absence of personal accountability measures for relevant managers and teams working on compliance and risk issues. In the absence of such measures, it is questionable that these proposals will deliver the changes in corporate accountability and risk ownership that is intended and manifestly required.

Code of Practice for staff

- We welcome Ofcom's proposal that regulated companies should adopt a Code of Conduct that sets standards and expectations for employees around protecting users from the risks of illegal harm. We also support the related measure that staff involved in the design and operational management of regulated products are sufficiently trained in the services approach to compliance.

- We agree with the regulator that Codes of Conduct and targeted training programs can be effective in ensuring that regulated companies effectively embed their risk management, mitigation and compliance approaches within the organisational culture. However, we note the regulator provides insufficient detail about the quality, composition or content of codes of either its proposed codes or training programmes.

- In the absence of further detailed requirements about Ofcom's expectations, it is not unreasonable to anticipate that companies may attempt to comply with these regulatory responsibilities in a light touch and ultimately ineffective way.

- We also question the decision not to extend these requirements to smaller services, in particular when there is ambiguity about whether medium-sized services such as Roblox, Twitch, Discord and Telegram will reach the threshold posed by Ofcom to be considered as a 'large' platform.

- There are clear and manifest safety risks associated with medium-sized platforms, and we are unclear how the regulator could determine that it is disproportionate or somehow

[98] Sarang, V (2022) Community Standards Enforcement Reports Assessment Results. Menlo Park: Meta. Blog posted 17/05/22

unnecessary to expect them to introduce such fundamental measures such as training or best practice requirements.

Bonuses

- We note that the regulator has declined to recommend measures relating to staff bonuses due to what it describes as 'limitations in currently available evidence that demonstrates the effectiveness and costs of these proposals.'

- We wish to remind Ofcom that other regulators are bringing forward measures that would prevent bonuses being paid to executives of firms that commit criminal acts. For example, Ofwat will prevent bonuses being paid to executives of water companies that commit criminal act of water pollution from 2024/25.[99]

- We strongly encourage the regulator to reflect on the particularly poorly aligned incentives in this sector for regulated companies to tackle illegal harms, both on a discrete basis and when in comparison with other markets; and note that if other regulators are identifying the need for measures that link bonus payments with action taken to avoid illegal behaviour, the market size of large companies and moral hazards associated with the tech sector are highly likely to necessitate at least similar action in this regime.

- We would also remind the regulator of the considerable deterrence value of these measures.

- We therefore strongly encourage the regulator to revisit its position on bonuses and online safety outcomes, and to apply measures that link bonus payments to the user experience of and exposure to illegal content on regulated services. This could beneficially form part of an effective harm reduction framework, as discussed further in the next chapter, in which bonuses should only be payable if and when a company can demonstrate annual reductions in the exposure to harm on its platforms.

- Ofcom might also choose to adopt a similar framework for the awarding of bonuses to its own online safety directors.

[99] Ofwat will consult on details of changes to its regulatory scheme later this year, in announcing the proposals the Environment Secretary Steve Barclay set out that bonus restrictions would apply to executives of any company that had committed 'serious criminal breaches.'

Section 4: Adopting a harm reduction framework

- Throughout our response MRF has extensively referenced internal company research commissioned by Arturo Bejar during his second period working at Meta. A full copy of the research (the BEEF framework) is included in appendix one.

- The data is particularly valuable because it provides a user-centred understanding of the experience of and exposure to harmful content among a key segment, teens aged 13-15.

- The BEEF survey highlights a clear disconnect between the metrics that Meta uses to externally report on the prevalence of harmful content to key audiences, including governments, regulators and advertisers; and the internal metrics it has at its disposal to track the exposure to and prevalence of harms among its users, including by key cross-breaks, content type and commercially sensitive categories such as whether users are categorised as content creators.[100]

- The BEEF metrics provide arguably the most robust means of assessing the prevalence of and exposure to harms on relevant platforms; and its research design minimises the risk that relevant metrics can under-report the prevalence of harmful content (whether purposely or by accident.)

- During the consultation period, Ofcom has emphasised that it intends to use a broad set of levers to ensure progress against its regulatory outcomes, with its supervisory, information disclosure and transparency powers being used in conjunction with each other as part of an overall regulatory approach.

- We see merit in the regulator making a more explicit connection between the operation of its risk profiles, codes of practice and transparency metrics.

- Specifically, we propose that the regulator should use transparency metrics in conjunction with the codes of practice to form an explicit, annualised harm reduction framework. Under this approach, Ofcom would use both its risk profiles and codes of practice to recommend appropriate measures to tackle and reduce exposure to harms; and through adopting a transparency programme modelled largely on the BEEF framework, it could then test whether the measures are successfully driving down rates of exposure to illegal harm.

- In this approach, the regulator would be able to set out additional measures that a platform should be expected to take, if either the exposure to illegal content among relevant groups of users increases or does not fall in line with specified thresholds.

- We also envisage that the model could be gradually tightened over time, setting more stringent measures to drive continual improvements in respect of the prevalence of and overall exposure to illegal content.

- We consider that this approach is more consistent with the regulatory approach envisaged by Parliament during its discussions on the Act. In many respects it is also closer in its application to the original Duty of Care proposed by the Carnegie UK Trust.

- As Perrin and Woods set out in their original Duty of Care proposal, a core harm reduction framework would 'create a pattern identifying harm, measuring it and taking action to reduce harm, assessing the impact of that action and taking further action. If the quantum of harms does not fall, the regulator [can work] with the largest companies to improve their strategies.'[101]

- Although Ofcom's approach implicitly suggests that it intends to work across a range of relevant levers, we are concerned that any decision by the regulator not to explicitly fuse together transparency metrics with its codes of practice may weaken the overall impact of the codes. A more siloed approach may also weaken Ofcom's ability to deliver continuous improvements in safety and the overall experience of users.

- We strongly encourage Ofcom to set out the merits of a harm reduction approach in its response, and to discuss the merits of adopting this approach when it publishes its response and finalised schemes.

[101] Perrin, W; Woods, L (2018) Harm Reduction in social media -a proposal. Dunfermline: Carnegie UK Trust.