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

July 2024

Summary

Molly Rose Foundation (MRF) welcomes the opportunity to respond to Ofcom's consultation on its regulatory scheme to protect children from online harmful content.

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

Suicide and self-harm should be major considerations for regulated services. A growing number of studies set out the relationship between exposure to harmful online content and resulting suicide and self-harm risks,[1] with suicide related online experience a 'common but likely underestimated antecedent' to suicide among young people'.[2] Suicide -related Internet use has been reported in 24 per cent of deaths by suicide among young people aged 10 to 19, equivalent to 48 deaths in 2023.[3]

Overall, we have substantive concerns about Ofcom's proposed approach. While the regulator's measures undoubtedly offer greater ambition than their previous proposals covering illegal content, in our assessment the measures are unlikely to be commensurate to the risk profile of many online services. As a result, the regulatory scheme being proposed is unlikely to prove sufficient to disrupt the scale and complexity of preventable harm on many regulated services.

[1] 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

[2] ibid

[3] 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

Ofcom has made a set of strategic decisions about how it intends to operate it regime that seem likely to blunt its overall impact. The regulator has made set of structural choices, including how it approaches proportionality, evidentiary thresholds and the precautionary principle, that demonstrably constrain the scope of its regulatory scheme.

These decisions may leave Ofcom fundamentally unable to deliver on the extent of the ambition envisaged by Parliament when it passed the Online Safety Act. MRF has substantial concerns about the impact of these structural measures.

In short, there are significant questions about whether these proposed measures will go far enough – and whether the measures can meaningfully translate into materially safer conditions for young people at risk of suicide, self-harm and poor mental health outcomes.

We consider it is questionable whether the proposals set out will deliver the 'strongest possible protection for children', as promised by the previous Government during parliamentary passage.

In its final proposals, we therefore encourage Ofcom to address a number of crucial structural and design issues, and to adopt a bolder and more assertive approach to how it identifies and tackles preventable online harms.

Key issues

- **Ofcom's approach falls significantly short of tackling the risk profile it is supposed to address**. As a result of Ofcom's structural and regulatory choices, regulated services could qualify for a 'safe harbour' while failing to address substantial ongoing risks to young people. The regulator's proposals demonstrate a lack of emphasis on safety-by-design, with an overreliance on prescriptive ex-post measures rather than proactive risk assessment and upstream mitigation requirements;

- Much needed **measures to prevent Primary Priority Content being recommended in children's feeds, and to reduce the amount of Priority Content being recommended to them, risk being undermined by how Ofcom intends to implement and enforce them**. Rather than setting a clear, outcome-based set of requirements, Ofcom will only require platforms to take action on content where it has evidence to believe that content is either Primary Priority or Priority Content – in other words, where wholly ineffective moderation arrangements are able to detect content that is likely harmful;

- The clear expectation of both Parliament and parents is that the Online Safety Act should enforce **minimum age requirements for online services**. We are dismayed that Ofcom has found itself unable to deliver a set of proposals that can effectively meet these expectations. MRF is wholly dissatisfied with the regulator's seemingly circular logic that it can enforce age assurance through risk assessment processes, and in effect incentivise companies to invest in highly effective age assurance solutions, when Ofcom itself has been unable to conclude these measures are robust yet (primarily because of its lack of research and delayed use of information disclosure powers);

- Ofcom's **proposals for content moderation processes are wholly unambitious and will likely have the effect of 'baking-in' deeply ineffective and inconsistent content moderation arrangements**. New MRF analysis underscores the substantial failure of current moderation approaches: large platforms such as Instagram and Facebook account for only 1 per cent of content moderation decisions relating to suicide and self-harm.[4] We present evidence that large platforms have significantly failed to address the risks posed by high-risk functionalities and features: for example, on both Instagram and TikTok, less than one in five suicide and self-harm content moderation decisions relate to video and image content – despite these being some of the highest-risk and most used surfaces on both platforms;

- The regulator demonstrates a few welcome examples of innovative and deeply ambitious thinking, but generally these are few and far between. We strongly support some genuinely innovative new proposals, such as the **requirement for teens to be able to offer negative feedback that can in turn actively inform the results of content recommendation systems**. However, we remain concerned that Ofcom's structural and strategic choices will disincentivise innovation overall and may result in decreased focus on safety-by-design, in favour of a more prescriptive, tick box approach to regulatory compliance.

Ofcom now has a crucial opportunity to reset its approach – and to assert a clear imperative for services to robustly test, identify and mitigate entirely preventable and reasonably foreseeable harm. The regulator must demonstrate it has a strong and coherent plan to achieve long-term harm reduction, and MRF looks forward to working closely with the regulator to help them achieve it.

Our response

MRF's response is structured as follows:

- in *section 1*, we set out our overarching concerns about Ofcom's proposed approach;
- in *section 2*, we explore the regulator's proposed approach to risk assessment and set out our concerns about the lack of a truly systemic approach;
- in *section 3*, we focus on the risk factors and drivers of harmful content, including how platform functionalities and features increase the risk of exposure to suicide, self-material and highly depressive content, and in turn the susceptibility of users to its harmful effects, and;
- in *section 4*, we respond to Ofcom's draft Code of Practice and set out a range of systemic recommendations for consideration.

[4] Molly Rose Foundation (2024) How effectively do social network moderate suicide and self-harm content? An analysis of the Digital Services Act Transparency Database

Section 1: overarching concerns

- Ofcom has made a series of strategic choices about its proposed approach to regulation that we fear may significantly constrain the ambition and impact of its regulatory scheme.

- We previously expressed concern about many of these choices in our response to the earlier consultation on illegal content. While we appreciate that Ofcom is in the process of considering this feedback, we consider it important to reiterate our concerns here.

- We have significant concerns that Ofcom's approach risks the overall effectiveness of the regulatory regime. Ofcom's overly cautious approach to a number of fundamental aspects of regulatory design could reasonably result in regulation that falls considerably short of the expectations of civil society, those with lived experience of preventable harm, and indeed of Parliament when it passed the Act.

- Ofcom has signalled that it sees its proposed approach as a first iteration and that it expects to expand on this framework over time. In practice, Ofcom appears to be adopting an inherently gradualist approach that stems largely from the constraining influence of its approach to regulatory design.

- We are particularly concerned that Ofcom's approach to proportionality, evidentiary thresholds, and its application of the precautionary principle, is likely to significantly inhibit its ability to oversee a regulatory scheme that is commensurate to the risk profile of many online services.

- In our assessment, Ofcom's structural decisions are likely to exert a significant, long-term constraining influence on the scope and ambition of its regulatory scheme, and they are unlikely to result in a regime that delivers 'the strongest possible set of protections for children', as promised by Government and set out in section 1 3 (b) of the Act.

Ofcom's structural choices

- Ofcom has made a number of structural choices which are inherently problematic, and that we are concerned will result in an overly prescriptive, tick box approach to governance and compliance.

- We are particularly concerned with the piecemeal way in which Ofcom has approached the selection of measures contained in the codes - only adding measures where it considers there is appropriate evidence - rather than considering the risk-based outcome that the legislation had clearly envisaged.

- We wholly agree with the Online Safety Act Network that unless the combined response to the illegal harms consultation and this consultation suggests a significant shift in approach, it is unlikely we will see a systemic regulatory approach, rooted in adequately realised risk assessment and safety-by-design principles, that Parliament intended when the Act was passed.[5]

- We set out our concerns about each of these design choices, and their potential negative impacts on regulatory outcomes, below.

Weak approach to 'safety-by-design'

- In our assessment, Ofcom's approach is insufficiently grounded in a safety-by-design approach, and we are therefore concerned that the choices that the regulator has made in developing its proposals inadequately align with the overall objectives of the Act.

- Parliament clearly envisaged that the Act would result in a systemic, risk-based approach that targeted upstream risk identification and mitigation i.e. safety-by-design. Most notably, Schedule 4 sets out a series of Online Safety Objectives, including the objective that: 'a service should be designed and operated in such a way that -
  (i) the systems and processes for regulatory compliance and risk management of effective and proportionate to the kind and size of service,
  (vi) the service provides a higher standard of protection for children than for adults, and
  (vii) the different needs of children of different ages taken into account.

- The Act also specifies that a service should be 'designed and operated so as to protect individuals in the United Kingdom who are users of the service from harm, including with regard to-
  (i) algorithms used by the service,
  (ii) functionalities of the service, and
  (iii) other features relating to the operation of the service

- Our assessment is that the regulator has failed to deliver an approach to 'safety-by-design' that is consistent with the objectives set out above. Ofcom focuses on individual and often ex-post measures, rather than building a set of measures that form part of a coherent and intersecting approach to upstream risk identification and mitigation.

- In places, the regulator's approach to safety-by-design is unacceptably thin. As it stands, Ofcom does not propose mandatory upstream product safety testing, nor is there a requirement to mitigate the harms caused by all functionalities integral to their product, even if a platform's own risk assessment identifies a potential for resulting harm.

[5] Online Safety Act Network response to Ofcom's Protection of Children consultation

- Ofcom should migrate towards an outcome-based approach. The regulator's current atomistic approach focuses on gradualistic strengthening of the regime, with the adoption of additional recommended measures over time and as underpinning evidence emerges. This is inherently a prescriptive, tick-box based approach, and it has the reasonably foreseeable effect of incentivising companies to focus on a discrete programme of compliance rather than the delivery of substantive and proactive safety-by-design.

- We are furthermore concerned that Ofcom's approach could disincentivise innovation among regulated services. In effect, platforms will be poorly incentivised to develop safety-by-design approaches that exceed the regulator's requirements - and may face a moral hazard if they choose to do so, knowing that new safety-by-design measures will provide the evidentiary basis to extend the regulatory burden placed upon them.

Economic application of proportionality

- Ofcom appears to take a highly precautionary approach to imposing measures on regulated firms, and it has largely opted not to recommend measures where it deems costs to be disproportionate (or where there is limited evidence about the proportionality for smaller or medium sized firms.)

- In contrast, Ofcom has seemingly chosen to underweight or even overlook the social and economic costs of technology-facilitated harms. The regulator seems to apply an implicit requirement that the costs associated with user or societal harms must be demonstrably identified and expressly proven as a precondition for the proportionality of acting on them 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 to users experiencing or at serious risk of online harm.

- Ofcom's application of economic proportionality is one of several significant judgements on which the regulator is not consulting, but that fundamentally affect the shape of the resulting framework. We previously raised our concerns in our response to the illegal content consultation. MRF continues to see no legal basis in the Act that requires Ofcom to attach greater primacy to the costs to be faced by companies of addressing harms, over the externalised costs of these harms being passed on to service users and society.

- Ofcom should be prepared to set out how it has approached the linked issues of proportionality and the economic basis for choosing to recommend measures. This should include a description of the economic model that informs and is actively underpinning its regulatory approach.

- In particular, Ofcom should articulate and justify its approach to how it assesses the magnitude of, and costs associated with relevant harms, including the economic

calculations that inform whether it determines that a measure being recommended to tackle a relevant harm is proportionate or not. 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 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 on illegal harm costed the social and economic cost of a death by suicide as £1.67 million in 2009 prices (£2.23 million in 2023 prices).[6] 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.'[7]

- MRF analysis suggests that if the J-value method is applied instead, the estimated value of a human life - and the corresponding economic justification for recommending measures to prevent fatalities - increases by more than fourfold.

- 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 48 deaths in 2023.[8] 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 (assuming £11.3 million per death in 2024 prices.)

- Ofcom must be prepared to set out how it approaches economic proportionality. The regulator provide assurance to civil society and bereaved parents that it is appropriately considering the social and economic costs of externalised harms.

- If Ofcom feels it cannot or will not appropriately weight the externalised costs of harms, Parliament may want to consider whether changes to the legislative framework may be required.

[6] Ofcom (2023) Consultation on Illegal Content

[7] 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.

[8] 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

Evidentiary thresholds

- 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 introduce them.

- 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 children. This approach is likely to have a significant constraining impact on the effectiveness of the regulatory regime, and put bluntly, is likely to result in children continuing to be exposed to entirely preventable harm.

- In our response to the illegal harm consultation, we expressed surprise that Ofcom had sought to develop its codes without appropriate consideration of the precautionary principle. We are deeply disappointed that a precautionary approach is still absent from the regulator's proposals. Invoking and applying this 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.[9]

- The precautionary principle creates the 'impetus to take decisions notwithstanding scientific uncertainty about the nature and extent of the risk'.[10] 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 harms referenced in the Code, and the well-observed evidentiary challenges associated with demonstrating a causal relationship between online platforms and many of the illegal harms in scope, Ofcom's approach has the effect of being overly cautious in favour of regulated companies, and restricting the protections that it offers to young people being exposed to online risk.

---

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

[10] ibid

- 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. 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.[11]

- 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.

- We strongly encourage Ofcom to 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 (as discussed in section 4).

- We remind the regulator that the precautionary principle is actively enshrined in European law, with a growing number of legal challenges in respect of insufficiently protective legislative and regulatory instruments.[12] Procedural regulatory arrangements, the scope of risk assessments and the distribution of regulatory authority are all matters that case law has determined can be legitimately underpinned by the precautionary principle being applied.

- In its consultation response, we therefore 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. The regulator should clarify what if any barriers it perceives exist in the legislative framework that may prevent it from adopting such an approach. This will enable Parliament and civil society to understand if there are legislative or drafting issues that may need to be revisited.

Application and balancing of fundamental rights

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

[11] 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).

[12] Leonelli, G (2021) Judicial Review of compliance with the Precautionary Principle from Paraquat to Blaise: Quantitative Thresholds, Risk Assessment, and the Gap between Regulation and Regulatory Implementation. German Law Journal, 22 (2), pp184-215

- 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 place disproportionately focus on the fundamental rights of users posting content, and inadequately considers the chilling impacts of harmful speech on the right to free expression and association of others, including children.[13]

- Ofcom is in effect interpreting Section 22 as a measure that constrains the overall ambition of its regulatory scheme. In multiple parts of Volume 3, the regulator's approach cites adverse impacts on free expression as grounds not to proceed with recommended measures.

- 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[14] demonstrates that the failure of the company to adequately prevent teen users being exposed to unwanted harms has had an adverse impact on their 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.[15]

- 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 relevant characteristics.[16]

- 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,[17] particularly where that person is a child.[18] It is difficult to conceive how Ofcom's proposals meet the positive obligations placed upon them. [19]

---

[13] Woods, L (2024) Ofcom's approach to human rights in the illegal harms consultation. London: Online Safety Act Network

[14] 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

[15] ibid

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

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

[18] 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

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

- 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. We therefore remind the regulator that it 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 to disrupt or reverse the scale and magnitude of many har have not been met.

Safe harbours

- The combined effect of the measures set out above is to grant a 'safe harbour' to platforms if they comply with a Code of Practice which in some respects falls substantially short of what is required to or reverse the scale and magnitude of preventable harm.

- Under Ofcom's proposed approach, the regulator will consider an online service to be compliant with its child safety duty set out 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. Section 41 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, as set out above.

- 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 through which meaningful improvements in online safety outcomes would result.

- In its final proposals, we therefore strongly encourage Ofcom to reassess the merits of its structural choices.

Section two: risk assessment and scope of children's duties

- Ofcom sets out its proposed approach to risk assessments in Section 12. Here, we set out a number of concerns about the regulator's proposed risk assessment framework.

- We are particularly concerned by the absence of high-quality product testing. In our assessment, product testing is a prerequisite for effective risk identification and mitigation, and the absence of relevant measures may substantially undermine the effectiveness of the risk assessment process.

Risk assessments and product safety testing

- Risk assessments are a crucial part of the online safety regime, and because the risk assessment outputs are directly enforceable under the Act, it is hugely important that the regulator sets out a stretching and ambitious approach to the underpinning process.

- The Act requires regulated services to complete risk assessments to a 'suitable and sufficient' standard. While we support Ofcom's proposals that regulated services should follow a structured and systematic approach, we are concerned that the regulator has developed a risk assessment process that places disproportionate weight on the resource implications for regulated firms, and that fails to appropriately incentivise or require risk assessment being produced to the necessary and/or highest quality.

- The regulator explicitly states it has sought to propose 'the least onerous approach to ensure that services understand the risk their service poses to children.' We wish to remind the regulator that effective risk assessment is an integral component of embedding safety-by-design outcomes into a regulatory scheme.

- The absence of product testing processes is a major concern. In our assessment, the lack of substantive product safety measures risks substantially undermining the effectiveness of the risk assessment process, and in our view falls this approach considerably short of the upstream product testing approach envisaged by Parliament when it passed the Act.

- We strongly encourage Ofcom to reconsider this part of its approach. It is wholly inappropriate to envisage product testing as an optional and/or enhanced input, as the regulator seems minded to do, rather than treating this a crucial enabler of good regulatory outcomes.

- We therefore fully endorse the recommendation made by the Online Safety Act Network that product testing should be undertaken as a core input, by all services that are large

and/or medium and high risk.[20] This should be accompanied by a corresponding requirement to redesign, limit or even restrict high risk functionalities or product features depending on the mitigations that can reasonably be achieved.

- In discussions with civil society, Ofcom has stated that because the Act is 'technology neutral', it feels unable to adopt a position that regulated companies should not offer high-risk functionalities, until and unless it can be confident the relevant functionality is safe. Ofcom should be prepared to set out how it has arrived at this position, and the regulator should set out which if any provisions in the Act it feels prevents it from adopting a more assertive, risk-based approach.

- We also encourage Ofcom to consider the application of an overarching product testing requirement, drawing on the Outcome Testing requirements in the FCA's new Consumer Duty. Under this part of the Duty, regulated firms are expected to test the likely impacts of their products from the design stage all the way through to real-world outcomes.

- The FCA's requirements create a continual set of feedback loops, with providers expected to use the outputs from Outcome Testing to continually update and refine their services.

- As Ofcom finalises its approach to its regulatory scheme, we strongly encourage the regulator to explore a set of similarly overarching approaches. Ofcom could positively adopt a set end-to-end measures that can effectively target harm reduction, incentivise continual improvement across the product life-cycle, and ensure appropriate resource is directed into risk mitigation and proactive safety by design. [21]

Size, risk and number of child users

- We remain concerned that Ofcom continues to envisage the impact of harm primarily as a function of a platform's size and user base. While we welcome the regulator's clarity that the number of children using a platform is only one of the various risk factors that services should consider when they determine their risk level, and that in some instances it may actually be a weak indicator of risk levels, it is striking that size is the only criteria that Ofcom proposes should result in a default assumption of high impact.

- We invite Ofcom to set out why it considers other high-risk design choices and functionalities should not automatically result in similar default categorisations, not least the high-risk functionalities set out in Volume 3.

[20] Online Safety Act Network (2024) Response to Ofcom's Protection of Children Consultation
[21] Grant Thornton (2024) Consumer Duty Outcome Testing; How to meet the FCA's expectations

- Ofcom appears to be taking an approach to platform size that assumes a platform has a significant number of child users, and should therefore be considered high risk, if it exceeds a threshold based on the total number of child users aged 0-17. Given most regulated services have a minimum age limit of 13, this seems a confusing approach.

- In our assessment, where a platform set a minimum user age of 13, the service should be considered high risk if the 7 per cent threshold is met in respect of its user base aged 13 to 17, rather than assessing this against the broader pool of child users aged 0-17.

Core and enhanced risk assessment inputs

- MRF broadly supports the principle that core and enhanced evidence inputs should inform a platform's risk assessment. However, we have several concerns about how Ofcom proposes to embed this in the risk assessment process.

- Firstly, Ofcom only sets out a limited range of core measures, seemingly driven by a focus on economic proportionality. The set of measures are so limited that there are significant questions about whether these measures will be able to adequately capture the nature and magnitude of relevant reasonably risks. The proposed list of core inputs set out in table 12.3 appear particularly poorly suited to identify new and emerging harms.

- We strongly disagree with Ofcom's assertion that in some instances a platform could be able to undertake a suitable and sufficient risk assessment solely by relying on core inputs. There are significant issues with the quality and reliability of many of the core measures that the regulator proposes.

- For example, of the eight main core inputs identified by the regulator, three of these – the outputs of content moderation systems, user complaint data, and data on the age profile of users – are wholly reliant on at best inconsistent and unevenly applied current processes. MRF cautions that Ofcom's first iteration of the Code is unlikely to materially improve the effectiveness of content moderation or user reporting flows, and Ofcom itself has said that highly effective forms of age assurance don't yet exist.

- Ofcom's list of enhanced outputs is likely to be a more effective means to identify and act on reasonably foreseeable harms, although we encourage the regulator to address several areas where we envisage that its proposals could be readily gamed.

- Firstly, while the regulator correctly identifies that platforms could use external audits or other risk assurance processes to support its risk assessment process, in practice we have seen several examples of firms already use this approach to 'bake-in' existing largely ineffective existing processes and/or to project questionable legitimacy onto future plans.

- For example, in autumn 2022 Meta commissioned a nominally independent human rights assessment of its proposed rollout of end-to-end encryption that found that the

potential proposals wouldn't directly cause adverse impacts. This conclusion was only reached because the report opted to classify Meta as being 'directly linked to', rather than 'contributing to', the potential adverse impacts on child sexual abuse and exploitation of its own proposals.[22]

- Secondly, Ofcom encourages regulated firms to consult with independent experts and representative groups. While the children's sector will of course wish to engage in good faith with any regulated firm that seeks to meaningfully improve their child safety outcomes, many groups may be legitimately concerned that regulated companies may seek to engage with us only a tick box basis, with engagement primarily being undertaken for performative rather than risk-based purposes.

- Thirdly, Ofcom needs to set out how it intends to manage the risks associated with academic and expert capture.[23] A number of regulated firms are likely to be incentivised by the regulator's guidance to actively commission external research, insight and data, but there is a reasonable chance this activity may be driven by an intent to add legitimacy to a platform's current moderation and design processes and/or to seek the skew the evidence base relating to exposure to harm on its service.

- Given the substantial and foreseeable risks that platforms may seek to game this process to reduce the magnitude of risks on their site, and in turn limit the scope of their compliance costs and enforcement risks stemming from Section 12, we strongly encourage the regulator to provide additional guidance about the quality, sufficiency and independence of third-party inputs.

- The regulator should be prepared to offer detailed guidance about how it intends to manage the interplay of moral hazards and conflicts of interest that may stem from the application of this part of its regulatory scheme, and how it intends to track and act against the unintended consequences that may result.

[22] BSR (2022) Meta's expansion of end-to-end encryption: human rights impact assessment.
[23] Academic and expert capture has been identified as a means through which corporate interests seek to influence the evidence base on which regulatory decisions are taken. See for example Abdalla, A et al (2021) The Grey Hoodie Project: Big Tobacco, Big Tech and the Threat on Academic Integrity. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics and Society, pp287-297

Section 3: risk factors and understanding of drivers of harm

- This section of our response focuses on Ofcom's understanding of the drivers and dynamics of content harmful to children, with particular focus on the risk profiles and register of risks set out in Volume 3.

- In its risk profiles, Ofcom sets out a range of ways in which platform design choices and product features may facilitate exposure to harmful content, including suicide and self-harm material, dangerous stunts which may lead to a child's death, and content that encourages or promotes the ingestion or consumption of harmful substances.

- We strongly support Ofcom's understanding of harm dynamics and archetypes, including its identification of risks associated with both isolated and cumulative exposure to harmful content. We agree with Ofcom that cumulative harm can occur where a child is both repeatedly exposed to harmful content and/or combinations of harmful content.

- Ofcom sets out a range of risk factors associated with specific functionalities and service characteristics. While this assessment is well developed in many places, there are a number of product functionalities that appear to have been omitted or where the level of detail falls short.

- We also consider that Ofcom has substantially underplayed the role that business models play in the commercial and design decisions of regulated services. In our assessment, business models and commercial profiles are a significant and arguably growing contributory risk factor for children's exposure to harmful content, particularly for Category 1 services.

- In our assessment, Ofcom is missing a substantial opportunity to deliver upstream harm reduction by focusing only on the risks associated with specific functionalities and service characteristics, rather than on a set of underlying mechanics and drivers of harm.

- In its final response, we strongly encourage the regulator to set out the likely psychological, cognitive, emotional and psychosocial effects of the design choices it identifies, and to identify that that a consistent set of risk factors (for example, assortative relating, emotional dysregulation, rumination and maladaptive emotional strategies) are likely to result from and apply across a number of high-risk functionalities and features.

- The regulator could usefully develop a set of risk exacerbation and harm pathways associated with relevant design features. These should set out the ways in which design features work in isolation or conjunction to increase the risks associated with harmful content. By illustrating the underlying cognitive, psychological and behavioural effects at play, these models would usefully support the shift from an approach focused on discrete risk identification towards greater proactive emphasis on safety-by-design.

- A set of suggested risk exacerbation pathway models is set out for consideration in Appendix 1.

- As Ofcom finalises its proposals, we strongly encourage it to recognise that the evidence base in respect of suicide and self-harm content is still actively developing, and that the currently available evidence is generally less advanced than some other harm archetypes, for example CSEA. In this context, Ofcom must resist conflating absence of evidence or harm with the absence of harm altogether.

- We therefore strongly encourage Ofcom to adopt a precautionary principle approach to the risks of harmful content. As an overarching approach, Ofcom should be prepared to require companies to act on potential risks wherever it is more likely than not to assume that reasonably foreseeable harm may occur.

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

The scale of and prevalence of harmful content

- 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. 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.[24]

- 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 Framework (BEEF), 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.[25]

- MRF research has found that substantial amounts of harmful suicide and self-harm content remain readily accessible and discoverable on major social networks.[26] Almost half of the most engaged posts on TikTok (49%) and Instagram (48%), and that were

[24] The Bad Experiences and Encounters Framework research can be found in appendix one
[25] Ibid
[26] 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

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.

- 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, 99% of the short form videos we were algorithmically shown on Reels, through watching a set of posts recommended by the app's autoplay function, contained at least one type of relevant harmful material. More than half of posts referencing suicide ideation (often through graphic and slickly produced memes.)

- We consider it likely that this these results stem predominantly from 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. [27]

Adverse effects of harmful online content

- Suicide is the third leading cause of death among 15 to 19-year olds,[28] and the most recent annual figures indicate that 524 people aged 24 under died by suicide in the UK.[29]

- Findings from multiple studies have raised concerns about the harmful effects of exposure to self-harm and suicide related online content;[30] the impact of engaging with material that promotes, glorifies or incites serious acts of self-injury;[31] 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'

---

[27] The Bad Experiences and Encounters Framework, publicly released by the Attorney General in New Mexico in January 2024

[28] Department of Health and Social Care (2023) Suicide Prevention in England: Five Year Cross Sector Strategy

[29] 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

[30] 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

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

to suicide in young people.[32] Suicide-related internet use has been reported in 24% of deaths by suicide among young people aged 10 to 19, equivalent to 48 deaths in 2023 [33]

- Suicide and self-harm related internet use has been reported in 26% of child hospitalisations relating to self-harm.[34] We agree with Ofcom that it is practically difficult to determine between suicide and self-harm online content that is often highly interconnected in nature, and we therefore support its approach to consider both harm types as interrelated. 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 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).[35] There were 42,793 admissions among young people aged 10-24.[36] In 2014, one in five female 16- to 24-year-olds reported non-suicidal self-harm, a threefold increase since 2000.[37]

- 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 of girls, and in cases affecting adolescents who are identified as LGBTQ+. [38]

- 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.)[39] Among young people aged 10-19, we estimate that in England alone over 8,100 annual admissions are associated with harmful internet material each year. [40]

---

[32] 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

[33] 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

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

[35] Nuffield Trust (2023) Hospital admissions as a result of self-harm in children and young people.

[36] Office for Health Improvement and Disparities (2024) Public Health Profiles: Self-Harm.

[37] 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

[38] ibid

[39] 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

[40] 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)

Mechanics and drivers of adverse impacts, including demographic risk factors

- 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 both harmful and protective effects, a recent systematic review concludes that harmful effects predominate.[41]

- 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;[42]

    - 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;[43]

    - engaging with self-harm content may result in *emotional, cognitive and physiological impacts*, which may trigger or exacerbate self-harm behaviours and suicidal thoughts;[44]

    - 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;[45]

    - 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; [46]

    - *an adverse 'assortative relating' effect, in which young people experiencing suicide ideation or thoughts of self-harm are more likely to identify and build*

[41] 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
[42] 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
[43] ibid
[44] 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
[45] ibid
[46] 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

relationships with other users experiencing similar actions and thoughts[47]. Although this technology-facilitated effect may provide adolescents with much needed immediate connection, validation, help and support[48], it also presents significant risks (including the potential for unintended consequences.) For example, self-harm may become 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'.)[49]

- Studies point to a higher risk of adverse impacts associated with suicide and self-harm content in adolescent girls[50] and those already experiencing poor mental health, including health conditions such as depression, anxiety and poor body image.[51]

- A recent systematic review found that adolescents with clinical level mental health problems may be particularly vulnerable to digitally mediated harm.[52] Young people diagnosed with depression reported more problematic internet use, as well as difficulties in regulating their digital engagement compared to their nonclinical peers.[53]

- Recent research suggests that the consumption habits of young people experiencing depression and other mental health conditions may interrelate with high-risk platform functionalities and features to increase the risks associated with harmful content, and that clearly observable digital harm pathways may result.

- For example, exposure to suicide, self-harm and depressive content may result in young people being exposed to a broad set of risk exacerbation mechanisms, and among children experiencing depression and/ or poor mental health, this may this result in them experiencing more negative emotional and cognitive framing.

- Once children start to engage with harmful content, platform design features may significantly accentuate the potential risks. For example, platform algorithms may reinforce and amplify depression-related search preferences, which in turn may lead to

---

[47] 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

[48] 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.

[49] ibid

[50] 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

[51] 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

[52] 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

[53] 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

a vicious circle in which children are more likely – as a result of their cognitive biases being reinforced - to interpret content more negatively.

- Some adolescents may be consequently more likely to experience negative rumination about content and/or negative experiences they've had. In turn, some may go on to require an excessive need for reassurance or approval, which may increase the risk they go on to experience or seek out more harmful or risky online interactions.[54]

- 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, particularly girls.[55] 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 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.

- There are significant risks associated with the ways in which content is curated, searchable and/or algorithmically recommended. For example, a substantial amount of young people may actively go use regulated services for help-seeking, validation or support,[56] but popular content posted by accounts and recommended using search terms may readily switch between content which is beneficial and that which may result in direct or cumulative harm.

- As a result, young people may opt to use regulated services for positive and potentially beneficial purposes, but as a result of the interplay between platform design choices, search curation and algorithmic recommendation systems, their search behaviour may actually result in maladaptive effects.

- In its final response, Ofcom should set out how it intends that companies should consider and mitigate the risks associated with such use cases, including in its risk assessments.

---

[54] Sonuga-Barke, EJS et al (2024) Pathways between digital activity and depressed mood in adolescents: outlining a developmental model integrating risk, reactivity, resilience and reciprocity. Current Opinion in Behavioural Sciences, 58

[55] 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

[56] Research shows most young people have a preference for self-reliance when experiencing personal and emotional concerns and are therefore more likely to use informal help sources, readily available online, if and when they seek support. Pretorious, C et al (2019) young people's online help seeking and mental health difficulties: Systematic Narrative Review. J Med Internet Research, 21(1)

Effect of functionality-driven risk factors

- We are pleased that Ofcom has extensively set out the risks associated with several functionalities and features that contribute to the risk of young people being exposed to harmful content. The regulator correctly identifies the risks associated with a broad set of high-risk design features, particularly the impact of personalised recommender systems.

- However, it appears that Ofcom has omitted a number of high-risk design choices, including several functionalities previously identified by MRF's research. In respect of some other features, the regulator's analysis appears to be underdeveloped.

- We assess of the omitted or underdeveloped functionalities in turn below.

Saving and sharing functionality

- 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.

- Our research suggests that users are being able to readily amass large albums of harmful content. Almost a third of harmful posts that we analysed on TikTok (30 per cent) had been saved by at least 30,000 users, and 4 per cent of harmful posts had been saved over 50,000 times.[57]

- 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 the onset thoughts of self-harm and suicide ideation.

- Saving and sharing functionalities are an integral part of clearly observable suicide and self-harm risk pathways: vulnerable teenagers can be exposed to large amounts of harmful material as a result of personalised recommended systems and other design features; this content can go on to have a detrimental impact on their mental health; and in some cases, adolescents will go on to amass substantial volumes of albums or collections of harmful content to 'binge watch' on-demand, including as a maladaptive coping response.

- We wish to remind the regulator that 'binge watching' content was a prominent feature of the circumstances surrounding Molly's death.

[57] 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

- In our assessment, saving and sharing functionality is one of the highest risk design choices for suicide and self-harm content. Ofcom must more substantively address the resulting risks in its first iteration of the Code.

Comments and discussion spaces

- 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 young people, with the most substantial risk profiles associated with both large social media sites and high-risk suicide discussion fora.

- While Ofcom addresses this issue in chapter 7, in our assessment this analysis is somewhat underdeveloped and inadequately considers the underlying mechanics that may result in or exacerbate harm among young people.

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

- MRF's recent research into the nature and prevalence of suicide and self-harm risks on TikTok 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.[58]

- 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. We strongly encourage Ofcom to recognise and set out the underlying psychological, cognitive and behavioural mechanics that may result in exposure to harm and deepen its impacts among young people.

- For example, research suggests that large volumes of comments may risk normalising self-harm as an acceptable coping strategy, trigger emotional dysregulation effects and/or may encourage adolescents to understand that suicide ideation and self-harm behaviours are more common than they actually are.

[58] 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

- Numerous studies have shown the risks associated with a lack of protective measures in such unmoderated or poorly moderated communities,[59] with a risk that young people may identify such spaces as helpful despite the range of unintended adverse consequences set out above. [60]

- 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.' [61]

- 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 harmful content and illegal activity can be readily observed.

- X has recently introduced its Spaces surface, group channels covering specific topics and interests that are algorithmically recommended in user feeds. These include a number of community spaces dedicated to suicide and self-harm. As recently as July 2024, MRF observed several Spaces that shared high-risk and potentially harmful content, including posts that promote suicide and self-harm behaviour.[62]

- One of these Spaces was dedicated to the hashtag #shtwt, which previous research has found is used to share an extensive range of suicide and self-harm content on the platform.[63] Our analysis found a significant proportion of harmful content being posted in the group, including graphic images of self-harm and material that promoted and encouraged suicidality and self-harm behaviours.

- MRF was algorithmically recommended the #shtwt group. This appears to contradict X's previous claims that it would no longer recommend this hashtag to users.

Search and discoverability features

- 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. This includes children who may be proactively seeking helpful or supportive content.

---

[59] Bell, J et al (2018) Suicide Related Internet use among suicidal young people in the UK: characteristics of users, effective use, and barriers to off-line help seeking. Arch Suicide Research, 22(2), pp263-277
[60] Mars, B et al (2015) Exposure to, and searching for, information about suicide and self-harm on the Internet: prevalence and predictors in a population-based cohort of young adults. Journal of Affective Disorders, 1; 185, pp239-245
[61] 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.
[62] Desk analysis undertaken by Molly Rose Foundation
[63] 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

- Our recent research observed a number of high-risk design features on both TikTok and Pinterest, including video search recommendations that were highly problematic. These included including hashtags and search terms linked to harmful content ('people also search for 'quickest way to end it.')[64]

- 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.'

- We note that autocomplete suggestions were treated as a distinct risk factor in Ofcom's illegal content response, and that in respect of the risks of financial fraud, the regulator recommended proactive measures to tackle resulting harm. However, autocomplete proposals do not feature prominently in these proposals at all.

- Pinterest uses a range of particularly intrusive 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.[65]

- Hashtags continue to play a substantive role in enabling users to readily access and discover potentially harmful suicide and self-harm content.[66] 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 has found virtually no sensitivity screens in place (in less than 1 per cent of harmful posts).

- MRF data suggests that sensitivity screens and other measures designed to introduce friction into the search experience have been applied highly infrequently if at all. Our forthcoming analysis of Meta's content moderation decisions, submitted to the Digital

[64] 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

[65] 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,

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

Services Act Transparency Database, suggests that neither Instagram nor Meta used sensitivity screens for any items of content between September 2023 and April 2024.[67]

- Similarly, and despite promising that it would start to age-restrict some forms of suicide and self-harm content from appearing in the feeds of under 18s at the start of this year, as of April 2024 Meta had failed to age-restrict a single item of its content.[68]

- In its DSA submissions, Facebook claims that it had downranked 12 per cent of suicide and self-harm content it had reviewed during our review period, but it has not set out further information about the grounds or criteria that prompted it to do so.

Biographical features

- Our recent research found multiple ways in which accounts distributing harmful suicide and self-harm content able to exploit platform design features, including the use of biographical features to embellish their mental health standing and expertise.

- 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 these accounts are 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. This is particularly problematic given the increased tendency of younger users to use social media to search for supportive or helpful mental health content. [69]

- 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.

- It is likely that biographical information may make it more likely young followers to engage with relevant forms of content. Evidence suggests that young people value online services run by mental health professionals[70]and that there is a keen interest in

[67] Molly Rose Foundation (2024) How effectively do social networks moderate suicide and self-harm content? An analysis of the Digital Services Act Transparency Database
[68] Ibd
[69] Pretorious, C et al (2019) young people's online help seeking and mental health difficulties: Systematic Narrative Review. J Med Internet Research, 21(1)
[70] Frost, M (2016) Who seeks help online for self injury? Are Suicide Research, 20 (1), pp69-79

obtaining help on social media from those who claim expert credentials.[71] Where their critical data and mental health literacy skills may be lacking,[72] some young people may be more likely to take exaggerated or inaccurate claims at face value.

- 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 and/or encouraging users to migrate onto and share content on private or restricted platforms.

Ephemeral stories and broadcast channels

- 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 commissioned 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 to share or produce harmful content.

- 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)[73] and substantial social learning effects (where there is a

[71] Birnbaum, ML et al (2017) Role of social media and the Internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. Early Interventions in Psychiatry, 11 (4), pp290-295
[72] Evidence finds that the ability of young people to access reliable, helpful information is influenced by their lack of mental health literacy and limited means to assess the credentials of informal online materials.
[73] ibid

risk that mood behaviours may be modelled or imitated based on exposure to the shared characteristics of the group.)[74]

- 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 its effects.

DMs, groups and private messaging

- 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.[75]

- 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 to share or produce illegal content, is clear.

- Recent evidence suggests that encrypted group chats are being used to share harmful material, including suicide, self-harm and highly depressive content. A recent BBC investigation found that WhatsApp groups were being used to share suicide and health self-harm content in a number of schools in NE England[76]. In a number of cases, it is understood that the groups were encouraging young people to commit acts of self-injury, and that a number of related hospitalisations occurred.[77]

- 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 encrypted private messaging services are likely to have a particularly high-risk profile for such content.

- Through our work with other bereaved families, including the Bereaved Families for Online Safety, we are aware that a number of 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

[74] 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
[75] See appendix one.
[76] BBC News (2024) Nine Year Olds Added to Malicious WhatsApp Groups
[77] Molly Rose Foundation discussion with journalists from BBC Newcastle

number of users in the months before her death, but without access to the relevant messages are unlikely to ever learn the underlying context.

Impact of business models on risk profile

- While Ofcom correctly recognises that advertising-based revenue models may contribute to the risk profile of user-to-user services, in our assessment Volume 3 significantly underplays the role that business models play in the commercial design decisions of regulated services.

- Ofcom's lack of emphasis on business models is problematic, with substantial evidence that the business models of regulated services are a primary driver for relevant harms, particularly category 1 services.

- As Ofcom sets out, regulated services currently use a range of user engagement mechanisms that are designed to maximise time spent using services. Following our recent research into harmful content on Pinterest and Instagram, both platforms have continued to send emails encouraging us to log-on to our accounts, featuring personalised recommendations that contain suicide, self-harm and highly depressive content.

- These emails are broadly identical to those Molly received in the months before her death.

- We note that Pinterest had committed to cease sending such emails in November 2022, as set out in its response to the Prevention of Future Deaths report issued following Molly's inquest.[78]

- More broadly, there is increasing evidence that the risk profile on Category 1 services is directly linked to their commercial profile. As set out above, there is a clear relationship between the commercial importance of the various product surfaces on Instagram and their respective significance to the company's revenue and growth targets.

- For example, Instagram has repeatedly stated it identifies Reels as a priority focus in its efforts to maximise user engagement on the platform. In its quarterly earnings calls, Meta consistently highlights the performance of Reels against its preferred user engagement metrics. In the company's Q4 2023/24 call Mark Zuckerberg stated that Reels now accounts for half of all time spent on the platform.[79]

[78] Letter from Pinterest to Senior Coroner Andrew Walker responding to the Prevention of Future Deaths report issued following the inquest into Molly's death
[79] Meta Q4 Earnings call, transcript

- It seems reasonable to assume that Instagram's focus on user growth on Reels is causing it to roll back safety thresholds on that part of its platform. Our research found a markedly increased risk profile on Reels than any other part of Instagram, with 99% of algorithmically recommended videos containing harmful content.

- It seems difficult to conclude that this is driven by any other factor than a race for short-form video market share. As Meta warns investors in its annual SEC filings, "If we fail to retain existing users or add new users, or if our users decrease their level of engagement with our products, our revenue, financial results, and business may be significantly harmed.[80]

- If Ofcom cannot conclude similarly, we would expect it to use its supervisory and information disclosure powers until such time it can suitably disprove this assessment.

- We also encourage the regulator to accept that Trust and Safety has typically been treated as a cost centre by regulated services, and that this is a direct function of the commercial and risk profile of major platforms. There are compelling grounds to consider this a market failure, with the UK's competition regulator having previously stated that user safety should be considered as a characteristic of a well-functioning market.[81]

- Extensive legal failings have demonstrated that Meta has consistently opted not to invest in youth safety and well-being, despite direct knowledge of the causal relationship between its product design and exposure to harmful self-harm and suicide content.

- For example, internal emails disclosed as part of the lawsuit brought by US Attorneys General shows that senior Meta staff were fully aware of the risks associated with suicide and self-harm content being recommended by its algorithms, with an internal report finding there was a 'palpable risk of "similar incidents [to Molly's death]" because its algorithmic features were "[l]eading users to distressing content.'[82]

- In the weeks following the initial media coverage of Molly's death, emails released by the US Senate show that members of Meta's Executive Team were expressly told that Instagram's product design meant that suicide and self-harm content was being presented in search results a way that contravened its content policies. [83]

- Despite knowledge of these direct safety risks, Head of Instagram Adam Mosseri and Meta CEO Mark Zuckerberg subsequently rejected two business cases to create additional youth wellbeing resource. At the time the first business case was rejected, Instagram had a total of only 0.2 FTE working on youth well-being issues, with no staff resource on the product and engineering side.

---

[80] Meta corporate accounts, highlighted in New Mexico vs Meta

[81] Competition and Market Authority (2020) Online Platforms and digital advertising: market study final report

[82] Internal Meta emails published by the US Congress following the Senate Judiciary Hearing in January 2024

[83] ibid

- In his email refusing to support additional investment, Adam Mosseri explicitly referenced commercial and staffing drivers when he wrote: 'I don't see us funding this any time soon, we recently ran a divestments exercise and are still too tight in a number of places.'

- Additional emails show that there was also strong pushback from senior management with direct access to Nick Clegg and Chief Product Officer Chris Cox. In an email sent to both of them in August 2021, they were strongly encouraged not to support additional investment in youth well-being because 'having too much central oversight demotivate local product and research teams.'

- This message also explicitly confirmed that youth well-being was seen as a lesser commercial priority by senior leaders in the company, stating that 'horizontal efforts are hard at Facebook and operationally would only make sense to do for big things e.g. creators, youth [compared to] smaller efforts like well-being.'

Non-designated content

- We strongly support Ofcom's proposed approach for classifying Non-Designated Content (NDC) and agree with its preliminary assessment that depressive content and body image content meet the relevant thresholds for designation.

- In order to be classified as NDC, Ofcom proposes that harm that a set of four conditions must be met: the content must cause harm; the harm must be significant; there is a material risk of harm occurring; and there must be an appreciable number of children affected.

- In respect of depressive content, we consider that each of these conditions have been met. MRF's research finds substantial amount of content on social networks that contains teams of intense depression, hopelessness and misery.

- Content is typically posted using hashtags that are used interchangeably to share or search for suicide and self-harm content, or that are shared by high frequency 'meme' style accounts that readily switch between posting depression, suicide and self-harm material.

- This presents a significant risk of cumulative harm, with relevant depression material likely to be consumed or algorithmically recommended in contexts where the user is being exposed to high-risk content that may incite or have the effect of encouraging acts of self-harm.

- There is increasing evidence that depressive material presents a material adverse effect. As set out earlier in the section, Sonuga-Barke et al[84] establish a harm pathway in which users seek out more negative content that aligns with their mood and/or that acts as a form of maladaptive coping; this makes it more likely they will be algorithmically recommended additional and increasingly high risk forms of content; in turn, this is likely to increase rumination over negative content and reinforce cognitive biases that mean content is likely to be interpreted more negatively; and finally, this may create an excessive need for reassurance or approval, which could push users to engage in higher risk forms of social encounters.

- In developing its approach, we encourage Ofcom to focus its attention on content that features intense themes of depression, hopelessness, misery and despair. There is growing evidence that feelings of isolation, hopelessness, and perceived burdensomeness can exacerbate suicide ideation,[85] and that when users are being recommended or able to 'binge watch' large volumes of posts containing these themes, often alongside or in conjunction with explicit suicide and self-harm content, there is a materially increased risk of them experiencing harmful outcomes.

- Similarly, a recent systematic study found that there was a substantial and statistically significant relationship between feelings of hopelessness and suicide risk among young people aged 18 to 30. Wolford-Clevenger et al find that the relationship between the perception of being a burden, hopelessness, and a sense of belonging is a significant predictor of suicidal ideation. [86]

- This is particularly important in the context of much of the depressive posts that MRF has analysed. Highly depressive content typically features consistent themes of hopelessness, misery and despair, and much of this content is specifically targeted at or posted in the guise of teenage girls.

- In developing its approach, Ofcom should encourage regulated companies to reflect on the types of content that may be harmful; the disproportionate risk of certain user groups being exposed to harmful content; and also the underlying risks posed by platform functionalities and design features, including how non-designated content is likely to be shared, recommended or seen.

- We remind the regulator that some companies have repeatedly argued that non-designated content can be beneficial to some users, and that in adopting this position, have often failed to take necessary steps to restrict the searchability or the volume at which such content is algorithmically shown.

---

[84] Sonuga-Barke, EJS et al (2024) Pathways between digital activity and depressed mood in adolescents: outlining a developmental model integrating risk, reactivity, resilience and reciprocity. Current Opinion in Behavioural Sciences, 58

[85] Bat Tonkus, M et al (2022) The relationship between suicide and hopelessness in young adults aged 18 to 30: a systematic review. Journal of Psychiatric Nursing, 2022, 13 (3),pp253-262

[86] Wolford-Clevenger, C et al (2020) Proximal correlates of suicidal ideation and behaviours: a test of the interpersonal psychological theory of suicide. Suicide Life Threat Behaviours, 50, pp201-210

- While we recognise that positive use cases do apply, this position typically fails to reflect the unintended consequences associated with the consumption and/or recommendation of potentially harmful content. [87]

- In particular, this stance fails to reflect the range of psychological, cognitive or behavioural mechanisms that may contribute towards harmful outcomes, for example assortative relating, descriptive normalisation, emotional dysregulation and reduced aversion.

[87] For example, Nick Clegg fiercely stressed the positive use cases associated with suicide related content on Radio Four's Today Programme when he was shown examples of harmful content from MRF's 'Preventable yet Pervasive 'research

Section 4: Harmful content Code of Practice

- This section focuses on the measures recommended by Ofcom to address the risks of harmful content, captured in its draft Code of Practice.

- Ofcom sets out that its proposed measures are expected 'to make a big difference to children's online experiences' and that its proposals 'form a strong set of foundations to protect children online.'

- We welcome many of Ofcom's proposed measures. The regulator's overall package is a stronger response to suicide and self-harm risks than set out in the previous consultation on illegal content. However, we are also concerned that Ofcom has primarily adopted an atomised rather than systematic approach to the risks of harmful content, with the regulator relying largely on a discrete set of largely ex-post proposals and placing insufficient emphasis on upstream harm reduction.

- As it develops its final proposals, we strongly encourage the regulator to place greater emphasis on safety-by-design and harm reduction outcomes. Ofcom's measures will by definition 'offer a higher standard of protection to children', as required by the Act, but it is deeply questionable whether the overall proposals represent the systemic, risk-based regime that Parliament envisaged when the OSA was passed.

- We remain concerned that the regulator's application of evidentiary thresholds, and its disproportionate focus on the costs to industry of addressing harms for which they should reasonably be held responsible, acts as a substantial constraining influence on the overall ambition and strength of the regulatory regime.

- We note the regulator's intention to hold an additional consultation on the opportunities presented by automated content moderation. Ofcom also references the potential impacts of generative AI, but at this stage makes no specific recommendations to address the risks that may result.

- We have previously set out our concerns that Ofcom's approach to the development and revision of its Codes of Practice risks being unhelpfully gradualist, and that its overall approach therefore risks being unacceptably reactive.

- In the following sections, we set our detailed views on the Codes of Practice and some of our primary concerns about the measures proposed.

Highly effective age assurance

- Section 12(4) specifies that regulated platforms must use highly effective age assurance to prevent children encountering Primary Priority Content on its service. Age assurance is a crucial part of the online safety regime.

- Ofcom's approach focuses on the use of highly effective age assurance measures to determine if a user is under 18, in order to prevent children accessing harmful content and/ or reduce the amount being recommended to them.

- The regulator establishes a set of principles that platforms should use to determine if they have highly effective age assurance arrangements in place. Regulated services will be expected to ensure their chosen age assurance arrangements fulfil each of the criteria of technical accuracy, robustness, reliability and fairness, to ensure these measures are considered highly effective.

- MRF has significant concerns about Ofcom's proposed approach, in particular its decision not to recommend measures in the Code that would require platforms to enforce their minimum user age, where one has been set.

- We are also concerned that Ofcom fails to recommend measures to offer an age-appropriate experience to children of different ages and developmental stages. Section 12(2) of the Act sets out a clear requirement for companies to 'mitigate and manage the risks of harm to children in different age groups'.

- Ofcom claims that this decision is rooted in a lack of independent evidence about the technical capability of current age assurance measures to determine that a user is a child, and to distinguish between child users of different ages, to a highly effective standard.

- In discussions with civil society, Ofcom has claimed that it nonetheless thinks it is incentivising platforms to adopt age assurance processes that will meet a highly effective standard. The regulator states that it envisages being able to take enforcement action against platforms if they fail to uphold their Terms of Service, or where their risk assessment process identifies a risk of under-age children using the service, in cases where a minimum user age has been specified.

- This assessment appears confused. It appears highly challenging to envisage circumstances in which Ofcom could successfully take enforcement action against a platform for failing to introduce highly effective age assurance measures, when the regulator itself has been unable to conclude that such measures yet exist.

- We also question Ofcom's seemingly circular logic in looking to resolve this evidentiary issue. While we understand that the regulator may feel unable to rely on performance data supplied by third-party commercial providers, the regulator has simultaneously

opted not to set out any alternative measures to seek to secure the accuracy, robustness, reliability and fairness that it says it needs.

- As we understand it, Ofcom has no plans to establish a regulatory sandbox, third party audit arrangements or any other proactive approaches to determine that highly effective age assurance processes currently exist.

- We also understand that the regulator has not sought to use its information disclosure powers, despite having been equipped with these powers for over 9 months, to establish evidence about the efficacy of age assurance measures that are already being used by many regulated firms.

- We wish to remind the regulator that the enforcement of minimum user age limits is a fundamental matter of public concern, and there is a reasonable public expectation that the Online Safety Act regime would result in companies actively enforcing their minimum age limits.

- We therefore strongly encourage Ofcom to move quickly to establish a roadmap setting out how it intends to generate the evidence it needs. This should include a commitment to work with age assurance providers to test their solutions in real world conditions, if necessary.

Recommender systems

- We strongly welcome Ofcom's emphasis on the risks associated with personalised content recommender systems.

- As the regulator sets out in Volume 3, recommender systems are a primary mechanism through which user generated content is disseminated across regulated services. In its proposals, the regulator sets out what it describes as a 'precautionary' approach to the use of content algorithms in children's feeds.

- We strongly support Ofcom's focus on the risks associated with cumulative harm, including content that my have the effect of causing harm (even if unintentionally). As MRF research sets out, there are substantial risks associated with how content can be presented in large volumes or in combination with other content categories.[88]

- We particularly welcome the regulator's expansionist reading of the categories of Primary Priority Content compared to the Act. Under Ofcom's proposed definition of Primary Priority Content, material will be considered to meet this threshold if it glamorises, glorifies, romanticises and crucially *normalises* suicide and self-harm behaviour.

88

- Ofcom also states that content can be considered harmful even if it does not intentionally or deliberately intend such an effect. Crucially, this means that substantial amounts of content that have the potential to cause harm, but which are not currently treated as harmful by platform Terms of Service, may now be considered as Primary Priority content.

- However, we have substantial doubts about whether the ambition behind Ofcom's proposals can realistically be delivered. Rather than implement its proposals as an outcome-based set of measures, regulated firms will only be required to restrict or reduce the prominence of harmful content *where this has already been identified* i.e. where content is already undergoing moderation or where there is other relevant information that may suggest a material likelihood that content is gone primary priority content.

- As such, a fundamental safety-by-design measure effectively becomes contingent on the efficacy of other largely ex-post measures, including content moderation. As we set out later in this section, forthcoming MRF research suggests that content moderation arrangements for suicide and self-harm content are currently exceptionally poor[89].

- We also have limited confidence that the measures relating to content moderation outcomes, as set out in chapter 21, can sufficiently incentivise the scale of improvement that is required.

- In Volume 5, Ofcom sets out a number of existing systems and processes that platforms may use to indicate that content is likely to be harmful to children. These include content identification processes such as content classifier and heuristics; user tagging, feedback and reports; and the use of other information gathered through its regulatory undertakings.

- Again, we cannot reasonably see how these measures can be expected to identify the majority – or in all likelihood even a substantial minority - of posts that may reasonably meet the Primary Priority Content and Priority Content thresholds.

- We are also deeply confused by the apparent disconnect between Ofcom's guidance to companies – that they should ensure recommender systems are designed to take a precautionary approach – with the highly prescriptive and restrictive measures that are actually set out.

- MRF is disappointed with the proposed response to certain types of Priority Content, including content that promotes dangerous stunts or incites or encourages users to ingest harmful or poisonous substances. While we recognise that Ofcom's proposals are consistent with the measures set out in the Act, both content types have been linked to significant harm, including a number of child deaths.

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

- We therefore conclude that the recommended approach – that regulated firms should reduce but not wholly cease algorithmically recommending this content to child users - is insufficient. In its final proposals, we encourage the regulator to review the risk profile associated with these content types, and to look again at whether stronger measures can be reasonably introduced.

Content moderation

- Given the design of Ofcom's regulatory regime, content moderation measures will play a particularly important role in tackling the risks of harmful content.

- In the absence of a greater emphasis on safety-by-design measures and given that some of the safety-by-design approaches that are proposed appear contingent on the effectiveness of content moderation arrangements, Ofcom's proposals in this area are intrinsically linked to the delivery of good regulatory outcomes.

- Ofcom recommends a set of seven content moderation measures for user-to-user services. The regulator focuses on a set of best practice standards that seek to establish a baseline approach to identifying harmful content, and that aim to ensure content is actioned in line with regulatory requirements and Terms of Service.

- MRF has substantial concerns about the effectiveness of Ofcom's proposals. The set of recommended measures appear to do little more than capture existing industry best practice, and in many cases will appear to enable regulated services to claim regulatory compliance while making limited if any meaningful changes to their moderation capabilities.

- Forthcoming MRF research highlights significant issues with the quality, consistency and impact of content moderation approaches, and forms the basis our concerns.[90]

- In our analysis of over 12 million content moderation decisions relating to suicide and self-harm, drawing on data from the EU DSA Transparency Database, we find that most Category 1 providers are failing to identify and act on suicide and self-harm content at a level that is commensurate to their likely risk profile.

- The research shows that almost all (98 per cent) of content moderation decisions are taken by just two platforms: Pinterest and TikTok. Instagram and Facebook each account for only 1 per cent of moderation decisions, despite both platforms having a substantial and arguably growing risk profile. X and Snapchat report just 0.13 per cent and 0.04 per cent of decisions respectively.

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

- There are significant inconsistencies and shortcomings in how major platforms respond to suicide and self-harm content across their platforms. For example, Instagram reports only 2 per cent of moderation decisions related to video-based content, despite MRF analysis that Reels is the highest-risk of any of Instagram's product surfaces research.[91] Reels now accounts for 50% of time spent by users on the platform.[92]

- Similarly, platforms appear to respond to suicide and self-harm content unevenly and take highly inconsistent actions if at all.

- For example, while TikTok and Instagram identify and decide on a substantial majority of content on the same day it is posted (94 and 87 per cent of decisions respectively), only two thirds of Facebook's decisions (67 per cent), and less than one fifth of decisions taken by Pinterest (19 per cent), related to content posted on the same day.

- Despite Meta committing in January 2024 to identify and restrict suicide and self-harm content from the feeds of children and young people, neither Instagram nor Facebook had recorded a single example of actually doing this by April 2024.

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

- In light of the inconsistencies set out above, Ofcom's recommended measures appear highly insufficient, and will likely have the effect of 'baking-in' wholly insufficient existing content moderation practices rather than incentivising better.

- While Ofcom's proposals require platforms to design their systems and processes to swiftly take action against harmful content, these proposals focus primarily on the speed rather than the adequacy of the platform response.

- Large and/or multi-risk platforms will be required to set out content moderation policies and set clear performance targets, but in the absence of either outcome-based measures or the adoption or a harm reduction framework that requires continual improvement against relevant specified safety outcomes, it seems difficult to see how these proposals will adequately incentivise platforms to improve upon their current arrangements.

- Our findings highlight the urgent need for Ofcom to develop measures that require platforms to develop content moderation approaches commensurate to their overall risk profile, but also the specific risk profile associated with high-risk functionalities and surface types e.g. video.

---

[91] 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.
[92]Comments made on Meta's Q4 earnings call.

- As it stands, the regulator's proposals seem unlikely to address the asymmetry between where harmful content is located and where it is actually detected. While platforms should be expected to identify and act on the risks associated with high-risk functionalities and features as part of their risk assessment process, appropriate measures should also be reflected in the relevant Codes.

- Put simply, the regulator's Codes must be able to provide confidence that it will address looming inconsistencies in the existing responses of major platforms.

- Our research shows that less than 15 per cent of TikTok's moderation decisions are taken in relation to video and image-based content, despite these clearly being the two major types of content consumed by users. The equivalent figure for Instagram is only 20 per cent.[93] The regulator's Codes must by definition provide capable of addressing this asymmetric response.

- Finally, we are surprised that Ofcom has concluded it is disproportionate to expect small and medium-sized platforms to adopt fundamental quality and safeguarding measures. For example, small and medium-sized firms will not be expected to ensure their staff are appropriately trained, nor even ensure their content moderation function is well-resourced.

- This appears to be a perverse application of proportionality over safety. Parliament clearly envisaged that statutory codes of practice would result in a strong and consistent set of measures, with the regulator appropriately adopting a risk-based response while also ensuring that basic safety and safeguarding measures would be put in place.

- We cannot envisage circumstances where it is disproportionate to expect that a user-to-user service either appropriately trains their staff or ensures there is appropriate safety resource to identify and respond to issues.

Search services

- As Volume 3 correctly sets out, Search services present a high risk of children being exposed to high-risk forms of harmful content, including suicide and self-harm related material. Research suggests that text-based queries via Search services are the most common means via which young people identity and access help-seeking content. [94]

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

[94] Pretorious, C et al (2019) young people's online help seeking and mental health difficulties: Systematic Narrative Review. J Med Internet Research, 21(1)

- Ofcom's research shows that search engines prominently return Primary Priority Content in their results.[95] According to the regulator's data, 1 in 5 of the most prominently displayed search results across all search services Primary Priority Content. These results are particularly concerning given that research suggests a substantial proportion of young people attribute quality based on superficial characteristics such as search result rankings.[96]

- We also note that the risks of being exposed to harmful content were found to be significantly higher across some media types. For example, around half of image searches displayed Primary Priority Content. Results featuring harmful content were typically interspersed with results for help, support and educational content.

- Given the strength of the regulator's own data, we are surprised that Ofcom is proposing seemingly less ambitious measures for Search services than those being recommended for user-to-user providers.

- Ofcom appears to have adopted a divergent approach to that being recommended for U2U services, particularly in respect of age assurance. Whereas user-to-user services will be expected to use highly recommend age assurance measures, search engines will only be expected to identify if 'users are believed to be a child'. Regulated firms will be encouraged to use techniques such as behavioural or content indicators to support their assessment of age, with no explicit requirement to use highly effective age assurance measures.

- Ofcom does not specify how or why it has arrived at this divergent approach, and the regulator provides no assessment of what proportion of children it expects to be correctly identified through this approach.

- Where search services determine that a user is likely to be a child, Ofcom will require these providers to filter out content that is likely to be Primary Priority Content. We support both this approach and the related recommendation that the 'safe search' setting cannot be switched off.

- However, in cases where a user is not believed to be a child, Primary Priority Content will *only* need to be downranked /and or blurred. This means that children who are not correctly identified through content and behavioural indicators will continue to be exposed to harmful content.

- The regulator's approach to Priority Content is unsatisfactory. Ofcom has decided that it should be for the search providers themselves to decide whether to take action on Priority Content and Non-Designated Content. In arriving at this decision, Ofcom states that search services should have regard to the prevalence of designated content, the severity of its potential harm, and the interests of all users (but particularly adults.)

---

[95] 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)

[96] Best, P et al (2016) Seeking help from everyone and no one: conceptualising the online help seeking process among adolescent males., 26 (8), pp1067-77

- It seems impossible to reconcile this approach with the statutory requirement to ensure children receive a higher standard of protection that adults. The regulator explicitly instructs search services to consider the interests of adults over those of children, which is wholly irreconcilable with the objectives of the Act.

- We are concerned that Ofcom's approach to Search services risks actively conflating the 'absence of evidence of risk' with 'evidence of the absence of risk'. Ofcom explicitly sets out that it has recommended a less onerous approach to Priority Content and Non-Designated Content because of insufficient evidence to support a stronger stance.

- In adopting this position, he regulator risks proceeding with an approach that will constrain its ability to recommend appropriate measures in this and future iterations of the Code, and that in the immediate term risks adopting a set of protections that are poorly aligned against a preventative and risk-based approach.

Disconnect between risk analysis and the Code of Practice

- Similarly to Ofcom's illegal harms consultation, the regulator has failed to recommend measures that address the full range of risks set out in its risk register. The regulator correctly identifies a set of high-risk functionalities and features in Volume 3, but then opts not to set out corresponding risk mitigations in Volume 5.

- In respect of a number of high-risk functionalities, Ofcom identifies the risks associated with suicide, self-harm and depression content, but then stops short of recommending measures that can tackle the resulting risks. We are particularly concerned by the lack of measures that target the risks of livestreaming, ephemeral messaging, hashtags, direct messaging and group messaging (including the potential adverse impacts of end-to-end encryption.)

- A number of other high-risk functionalities, including content shares, broadcast channels and account biographies, were not identified at all in Volume 3; and as a result, appropriate risk mitigations for these high-risk functionalities are also missing.

- In discussions with civil society, Ofcom has suggested that regulated companies will still need to address these measures through their risk assessments. While that is true, this logic places undue emphasis on one part of the Act (risk assessments) over another (the child safety duties and Codes.) The Act makes no obvious provision for the regulator's approach.

- The clear expectation of Parliament was that regulated services would need to comply with their risk assessment duties, but also adopt strong systems and processes - directly enforceable through statutory Codes - to address the risks of harmful content faced by young people.

- We remain concerned that the regulator appears to be setting out a post hoc justification for its approach, rather than addressing the fundamental issues of its approach to proportionality, the precautionary principle and evidentiary thresholds, each of which are exerting a significant and unhelpful constraining influence on the development of its regulatory scheme.

- Ofcom has claimed that because the Act is 'technology neutral', it doesn't feel able to adopt a position that states regulated companies should not offer high-risk functionalities, such as livestreaming, until and unless it can be confident the relevant functionality is safe.

- In discussions with civil society, the regulator has somewhat curiously expressed this argument through the lens of children's rights. For example, Ofcom concludes that children actively enjoy using livestreaming functionality, and therefore considers that any approach that seeks to restrict their access to such functionality may well be disproportionate.

- In making this argument, the regulator appears to attach greater primacy to children's rights to free expression and association over their fundamental right to safety. As an online safety regulator, it is unclear on what basis Ofcom has made this balancing assessment.

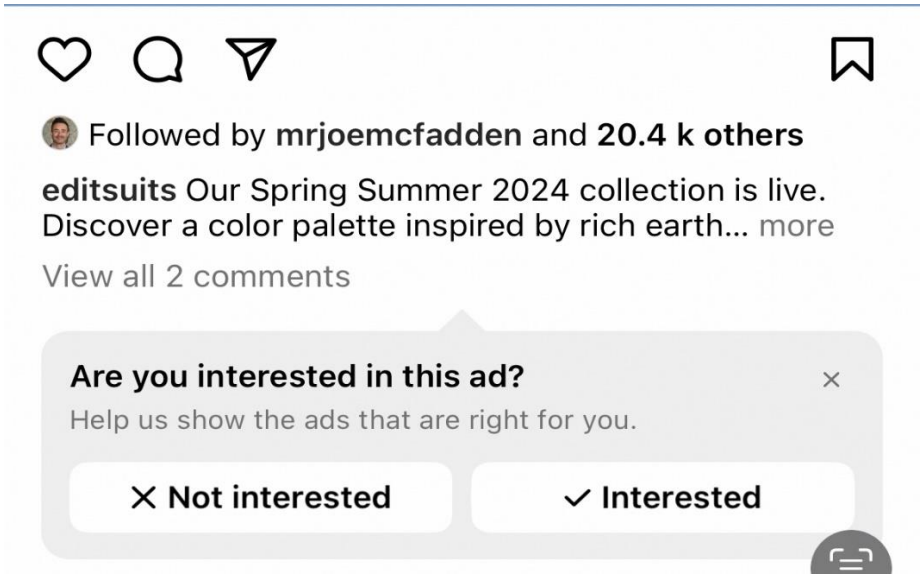
- We also struggle to reconcile this with the positive obligations to protect the physical and emotional safety and wellbeing of young people, and to protect their safety, privacy and dignity, each of which are expressly set out in relevant ECHR case law.

Feedback from child users on algorithmically recommended content

- We strongly welcome Ofcom's proposal that children should be provided with a means of expressing negative feedback on content they have been algorithmically recommended, and that in turn, this should inform what they are subsequently shown.

- If implemented correctly, this measure can effectively incentivise regulated services to provide safer and more responsive content recommender systems. The Meta whistle-blower Arturo Bejar has spoken powerfully about the substantial potential benefits

- As shown below, several Category 1 services have already user-tested real time feedback mechanisms for algorithmically recommended advertising content:



- Under the regulator's proposals, children should be able to express negative sentiment on an individual item of content, and this should result in similar content being limited in prominence.

- Ofcom takes 'similar content' to mean content that share similar characteristics with content about which negative feedback is made. The regulator sets out that significant characteristics may include, but are not limited to:

    - *Subject matter*: the topics, themes or issues were addressed in the content, and;
    - *Metadata*: relevant information about the content such as hashtags, categories and keywords associated with it

- Ofcom expects that regulated firms should provide means for under 18 to offer explicit feedback on recommended content, but that these services should also develop processes to analyse and respond to *implicit feedback*. This includes the number of times a user clicks on an item, the amount of time they spent interacting with it, and behavioural dynamics such as scrolling time.

- Ofcom's ambitions in this area are commendable, but we encourage the regulator to further refine its proposals.

- Firstly, Ofcom should specify a set of criteria against which the suitability and sufficiency of platform feedback mechanisms can be assessed. As a minimum, we would expect the regulator to require platforms to design and test user feedback mechanisms to a high standard, and to subsequently report on the usage and uptake of measures that are rolled out.

- Extensive previous research shows how online services have used 'dark patterns' and other well-established user design techniques to encourage users to take content and consent decisions that may not be in their own best interests.[97]

- Secondly, regulated companies should be expected to demonstrate how they have taken account of user feedback when designing and operating their content recommender systems. Companies could usefully be expected to commission and publish independent user research, with a metric setting out whether under 18s are experiencing more or less harmful or upsetting content in their feeds, in order to offer a useful long-term longitudinal measure of the impact of this measure.

- Thirdly, the regulator should consider additional sources of metadata, including content being recommended by surface type. This data can provide regulated firms with an additional data source setting out the volume and types of content about which young people are providing negative feedback.

- Finally, the regulator should incentivise companies to use this data to develop risk scoring approaches e.g. to identify if a significant number of reports are being made by particular users or category types. This could effectively regulated firms to identify potentially problematic users or content types and to take appropriate actions to restrict, downrank or where necessary remove their content.[98]

[97] Hilton, M (2023) Dark Patterns and User Mental Health: Identifying Theoretical Impacts of Deceptive Design on Vulnerable Demographics. Proceedings of the Human Factors and Ergonomics Society Annual Meeting

[98] In Ofcom's illegal content measures, the regulator declined to recommend risk scoring measures at this stage, citing a lack of evidence about existing approaches. Ofcom should build upon their increased ambition around this measure and incentivise risk scoring mechanisms as part of this approach.