FULL RESPONSE TO OFCOM'S CONSULTATION ON PROTECTING PEOPLE FROM ILLEGAL HARMS ONLINE

Introduction

About the OSA network

The Online Safety Act Network brings together over 60 civil society organisations, campaigners and advocates with an interest in the implementation of the OSA. More details about our work are here.

Structure of our submission

This submission is divided into ten sections, each of them covering a specific issue arising from Ofcom's consultation. We set out in our summary analysis (at pp2-7) how these issues intersect and propose a specific recommendation for Ofcom to address them before the illegal harms codes of practice are issued later in the year.

Sections

Each section is structured in the same way, which we hope provides a consistent approach and will enable Ofcom to make best use of our analysis in its entirety as well as within the individual teams leading on different parts of the consultation:

- Issue
- What the Act says
- Parliamentary debate
- Ofcom's proposals
- Evidence
- Recommendation

We also submit a number of annexes to this response, including:

- Annex A: Analysis of volume 2 functionality risks vs volume 4 mitigation measures
- Annex B: an evidence table, which follows the themes set out in this main paper
- Annex C: A PDF version of the freedom of expression analysis we summarise in section 5
- Annex D: a PDF version of the published analysis of the Illegal Content Judgements Guidance in section 2
- Annex E: a submission on issues relating to Violence Against Women and Girls, summarised in section 9.
- Annex F: a briefing note prepared for Ofcom by Peter Hanley and Gretchen Peters arguing for a product safety assessment approach. This is referenced in section 8.
- Annex G: the statement co-signed by 22 organisations detailing our shared concerns with the consultation

Organisations within our network will be submitting their own individual responses to this consultation. We do not repeat the expert analysis and evidence on their particular areas of interest in our submission but would see much of this as supporting evidence for the broad, structural themes we have focused on here. We would draw Ofcom's attention to the public statement on those themes which was signed by 22 of the organisations and experts in our network and which we submit as an annex for completeness.

February 2024

Contact: Maeve Walsh - maeve@onlinesafetyact.net

Summary of our response

It is to Ofcom's immense credit that this first consultation was produced so quickly after the Online Safety Act received Royal Assent. The protracted passage of the Bill through Parliament undoubtedly afforded much time to prepare in some areas – for example, undertaking calls for evidence and commissioning research, recruiting staff and building up expertise. But it also meant that there were many legislative moving parts, political U-turns and last-minute policy shifts that this consultation has been unable to accommodate. This is not just evident in some of the gaps – many of which Ofcom acknowledges – but in the very different tone and areas of emphasis between some of the initial sections of the consultation (e.g. the overview, approach and background (volume 1) documents) and the detail that follows.

However, in some fundamental areas (such as the differential treatment between large small services, which we discuss in detail below), the approach Ofcom has taken is at odds with the intent of other parts of the final Act; in many, the narrative tone and the decisions do not align with the expectations of Parliament (particularly the House of Lords, where scrutiny of the Bill was exemplary in its detail and its cross-party collaborative approach) or the reassurances offered by the Government in many of the debates, as we demonstrate by reference to Hansard.

Ofcom has stressed – in public stakeholder meetings and in private sessions with our network – that they have had to trade off speed (getting the consultation out) over comprehensiveness (getting everything right first time). The consultation frequently mentions that the draft codes of practice are first iterations and will be updated based on the evidence received in this consultation and as new information emerges. Ofcom's information-gathering powers only came into effect via a commencement order from 10 January and it is clear in statements made by its team and senior management that they see these powers as a route to amassing much more of the evidence they need to fill in the gaps and/or provide more evidence-based measures for further versions of the codes. That will take time.

We fully appreciate the challenges here. However, Ofcom's Chief Executive wrote to Peers at the start of the House of Lords Committee Stage of the Online Safety Bill in April 2023 and reassured them that in relation to the phase one consultations, including on illegal harms, they can "move very quickly here because this part of the Bill has remained unchanged for some time and the illegal harms are defined in existing law. **The Government's and Parliament's intentions about what they want platforms to achieve are clear**. We launched a call for evidence on illegal harms in July 2022, and **are well-advanced in gathering the necessary evidence, including on consumer experiences of those harms, drivers of risk, and the systems and processes available to services to address them." (Our emphasis)**

These reassurances – particularly on the last point – are not borne out by the material produced. We therefore have significant concerns that the approach that Ofcom has chosen to take – and the codes that have been drafted as a result – amounts to a missed opportunity to start the new online safety regime off on a robust footing. In fact, we would go so far as to query whether the approach – taken in the round – that is set out in Ofcom's proposals will even deliver the first step required by Section 1 (1) of the Act, providing for a "new regulatory framework which has the general purpose of **making the use of internet services regulated by this Act safer for individuals** in the United Kingdom".

In short, Ofcom has not been bold enough. Arturo Bejar, the Meta whistleblower who has recently testified to Congress, observed: "Social media companies are not going to start addressing the harm they enable for teenagers on their own. They need to be compelled by regulators and policy makers to be transparent about these harms and what they are doing to address them."

Yet, a scenario where it is justifiable for companies to be left – for quite a while longer – to take their own decisions *even on addressing illegal content* seems to have been accepted within Ofcom, despite the fact that there is now legislation and a regulator with powers to compel them to do otherwise. (See, for example, the comments made by an Ofcom principal at a recent webinar: "Actually, to be candid, for quite a while some of those voluntary principles are going to go further than we're going to be able to go on the codes until we're able to catch up … **It's going to be easier to recommend something as a voluntary principle than it is to have to meet the bar of evidence to codify that in a code of practice**." (WE Communications webinar: Navigating Tech Regulation in the Wake of the Online Safety Act – 31 January 2024; this extract is at 36 minutes in)

Ofcom has made a number of choices in how it is approaching the legislative framework that it has not fully justified and which we argue are not required by the language of the Act; there are inconsistencies between its analysis of the harms it has evidenced and the mitigation measures it proposes; and there are some significant judgements (such as the primacy of costs in its proportionality approach) on which it is not consulting but which fundamentally affect the shape of the proposals that flow from them. Moreover, we are concerned that the framework as proposed at this stage will not be "iterated" in subsequent versions of the codes: the combination of the takedown-focused illegal content judgements guidance and the rules-based, tick-box approach to governance and compliance proposed here will become the baseline for the regime for years to come. The piecemeal basis in which Ofcom has approached the selection of measures contained in the codes – only adding those where there is enough evidence – rather than stepping back to consider the risk-based outcome the legislation compels companies to strive to achieve concerns us. There is a significant risk that the chance to introduce (as Parliament intended) a systemic regulatory approach, rooted in risk assessment and "safety by design" principles will be lost. We think, however, that there is a relatively simple to this issue building on the existing proposals, as we set out in section 1 below. Adopting this approach will allow

Ofcom to bridge the gap between what is evidenced now and our future knowledge base without exposing users to unnecessary risk of harm.

Our analysis

We set out below our analysis of the building blocks of the regime proposed by Ofcom, provide evidence (where it is available) for alternative approaches and recommend specific revisions to the codes of practice that we believe can be made in their first iteration, rather than waiting for a second round of consultations.

We start with analysis of five fundamental issues which run through the whole regime and therefore provide the basis on which many of the specific recommendations are made. These issues are not out for consultation but we would urge Ofcom to consider the choices they have made here and review the impact they are having on the consequential measures recommended and their likely impact. These issues are:

- 1: Weak "safety by design" foundations;
- 2: The approach to the illegal content judgement guidance;
- 3: Decisions on the burden of proof/evidence threshold;
- 4: The approach to "proportionality";
- 5: The approach to human rights.

We then look at a series of specific implementation issues that are a concern and cover some gaps in the final section.

We are grateful to the 60+ organisations, experts and academics in our network for their comments and inputs to the series of discussions which have informed our analysis and to the Ofcom representatives who have met with us bilaterally or as part of larger group discussions. We do not speak on the network's behalf but many of its , have joined us in publishing a joint statement that sets out the key areas where we collectively have concerns. Where we do not have specialist expertise (for example on user verification, CSAM-mitigation measures or the impact of specific offences and illegal harms) we refer Ofcom to other organisations' submissions and analysis, where available.

We are also submitting parts of this written response via the proforma, where they are relevant to the specific questions contained there.

The legislative benchmark

We found it instructive to take Ofcom's proposed outcomes (set out on p7 of "Ofcom's Approach to Implementing the Online Safety Act") as a benchmark for assessing many of these thematic areas. Ofcom says:

> "Specifically, we anticipate implementation of the Act will ensure people in the UK are safer online by delivering four outcomes (Figure 1):
>
> • stronger safety governance in online firms;
> • online services designed and operated with safety in mind;
> • choice for users so they can have meaningful control over their online experiences; and
> • transparency regarding the safety measures services use, and the action Ofcom is taking to improve them, in order to build trust."

While the third and fourth outcomes are more applicable to other parts of the regulatory regime on which Ofcom will be consulting at a later stage (e.g. on the user empowerment duties and the transparency reporting), the first and second outcomes should be fundamental to the approach to illegal content given that this is one of the pillars of the legislation. We do not think Ofcom is providing enough on these first proposals to deliver those outcomes.

In relation to the first and second outcome, it is also worth reading our analysis with reference to the Online Safety Objectives set out in Schedule 4 of the Online Safety Act.

(a) a service should be designed and operated in such a way that—

(i) the systems and processes for regulatory compliance and risk management are effective and proportionate to the kind and size of service,

(ii) the systems and processes are appropriate to deal with the number of users of the service and its user base,

(iii) United Kingdom users (including children) are made aware of, and can understand, the terms of service,

(iv) there are adequate systems and processes to support United Kingdom users,

(v)(in the case of a Category 1 service) users are offered options to increase their control over the content they encounter and the users they interact with,

(vi) the service provides a higher standard of protection for children than for adults,

(vii) the different needs of children at different ages are taken into account,

(viii) there are adequate controls over access to the service by adults, and

(ix) there are adequate controls over access to, and use of, the service by children, taking into account use of the service by, and impact on, children in different age groups;

(b) 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.

As we set out above and in our specific areas of focus, the choices that Ofcom have made in developing their proposals do not align with the overall objectives of the Act, especially the central element of safety by design. The apparent piecemeal approach to each component of this consultation suggests that there has not been a stepping back to look at how these proposals stack up collectively, how they intersect with each other and – crucially – what the impact is of the very many gaps where Ofcom has determined that evidence is insufficient to make recommendations for measures in the codes of practice.

This atomistic approach to the codes creates a structural problem: Ofcom is thinking about adding bits on as and when evidence is available rather than stepping back and thinking about how to approach safe design based on a risk assessment: e.g. how do you make a product or a service so that it orientates itself towards safety? We do not think that it is acceptable to address this by promising further iterations to fill the gaps or to add on more individual pieces to the codes.

We therefore propose, as our key recommendation, that Ofcom add a further requirement to the codes in their first iteration to put a risk-based, outcome-focused requirement on platforms, of all sizes, to put in place a system to identify appropriate measures to address the risks arising from the design and functionality of their service, as identified in their risk assessments, that are proportionate to the size and type of service, bearing in mind best practice and the state of the art. This, we contend, is in line with the parameters of the Act and, just as importantly, justified by Ofcom's collected evidence of harm. It would provide a stop-gap, catch-all measure while Ofcom continues – via its information-gathering powers – to collect evidence on specific measures that work to mitigate harm while, in and of itself, helping to provide some of that evidence via companies' compliance activities.

Issue 1: weak "safety by design" foundations

Issue

At a relatively late stage in the progress of the Online Safety Act, the Government inserted a new "clause 1" which set out the overall objectives of the legislation, including a duty on providers to ensure that services are "safe by design". The Act makes numerous other references to design, as well as "systems and processes" in relation to companies' risk assessment and safety duties. Indeed, Ofcom has adopted as one of its outcomes for implementation that "online services [are] designed and operated with safety in mind".

Much of our analysis in this consultation response is interlinked, providing evidence of the choices that Ofcom has made which – taken together – we believe will not deliver this stated outcome, notwithstanding the fact that these proposals are just one part of a jigsaw that will not be complete for a number of years. We do not cover all the aspects in this section and provide cross-references where appropriate. It is, however, an area where we think a small but significant addition to the codes of practice could address many of the inter-related concerns we have in their first published iteration. We set this out below.

What the Act says

Section 1, which was inserted at Lords report stage, sets out the overall objectives of the legislation – "the general purpose of making the use of internet services regulated by this Act safer for individuals in the United Kingdom" – and specifies at 1(3) that:

> Duties imposed on providers by this Act seek to secure (among other things) that services regulated by this Act are—
>
> a. safe by design

Section 10 (4), which describes the illegal content duties, says that

"The duties set out in subsections (2) and (3) apply across all areas of a service, including the way it is designed, operated and used as well as content present on the service, and (among other things) require the provider of a service to take or use measures in the following areas, **if it is proportionate** to do so—

> (a) regulatory compliance and risk management arrangements,
>
> (b) design of functionalities, algorithms and other features,

(c) policies on terms of use,

(d) policies on user access to the service or to particular content present on the service, including blocking users from accessing the service or particular content,

(e) content moderation, including taking down content,

(f) functionalities allowing users to control the content they encounter,

(g) user support measures, and

(h) staff policies and practices.

Schedule 4 (the Online Safety Objectives) sets out at para 3 how "OFCOM must ensure that measures described in codes of practice are compatible with pursuit of the online safety objectives". At para 4 a), it lists these objectives, which include a number framed as "systems and processes" requirements and at para 4 b) states 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"

Also relevant here is part of the new duties on Ofcom, set out in section 91 which amend Section 3 of the Communications Act 2003, including:

(2) In subsection (2), after paragraph (f) insert—

"(g) the **adequate protection** of citizens from harm presented by content on regulated services, through the **appropriate use by providers of such services of systems and processes designed to reduce the risk of such harm**" (our emphasis)

Parliamentary debate

When the Online Safety Bill reached the House of Lords in February 2023, the opening paragraphs of the first statement at Second Reading from Lord Parkinson said: "This legislation establishes a regulatory regime which has safety at its heart. It is intended to change the mindset of technology companies so that they are forced to consider safety and risk mitigation

when they begin to design their products, rather than as an afterthought." (Hansard 1 February col 687)

On introducing a new "clause 1", Lord Parkinson, said that "Subsection 3 of the proposed new clause outlines the main outcomes that the duties in the Bill seek to secure. It is a fundamental principle of the legislation that the design of services can contribute to the risk of users experiencing harm online. … **I am pleased to confirm that this amendment will state clearly that a main outcome of the legislation is that services must be safe by design. For example, providers must choose and design their functionalities so as to limit the risk of harm to users. I know this is an issue to which we will return later on Report, but I hope this provides reassurance about the Government's intent and the effect of the Bill's framework**." (our emphasis) (Hansard 6 July column 1320)

Later in that debate, Parkinson said again, in response to a proposed amendment from Lord Russell: "He and other noble Lords spoke about the need for safety by design. I can reassure them this is already built into the framework of the Bill, which recognises how functionalities including many of the things mentioned today can increase the risk of harm to users and will encourage the safe design of platforms." (Hansard 6 July col 1382)

The implication of the "by design" approach is that relying on content takedown to mitigate the risk of harm, whether illegal or not, is the end of the road – the harm has already happened; designing better should be the starting point. It also raises the question of whether, where there is evidence of harm connected to particular features, that the obligation should be on the companies should be subject to the burden of rectification – even to the point of rolling back specific features (e.g. push notifications which have given rise to concerns about addiction in the US) until the evidence is there to make them safe enough: product withdrawals are known in other industries.

Ofcom's proposals

Ofcom's Approach document, which also includes its statement of "outcomes", says "Our role is not to instruct firms to remove particular pieces of content or take down specific accounts, nor to investigate individual complaints. Our role is to tackle the root causes of online content that is illegal and harmful for children, by improving the systems and processes that services use to address them. Seeking systemic improvements will reduce risk at scale, rather than focusing on individual instances." (p5). This is heartening – and reflects the Government's intention, as set out in Parkinson's above statement. But it does not flow through to the

volumes that follow this (including the approach to governance and risk assessment, proportionality decisions and the differentiated approach to size) nor to the codes themselves.

Moreover, with specific reference to measures that could be seen as touching on "safety by design" (including written statements of responsibilities or expectations of product testing), Ofcom makes an upfront judgement that these can only be reasonably expected of large or multi-risk companies – thereby undercutting at the outset the overarching legislative objective in the Act. Indeed, as we set out below, the Government's Impact Assessment makes reference to the fact that building in safety-by-design is a way for smaller platforms to reduce regulatory compliance costs.

For example, in volume 3 on governance, Ofcom sets out that a written statement of responsibilities for senior members of staff would:

> "include ownership of decision-making and business activities that are likely to have a material impact on user safety outcomes. Examples include senior-level responsibility for key decisions related to the management of risk on the front, middle and back ends of a service. **This would include decisions related to the design of the parts of a product that users interact with (including how user behaviour / behavioural biases have been taken into account), how data related to user safety is collected and processed, and how humans and machines implement trust and safety policies.** Depending on a service's structure, key responsibilities in online safety may fall under content policy, content design and strategy, data science and analytics, engineering, legal, operations, law enforcement and compliance, product policy, product management or other functions." (Vol 3, 8.64) (our emphasis)

This is comprehensive in terms of the expectations of a well-run company, that places a value on governance and accountability. Even in the smallest companies, responsibilities for these decisions would fall to someone even if the organisational structure did not allow for the formal separation of roles or teams to decide upon or deliver them.

So, it is perhaps surprising that, in the codes of practice, Ofcom only recommends these statements of responsibilities for large or multi-risk services. (Annex 7, Measure 3 C) But the instructive point here is the clear statement that "decision-making and business activities are likely to have a material impact on user safety outcomes". That goes to the heart of safety-by-design especially when the OSA's general mitigation duty (s 10(2) is to deal with "risk of harm to individuals".

Finally, we make a few observations here with regard to the approach to product safety testing

in the consultation proposals. Ofcom make a few brief references to product safety testing, which we would include as a component of an overall "safety by design" approach. In Volume 3, Ofcom says: "Our goal is that services prioritise assessing the risk of harm to users (especially children) and run their operations with user safety in mind. This means putting in place the insight, processes, governance and culture to put online safety at the heart of product and engineering decisions." (Vol 3, 9.8).

Then, in a table suggesting a number of "enhanced inputs" to help companies build up their "risk assessment evidence base", "results of product testing" are included:

> "We use 'product' as an all-encompassing term that includes any functionality, feature, tool, or policy that you provide to users for them to interact with through your service. This includes but is not limited to whole services, individual features, terms and conditions (Ts&Cs), content feeds, react buttons or privacy settings. **By 'testing' we mean services should be considering any potential risks of technical and design choices, and testing the components used as part of their products, before the final product is developed.** We recognise that services, depending on their size, could have different employees responsible for different products and that these products are designed separately from one another" (Table 9.5) (Our emphasis)

This is an "enhanced input": an expectation for larger services only. Ofcom's rationale for this distinction between "core" and "enhanced" inputs is: "All else being equal, we will generally expect services with larger user numbers to be more likely to consult the enhanced inputs (unless they have very few risk factors and the core evidence does not suggest medium or high levels of risk). This is because the potential negative impact of an unidentified (or inaccurately assessed) risk will generally be more significant, so a more comprehensive risk assessment is important. **In addition, larger services are more likely to have the staff, resources, or specialist knowledge and skills to provide the information, and are more likely to be the subject of third-party research**." (Vol 3, 9.113)

This therefore means that not only is product testing to ensure user safety not expected of smaller companies, it is not something that Ofcom feels should be carried out to inform a risk assessment to inform the measures that smaller services might feel they need to take in order to make their products safe. (We set out more on the implications of the differentiated approach to size in Ofcom's proposals in section 7, below.)

This seems to run counter to a "safety by design" approach. It is in marked contrast to the approach of the CMA and the ICO who suggest in a joint paper that testing is key to prevent

harmful design in choice architecture; the paper notes that there are different ways of testing. The resources available to a service provider could thus inform the sort of testing rather than the question of whether service providers should test.

Evidence

Safety by design

In 2021, shortly before the draft Online Safety Bill went through pre-legislative scrutiny, the Government published a series of guides setting out the principles of "safety by design" for online platforms. It introduced them by stating:

> "Safety by design is the process of designing an online platform to reduce the risk of harm to those who use it. Safety by design is preventative. It considers user safety throughout the development of a service, rather than in response to harms that have occurred.
>
> The government has emphasised the importance of a safety by design approach to tackle online harms. The government's response to the Online Harms White Paper highlighted the importance of a preventative approach to tackling online safety, including through safer platform design. In response to this, the government committed to publishing guidance to help UK businesses and organisations design safer online platforms.
>
> By considering your users' safety throughout design and development, you will be more able to embed a culture of safety into your service."

Ofcom makes no reference to this work in its risk profile evidence (volume 2) or in the "best practice" that it uses to inform the measures recommended in the codes of practice. We set out for reference the "best practice" guides that the Government produced:

- private or public channels
- live streaming
- anonymous or multiple accounts
- search functionality
- visible account details or activity

The omission of these principles by Ofcom is even more surprising given that – in the

Government's own Impact Assessment, it made reference to them as a means by which smaller businesses could reduce the costs of regulatory compliance and get ahead of the introduction of the legislation:

> **"**While per business costs are expected to be higher for medium and large businesses, it is important to consider the possibility that some in-scope SMBs will have limited resources for compliance. To minimise burdens on SMBs, it will be vital for Ofcom to work with businesses and to ensure both requirements and enforcement are proportionate to the risk of harm and resources available to businesses. Proportionality in the context of effective safety measures must be balanced against the risk of harmful content being displaced to smaller and less well-equipped platforms. The government and Ofcom will work with SMBs to ensure that steps taken are effective in both reducing harms and minimising compliance costs. **The government's Safety by Design framework and guidance is targeted at SMBs to help them design in user-safety to their online services and products from the start thereby minimising compliance costs.**" (our emphasis)

Relatedly, as we set out in this blog post, Ofcom has made no reference to other "by design" approaches such as its own "media literacy by design principles" on which it was consulting in parallel with the early stages of this consultation.

There are other examples, from which Ofcom could draw to support a general "safety by design" measure. The Australian e-Safety Commissioner produced Safety By Design principles – developed in conjunction with industry and adopted by the World Economic Forum – and introduced them in the following terms:

> "Rather than retrofitting safeguards after an issue has occurred, Safety by Design focuses on the ways technology companies can minimise online threats by anticipating, detecting and eliminating online harms before they occur. This proactive and preventative approach focuses on embedding safety into the culture and leadership of an organisation. It emphasises accountability and aims to foster more positive, civil and rewarding online experiences for everyone."

We note that the Australian e-safety Commissioner provided evidence on safety by design for Ofcom's call on the illegal content duties but this is only referenced, briefly, twice in volume 4.

In addition to the joint CMA and the ICO work on choice architecture, the CMA has referred to online choice architecture/nudges in relation to competition and consumer harm.

The National Cyber Security Centre has set out a series of "cyber security design principles" that focus on red-teaming design processes (described here) as a means to pre-empt problems. The Ministry of Defence has also produced a handbook on red-teams

Beyond regulators and government bodies, IBM has looked at technology design principles that would address domestic violence and work has been done on abusability testing frameworks described here, with examples set out here and here.

NGOs and other researchers have looked in depth at the specifics of platform design in relation to a wide range of harms, such as CCDH's work on TikTok and eating disorder and self-harm content; ISD's research into how TikTok's functionality spread hate and extremism; Amnesty's work on TikTok also illustrate that its recommender system amplifies depressive and suicidal content. The prevalence of anxiety-inducing health-related information has been seen on other platforms, along with the finding that user response tools are designed to provide relatively weak signal to the recommendation systems.

Harmful design

We set out elsewhere and in annex A the commendable work Ofcom has done to bring together in volume 2 evidence on how functionalities on services can contribute to a series of harms and have provided references above to how a "safety by design" approach might surface and help address some of those harmful design choices. We would also refer Ofcom to the series of recent US court filings and whistleblower reports that have recently laid out what happens when a "safety by design" approach is not embedded in companies' culture and the impact of platforms' design choices on the harms that are caused to users, particularly children. What is relevant here is that these documents also demonstrate platforms' awareness – over a number of years – of the harms that are being caused by design and their decision to implement features notwithstanding this awareness and their apparent unwillingness to redesign their services to prevent them; this is the exact opposite of safety by design. In the UK, coroners' reports have also identified where platform design has had a direct role in creating the conditions in which individuals have decided to take their own lives.

We list some of these documents here for Ofcom's reference and would recommend that these are reviewed as part of their evidence base, not just for application to the measures recommended for addressing illegal content but for the development of the proposals for the children's codes.

We are concerned that – while Ofcom knows that functionalities and service design choices cause harm – it has chosen not to place the responsibility for mitigating the risks of those functionalities and design choices on the regulated services. Instead, by deciding that it must judge for itself whether measures that work to reduce the harm of those design choices (in effect, ex post rather than ex ante measures), it has limited the imperative on regulated services to use the evidence that they already have (whether formally collected or not) on how their services are designed as a means to inform suitable measures to mitigate the risk of harm occurring as a result.

US court filings

- New Mexico Attorney-General case against Meta - January 2024
- Bad Experience and Encounters Framework (BEEF) survey - Instagram internal research - unsealed as part of New Mexico court case - January 2024
- California Superior Court Opinion re dismissal of Fentanyl Case re Snap - January 2024
- Multistate Complaint re Meta - largely unredacted - Nov 2023
- Second amended complaint re Fentanyl and Snap - July 2023
- California Master Complaint in re Adolescent Social Media Addiction - May 2023
- Class action against Tinder et al – February 2024

Whistleblower material

- Arturo Bejar testimony to Congress - November 2023
- Sophie Zhang oral evidence to Parliament & written evidence- October 2021
- Frances Haugen evidence to Congress & transcript - October 2021
- FB Archive - searchable repository of the Frances Haugen papers

Coroners' reports

- Prevention of Future Death Report: Chloe McDermott - December 2023
- Prevention of Future Death Report: Bronwen Morgan - November 2023
- Prevention of Future Death Report: Luke Ashton - July 2023
- Prevention of Future Death Report: Molly Russell - October 2022
- Prevention of Future Death Report: Joseph Nihill – September 2020
- Prevention of Future Death Report: Callie Lewis - December 2019

Some extracts are included here as examples as to the types of features that a "safety by design" requirement would capture, placing the responsibility on platforms (as the Act requires) to mitigate the risk of harm arising as a result of those design choices.

For example, from the New Mexico Attorney General filings:

> "The harms laid out in the complaint are tied to Meta's actions, failures, and design decisions, including, but not limited to: (i) implementing design features and policy choices that fail to ascertain or apply the actual age of users; (ii) preventing effective parental controls and reporting mechanisms; (iii) permitting predators to identify, contact, and groom children and to develop CSAM through these contacts; (iv) designing algorithms that serve up child sex exploitation content to children and to predators; (v) failing to warn and affirmatively misleading parents and children about the presence of young children and about sex trafficking and sexual exploitation content on the platforms; (vi) failing to identify and report CSAM; and (vii) creating and sending harmful notifications that encourage addictive use of its platforms. **Correcting these activities does not require Meta to edit or withdraw third-party content, but rather to design its product differently—namely, safely—and describe it honestly**" (our emphasis)

The lack of safety testing – and the awareness that this is a problem – is evident in many of the other US court filings and whistleblower materials. For example, again from the New Mexico Attorney General filings:

> "Meta launched Reels in order to attract teens who were transitioning to competitors, like TikTok, that already featured a video service. Internal Meta documents confirm that the launch of Reels was rushed in order to preserve engagement among Meta's teen users. One employee noted in a 2020 message: "The fact that we're shipping reels without a clear picture of the ecosystem impact is pretty mind boggling." Another employee echoed that sentiment: "it is scary the speed we are moving . . . we either do things WAY TOO FAST without Data. Or do things WAY TO[O] SLOW because of Design/Principles." These product designers were aware of the harm that could result from Reels, with one stating "I am worried that the cumulative effects are going to be bad."" (p163)

Additionally, Section XIV of the New Mexico case is entitled "META WAS ACUTELY AWARE OF THE HARM TO YOUTH WELL-BEING RESULTING FROM ITS DESIGN CHOICES, BUT FAILED TO DEVOTE SUFFICIENT RESOURCES TO ADEQUATELY ADDRESS THE HARM TO YOUTH" (p168 onwards) and begins:

> "At the same time that Meta was making these design choices, internal documents confirm that Meta was aware of the harmful effects that its products were having on the wellbeing of children and teenagers. Meta performed numerous studies and analyses concerning teen usage and the effects resulting therefrom, but systematically ignored internal red flags in favor of chasing profits." While this section of the filings relates to harms to children's mental health and wellbeing, the evidence of the platforms' awareness of how their systems and processes cause harm apply to the discussion on illegal harms – and indeed the earlier section of the New Mexico report on CSAM (extract above) clearly sets out some of the failings in this regard of which, again, the platforms are likely to be aware.

We would also refer Ofcom here to the important evidence in the recent report from Revealing Reality on Snapchat which suggests that its "design features not only enable the sharing of unpleasant and illegal material, but in some cases shape the behaviour that leads to its creation".

> "No one claims Snapchat set out to facilitate criminality or harm, or deliberately designed its platform to encourage children to share CSAM or to film fights. But, as we've seen, Snapchat's features and functions make it possible. At the same time, as we've seen, putting the onus on these children to report unsuitable content is not realistic – they won't. So what is the answer? Just like any product, features and functions can be changed. Moderation can be increased. Vulnerable children's experiences don't have to be this way. Design choices are just that – choices."

Recommendation

Supported by the evidence and analysis we provide in this and subsequent sections, **we recommend that Ofcom makes a small but significant change to its draft codes of practice before they are published in their final form later this year**. This would put a requirement on all regulated companies specifically to take measures to address harms that have been flagged in their risk assessment that arise from the features and functionalities of their service, drawing on current good practice, and to regularly monitor the measures' effectiveness. (Current good practice could include interventions that Ofcom has discussed but for which the evidence base is missing at the moment.) This provides an interim step, in the absence of the evidence Ofcom feels it requires to recommend specific measures, that would go a long way to ensuring that the regulatory regime begins on the right footing and starts, from the outset, delivering the "safety by design" intent of the Act and the general mitigation duty at section 10 2(c) for user-to-user services and 27 2 for search.

We also recommend that product testing should be included in the codes of practice, appropriate to the size of the company and the risks its products pose, and that the results of this testing should be a core input to the risk assessment.

We suggest the following wording is inserted in the draft codes, between the section on governance and accountability and the section on content moderation, which follows the order of areas in which measures should be taken identified in section 10 (4) and section 27 (4) of the Act.

> "Design of functionalities, algorithms and other features
>
> *Product testing*
>
> For all services, suitable and sufficient product testing should be carried out during the design and development of functionalities, algorithms and other features to identify whether those features are likely to contribute to the risk of harm arising from illegal content on the service.
>
> The results of this product testing should form a core input to all services risk assessments.
>
> *Mitigating measures*
>
> For all services, measures to respond to the risks identified in the risk assessment should be taken, including but not limited to, providing extra tools and functionalities, by redesigning the features associate with the risks, by limiting access to them where appropriate or where the risk of harm is sufficiently severe by withdrawing the function, algorithm or other feature.
>
> Decisions taken on mitigating measures, as part of the product design process or as a response to issues arising from the risk assessment, should be recorded. (Note: this would be included in the record keeping duties under section 23 (u2U) and section 34 (search).)
>
> *Monitoring and measurement*
>
> All services should develop appropriate metrics to measure the effectiveness of the mitigating measures taken in reducing the risk of harm identified in the risk assessment. These measures should feed back into the risk assessment."

Issue 2: the approach to the illegal content judgements guidance

Issue

The safety by design approach is central to the regime and should influence the implementation of both the illegal content safety duties and the children's safety duties, on which Ofcom will be consulting in phase 2 later this year. The illegal harms consultation, as the first component in the regime, should provide the framework on which these further consultations can build.

Yet, the guidance focuses primarily on individual items of content and assessing whether they should be taken down – it even refers in the draft Guidance to the obligation being "to take content down" (Annex 10, A1.14), rather than, as s 10(3) says, to operate a proportionate system designed to have that effect. While there are parts of the consultation which reflect the obligation correctly - for example, in the "Overview" document where Ofcom says "A new legal requirement of the Act is for all services to swiftly take down specific illegal content when they become aware of it" – the Act's systemic language is generally ignored in the draft guidance itself. Choices about design happen before you get the content flowing across them. There is also no real consideration of scale - the sheer volume of information that is potentially involved. This then defines the scope of Ofcom's overall illegal harms approach, with a focus on ex-post measures, such as content moderation and take down, which we discuss in more detail below.

Furthermore, by requiring that a criminal offence has taken place each time content is posted (rather than acknowledging that content which has been deemed illegal remains illegal when shared as it is still connected with the original offence), an unnecessarily limited view of relevant content is baked into the proposals compounded by an approach that sets the standard of proof at a high threshold – in some instances close to the criminal level – at odds with what is a civil regulatory regime. Again, this approach does not sit well with a systems-based approach. Moreover, this is especially problematic given that some criminal offences operate to protect individuals' fundamental rights; the rights balance here is, again, one-sided (see more general discussion here and in section 5 below and attached as a PDF). It is also unfortunate that Ofcom has not considered any of the existing non-priority offences, specifically s 127(1) of the Communications Act, which (unlike 127(2) Communications Act) has not been repealed.

We have published a detailed analysis on this issue by Prof Lorna Woods and include extracts from it below.

What the Act says

Section 59 defines "illegal content", specifying content the use of which "amounts to" or the possession, viewing, publication or dissemination of which "constitutes" a relevant offence. A relevant offence is either a priority offence (listed in Schedules 5, 6 and 7) or one that satisfies the criteria in s 59(5).

Given that content on its own is not a criminal offence, but rather certain behaviours linked to the content, and the difficulties of identifying the other elements of the offence, section 192 sets out "Providers' judgements about the status of content" and Section 193 requires Ofcom to produce Guidance on it.

Section 192 (1) identifies the purpose of the section, which is to set out the approach to be taken where:

> (a) a system or process operated or used by a provider of a Part 3 service for the purpose of compliance with relevant requirements,
>
> (b) a risk assessment required to be carried out by Part 3, or
>
> (c) an assessment required to be carried out by section 14,
>
> involves a judgement by a provider about whether content is content of a particular kind.

Section 192(5) specifies that a provider "must have reasonable grounds to infer that content is content of the kind in question", based on, according to s 192(2), "all relevant information that is reasonably available to a provider".

The provision relating to illegal content judgements was added to the Bill in July 2022 at Commons Report stage.

Parliamentary debate

In response, to concerns raised by Lord Allan (here) Lord Parkinson commented:

> I know that the noble Lord is concerned that this provision could encourage overzealous removal of content, but the Government are clear that the approach that I have just outlined provides the necessary safeguards against platforms over-removing content when complying with their duties under the Bill. The noble Lord asked for a different standard to be associated with different types of criminal offence. That is, in effect, what we have done through the distinction that we have made between priority and non-priority offences.

To assist services further, Ofcom will be required to provide guidance on how it judges the illegality of content. In addition, the Government consider that it would not be right to weaken the test for illegal content by diluting the content moderation provisions in the way that this amendment would. Content moderation is critical to protecting users from illegal content and fraudulent advertisements. (Lords Committee debate 17 July 2023)

Ofcom's proposals

Volume 5 of the illegal harms consultation sets out their high-level approach to the Guidance ("How to judge if content is illegal or not?") and the draft Guidance itself is in annex 10 ("Online safety guidance on judgement for illegal content")

Safety by Design

The safety-by-design approach is central to the regime (s 1(3)) and should influence the implementation of both the illegal content safety duties and the children's safety duties; Ofcom will be consulting on the latter in phase 2 later this year. As set out in section 10, there are a range of duties applying to illegal content (notably a general duty to mitigate) and some further duties applying to priority illegal content. These could be ex ante measures – for example, design choices (e.g. increased friction; approach to weighting of recommendation tools and revenue sharing policies), proactive measures (e.g. chatbot interventions to reduce racist messages) or ex post measures such as content curation and/or moderation (e.g. systems for downranking, takedown or account suspension). All the OSA duties relate to the design of the service (broadly understood) or its operation, and not to individual items of content (ss 10 and 27). **This is not fully reflected in Ofcom's discussion of the meaning of illegal content in Volume 5 or the approach taken in the draft Guidance (Annex 10) – even when taking into account the constraints of the Act.**

Ofcom's discussion focuses on individual items of content – to the point of saying in the draft Guidance that the obligation is to take content down (Annex 10, para A1.14), rather than - for user-to-user services - to operate a proportionate system designed to have that effect. It is unclear in the main how this becomes relevant to search (e.g. auto-complete suggestions, personalisation). Of course, the operation of a content moderation system does imply the application of the rules to individual items of content, but that is not the primary obligation in the Act. Lord Parkinson made this clear at Lords Report stage in the discussion on this clause.

> "My Lords, I start by saying that accurate systems and processes for content moderation are crucial to the workability of this Bill and keeping users safe from harm." (Hansard July 17 2023 col 2141)
>
> And:
>
> "platforms will not be penalised for making the wrong calls on pieces of illegal content. Ofcom will instead make its judgements on the systems and processes that platforms have in place when making these decisions." (Hansard July 17 2023 col 2143)

Earlier in the Lords, at Committee stage, Lord Parkinson also said:

> "To be clear, the duty requires platforms to put in place proportionate systems and processes designed to prevent users encountering content. I draw my noble friend's attention to the focus on systems and processes in that. This requires platforms to design their services to achieve the outcome of preventing users encountering such content. That could include upstream design measures, as well as content identification measures, once content appears on a service." (Hansard 27 April 2023 col 1359)

Ofcom could have chosen a different approach – one which fits with the systems approach – within the terms of the Act. The definition of illegal content in section 59 does not specify whether the requirements of the offence are to be defined in the abstract or in the applied context. When considering the application of the rules in a system (e.g. in the take-down context where an individual item of content is in issue) it might be relevant to assess how the behaviours required by the criminal offence may map on to those of the user. When we are looking at the design of the systems and processes, however, considering the offence more generically makes more sense. The effect of the Act being drafted in systems language means not only that understandings of illegal content must be considered at scale but also before those items of content have come into being. The recognition in the draft Guidance that services may be dealing with content "in bulk" (Annex 10, para A1.15) is not quite the same point.

Moreover, the word "content" is ambiguous. In line with usual statutory interpretation, the singular includes the plural, but could also extend to types of content. Section 192(4)(a) distinguishes between content and kinds of content suggesting that both approaches should be covered. The Guidance has considered when inferences could be made, focussing on an item-by-item approach. It should also consider what the signals for inference (the "reasonable grounds" in s 192(5) and (6)) about the mental element of crimes under s 192 are in relation to systems design. These cannot be the same for systems as in individual items of content (where

Ofcom suggests that decisions should be made on a case-by-case basis (Vol 5, para 26.24, 26.82) – which would in any event be hard to scale, even ex post).

Significantly, the Guidance emphasises that inference is based on the substance of individual items of content in relation to all priority offences. While this is a starting point, the service could have other sources of information on which to base its judgement. While a broader range of sources of information are noted in the Guidance (Annex 10, A1.66 and for examples see A6.11, A6.24 and A 6.37 in re fraud), these are not consistently considered. Contextual information is important when determining categories of content – for example, patterns of posting (e.g. frequency and timing of posts could be relevant for understanding harassment; cf Annex 10, A3.100-102), the existence of networks in addition to the content-based context of what was in the post before or after the impugned item (which Ofcom notes in some instances).

Of course, the precise significance of different types of contextual information may vary between offence type (for example, there is - as Ofcom notes - less to understand in the context of CSAM than threats). Nonetheless, Ofcom's focus in the Guidance seems to be on the context of the content itself (see e.g. examples in Annex 10, A 2.17, A6.65) rather than a wider range of metadata context information which may be of particular use when identifying categories of content and designing systems.

There is a significant gap in the Guidance here. While Ofcom rightly notes that in practice user-to user services' terms of service cover more content than that which would be defined as illegal content (on relationship between illegal content judgements and terms of service see 26.17 – 26.18 and 26.43 – and note that search engines do not have to have terms of service), the definition of illegal content is important beyond providing a floor for those terms of service. **Illegal content defines the scope of the regime as regards the illegal content duties.**

Volume 5 and the draft Guidance focus – as do Volume 4 and the draft Codes – on ex post measures, specifically takedown (see e.g. para 26.43 and in the draft Guidance that search can only ever look at individual items of content – Annex 10, para A1.16). They do not consider ex ante measures (which are not limited to proactive technologies within the meaning of s 231) nor safety by design, and so do not consider how measures which are not targeted to particular content but are aimed at generally removing risk/improving safety (and thereby impact across a range of harms caused by different types of illegal content) will fit in. For example, recalibrating the weighting on recommender tools (perhaps in line with the adaptation of such tools under the DSA as suggested in recitals 87-89) or taking steps to deal with data voids.

These questions are important as it is on the design/operation of system that the service can satisfy its duties and not on the taking down (or not taking down) of specific items of content. So while the draft Guidance covers some of the ground, Ofcom should consider how to understand content by reference to systems, and make clear that the Guidance, as is, is not exhaustive in that regard.

Content not Conduct

The illegal content safety duties are triggered by content linked to a criminal offence, not by a requirement that a criminal offence has taken place. Indeed, the Consultation states that it is not the purpose of the regime to make decisions on whether a criminal offence has taken place. The requirement for reasonable grounds to infer a criminal offence each time content is posted, as outlined in Vol 5 (para 26.44 et seq) and the draft Guidance, presents an overly restrictive interpretation of relevant content. Such a narrow perspective is not mandated by the language of section 59, which necessitates the existence of a link at some stage, rather than in relation to each individual user. The significance of this can be seen in the example given of the reposting of intimate images without consent – the re-post is still the content linked to the original offence, it has not changed its nature. Contrary to the views expressed in Annex 10, para A1.59, there is a difference between the same content and altered content. There is no obligation in the Act to look at the mental state of each individual disseminator of the content. Moreover, this point needs to be understood against those made about the systems obligations, when design choices are made in relation to types of content rather than specific items.

Burden of Proof in a Civil Regime

The Act introduces a civil regime not a criminal one. The Consultation recognises that "'Reasonable grounds to infer' is not a criminal threshold", and further notes that this test is the relevant test rather than beyond reasonable doubt (see Vol 5, para 26.14). Given that the regime is a civil regime this threshold should be understood against the civil burden of proof – that is on the balance of probabilities. This means the threshold for proof is lower both as to the types of evidence considered to give rise to an inference and the amount of evidence required. There is also the question of how to approach inference in the absence of evidence being reasonably available (which is permitted – it might even be required – by the Act) – to what extent (especially with content implicated in more serious crimes) should such an inference be made; it may be that there is greater scope for some offences (e.g. those with low mental element thresholds) than where there are more specific requirements.

In this, we should note that there is a wider body of potentially relevant information – the wider context Lord Parkinson referred to. So, for example, where we have evidence about widespread negative impact of a behaviour (e.g. cyber-flashing), with little in the way of countervailing interests (contrast for example the difficulties around the suicide offences), we could infer that the mental element was met. Evidence, albeit limited, indicates that a proportion of men know that the images cause distress. Moreover, Ofcom's own evidence gathering, set out in Volume 2, says that "Cyberflashing is not a product of technology and online behaviour alone; it is a manifestation of existing patterns of sexual violence and abuse. McGlynn argues that cyberflashing should be understood as part of a continuum of sexual violence. As with all forms of sexual violence, perpetrators of this abuse are motived by a desire to exert power, and victims and survivors experience feelings of fright and vulnerability." (Vol 2, 6S.19).

Given this understanding of the nature, extent and severity of the harm, is it not reasonable to infer on the balance of probabilities that the content is linked to criminal behaviour? In the context of articles used for fraud, Ofcom proposed "when considering the user's state of mind, services should ask themselves whether there is any possible use of the article concerned which is not for fraud" (Annex 10, A6.66). Yet for cyberflashing Ofcom suggests – without explaining why – that it would be hard to infer the mental element (Annex 10, A 10.43). The approach Ofcom has taken here is unnecessarily restrictive – especially as Ofcom has in relation to terrorism suggested that the threshold of recklessness is reasonably easy to infer (Annex 10 A2.55, A 2.69). As a consequence, Ofcom fails to deal with the harms caused by cyberflashing. It also raises the question as to whether the standards of proof in the Guidance are consistently those of the civil regime, or whether in some instances, a narrower approach has been adopted.

Moreover, the difficulties in these areas are compounded because Ofcom has considered inference in the face of a lack of evidence in respect of a moderator on a case-by-case basis (Vol 5, para 26.82). As noted, signals on which inferences may be made, may need to be understood differently than in the context of a case-by-case analysis.

Missing Offences

Ofcom is right in asserting that identifying the most serious or most specific priority offence is not the most effective way to think about how the regime works; for the purposes of the regime, it is sufficient if any priority offence is triggered and so the broader priority offences are the most significant when it comes to triggering the regime. So, when an offence (and the consultation gives the example of racial hatred) is committed, for the purposes of applicability

of the illegal content duties and enforcement it does not matter whether it is the aggravated offence or the base offence.

Against this recognition, it is unfortunate that Ofcom has not considered any of the existing non-priority offences, specifically s 127(1) Communications or the Obscene Publications Act 1959 (listed as priority in Sch 6 in relation to those offences only). Much content falling out of more specific offences will be caught by the Obscene Publications Act or by s 127(1), and therefore some safety duties would apply, notably the base level of mitigation (s 10(2)(c)) and having a system to take content down (s 10(3)(b)). The existence of these offences should be flagged so that they are not forgotten or overlooked, especially as Ofcom has suggested it is not proportionate for providers to anticipate all non-priority offences (Vol 5, para 26.70) and that (in relation to terrorism offences) the giving of guidance in relation to some offences and not others is to suggest to providers where they should focus their attention (Vol 5, para 26.64). This approach makes sense where an offence is unlikely to occur; much less so where there are offences which are quite likely to be relevant, as is the case with the two offences here. Moreover, the selection of the non-priority offences in respect of which guidance is given is not based on the likelihood of them being relevant, but on their newness (Vol 5, para 26.72).

The Impact of Rights

It should be noted that Ofcom has an obligation to take into account fundamental rights, noted in para 26.8 and reflecting the requirements of the Act, and this has weighed towards a narrow interpretation of illegal content. However, the terms of the act cannot remove the Ofcom's obligations in relation to other fundamental rights. While intrusions into Article 10 must be carefully considered, three counter points should be noted.

First, the speakers are *not* being *criminalised* by the application of the regime – this means there is a lesser intrusion into their rights than there would be were criminal penalties to be imposed. Even the takedown of content for legitimate reasons is a more proportionate response than the imposition of a criminal penalty. Indeed, takedown has been found to be a proportionate response in relation to civil actions; account removal – which has a greater impact on the user's speech rights - has in the case of persistent violation, been found appropriate in a regulatory regime (see *NIT S.R.L. v. the Republic of Moldova* (28470/12)).

Secondly, survivors of online harms have rights too, which should be considered, as is discussed here This Ofcom has completely failed to do – especially as regards the well-documented silencing effect some content has on others, especially those in minoritised groups. (See analysis on Ofcom's approach to human rights in the illegal harms consultation here.)

Further, while this point may be implicit in some of Ofcom's analysis, it should be expressly recognised that some content is likely to be more worthy of protection than others – and that this affects the impact of freedom of expression concerns on scope of offence. While it is possible that abusive speech could in some instances be considered to be political speech which attracts significant protection from Article 10 (see e.g. *In re S (FC) (a child)* [2004] UKHL7, albeit in the context of a civil law claim), it is hard to think that sharing deepfake porn would do so (even though it formally falls within Article 10).

Evidence

We set out some extracts from the recent US court filings below which have relevance to this particular topic.

For example, the issue of scale is demonstrated in the New Mexico Attorney-General US filings;

> "Meta knew about the huge volume of inappropriate content being shared between adults and minors they do not know; a 2021 presentation estimated 100,000 children per day received online sexual harassment, such as pictures of adult genitalia" (p95)

The problem with not addressing design is also evident in at p101 here:

> "[Meta's] algorithms can readily detect and recommend users groups or users with attributes similar to those a user already selected, but that same computing power does not identify illegal material appearing on the website, and, instead, compounds the problem by directing users to additional illegal material that should have been removed from the site in the first place."

And here:

> "Instagram is well aware that users on its site post, distribute and advertise CSAM. When a user searches using known CSAM keywords, Instagram displays "an interstitial alerting the user of potential CSAM content in the results." That warning reads: "These results may contain images of child sexual abuse. Child sexual abuse or viewing sexual imagery of children can lead to imprisonment and other severe personal consequences. This abuse causes extreme harm to children and searching and viewing such materials adds to that harm. To get confidential help or learn how to report any content as inappropriate, visit our Help Center."29 192. However, even though that warning acknowledges the illegality and harm stemming from searches and displays of "child sexual abuse" imagery, Instagram nevertheless permits users to view the material by including a link entitled "See results anyway" at the bottom of the warning. A user who

> clicks on “See results anyway” is taken to the very content that Instagram warns and knows is forbidden and/or harmful, thereby rendering its “warning” largely ineffective”

The recent testimony from Meta whistleblower Arturo Bejar is relevant to Ofcom’s overall approach:

> “Underlying this approach is the belief that in order to reduce distressing experiences for people, the most important area social media companies should work on is social norms. The current approach, based on setting legalistic definitions within policies and reactively removing content, is not sufficient. More importantly, it does not address the majority of the distressing experiences people face. What must guide the design of features to make people feel safe with each other in social media should be the actual experience of users.”

Recommendations

While the draft Guidance covers some of the ground, Ofcom should consider how to understand content by reference to systems, and make clear that the Guidance, as is, is not exhaustive in that regard. This would reinforce the improved focus on safety by design that we are also recommending (see section 1, above, including our specific recommendation for new measures to be added to the draft codes).

We would urge Ofcom to review the approach it has proposed in the light of the analysis above. We recommend that Ofcom consider how to revise the Guidance before it is published to address the risks that the current focus on a piece-by-piece approach to content will have for the effectiveness of the regulatory regime as a whole. At a minimum, an additional focus on the application of systemic and by-design measures – as provided for in the Act - should be added to the Guidance to ensure providers can apply it at scale. In addition, we would recommend a specific requirement to record and monitor the approach to illegal content judgments where they are made to verify if there are problems with the signals being used.

Issue 3: decisions on the burden of proof/evidence threshold

Issue

Much store is set in the consultation document narratives by the amount of evidence already collected to support the proposals e.g. the risk management approach, and on the "best practice" already provided by platforms to justify the approach. Conversely, where there is weak or limited evidence relating to the potential for a particular measure to address a particular outcome, this is given as a reason not to include it within the codes until more evidence becomes available (though this approach is not required by the Act). (See section 6 on measures and the codes below.) This approach reinforces the status quo, setting a "lowest common denominator" approach to a piecemeal, process-driven regime, rather than one that is focused on the outcomes described in the Act.

What the Act says

The Act makes no mention of the evidence on which Ofcom must base its recommendations for measures in the codes. There is a requirement that the measures must be technically feasible (Schedule 4, section 2 (c)) and age verification has some standards about effectiveness (Schedule 4, section 12 (3)). In terms of proactive tech, Ofcom is required to "have regard to the degree of accuracy, effectiveness and lack of bias achieved by the technology in question" and may refer to industry standards". (Schedule 4, section 13 (6))

Parliamentary debate

The growing weight of evidence of the nature and prevalence of online harms was a significant driver in the Government's decision to legislate, announced in May 2018. The opportunities for evidence to be submitted – from industry as well as the academic and civil society research communities – to influence the scope of the policy development and the legislation were provided at many stages between 2017 (the publication of the Government's Internet Safety Strategy Green Paper) and Royal Assent. These included pre-legislative scrutiny by a Joint Committee in 2021 of the draft Online Safety Bill and then Committee stages during the Parliamentary passage of the Bill between 2022-2023. A summary of, and links to, the Parliamentary stages is provided here and related research and commentary during that period is summarised here. Numerous Parliamentary inquiries on related topics took place during this time, each one accumulating more evidence via written submissions and oral testimony.

Ofcom's proposals

Evidence has been crucial to the decisions Ofcom has made, both as regards the risk register in Volume 2 and the underpinning analysis for the codes of practice in Volume 4.

We note that in Volume 2, para 5.10 it specifies that Ofcom – which relies here on third party evidence from a range of actors – has considered the evidence by reference to certain criteria: "method robustness, ethics, independence and narrative". It provides further information on these criteria, including the methodology of the studies, size and coverage, ethics (e.g. handling of personal data), whether stakeholder interests might have influenced findings and whether the commentary in the output matched the data found. By contrast, there is no such clear methodology for Volume 4 (and the methodology in Vol 2 is expressed so as only to apply to Vol 2). There is also a question as to whether the standards required for an academic research project should be the benchmark for policy making in this area because so much has not been investigated, not been proven or cannot be proven due to complexity. Again, we refer back to the merits of a "by design" safety obligation on companies to develop their own measures to address the risks it can see (via its own evidence) arising on their services.

There is a heavy reliance throughout the consultation document on statements from companies providing regulated services. "Best practice" examples are cited. But in many other areas, Ofcom refers to "limited" or "patchy" evidence for measures that work. This is particularly important given the increasing evidence from whistleblowers (e.g. Frances Haugen) and from litigation in the States (see our references provided in section 1, above) that some of the biggest social media companies have suppressed evidence and – it is claimed – sought to mislead both users and legislators. We include some of this evidence below.

We appreciate that Ofcom has only recently received its information-gathering powers and fully intends to use them to expand its evidence base in order to inform future iterations of the codes. This has been emphasised to us in a number of meetings and is set out volume 6 of the consultation documents (see e.g. 28.1-28.2,28.2).

> The statutory information gathering powers conferred on Ofcom by the Act give us the legal tools to obtain information in support of our online safety functions. These powers will help us to address the information asymmetry that exists between Ofcom and regulated services and to discover, obtain and use the information we need, including for monitoring and understanding market developments, supervising regulated services, and investigating suspected compliance failures. (Volume 6)

This is welcome. But it is not clear how Ofcom has determined how evidential thresholds had been satisfied, especially in relation to Volume 4. We also note that there are some concerns about whether solutions are proven to be effective, but we do not see a discussion of what the threshold is for that.

For example:

> "We also considered several other options regarding how services can assure their measures to mitigate and manage illegal content are effective. These could be alternatives to the options discussed above or could supplement them. However, we do not consider there is currently enough information on the effectiveness of other possible measures to be able to recommend them in Codes at this stage." (8.131)

See similarly Vol 4, 11. 14 and 11.15, c and 12.34.

> "**We are not proposing to recommend some measures which may be effective in reducing risks of harm**. This is principally due to currently limited evidence regarding the accuracy, effectiveness and lack of bias of the technologies that the measures refer to. We recognise that some of these measures may be proportionate for certain services to take, and welcome further innovation and investment in safety technologies to support ACM. We plan to consider further ACM measures for future versions of our Codes." (14.12)

More worryingly, even in some areas – such as "beacon platforms" – Ofcom have evidence of what may be effective but then make a judgement that this is not enough to recommend measures. "However, at this stage, we consider we require further evidence in order to propose a recommended measure tackling the harm created by the dissemination of these links, in particular about the following areas" (Vol 4. 14.221).

While there is a clear rationale for not recommending proven ineffective measures, this approach is worrying where there is some evidence of effectiveness. Moreover, absence of evidence is not evidence of ineffectiveness and responses in respect of which there is no evidence should not be excluded from the field of possible measures. More worryingly, Ofcom has also used lack of evidence in relation to its assessment of costs to justify the non-inclusion of tools in relation to smaller services.

> "Many of the measures we propose are for large services. This is often because we do not yet have enough information on the potential costs and benefits to know whether the measures are proportionate for smaller services at this point. As our understanding

> develops, it may be appropriate in future iterations of the Codes to expand the range of services for which some measures are recommended." (Vol 4, 11.16)

This begs the question as to why they have created this threshold for themselves when it so clearly prevents the recommendation of mitigation for a known, evidenced harm. Not only is there a question as to the appropriate evidence threshold, but the problem could have been avoided had Ofcom started from the premise that companies should address the issues arising from their risk assessment systemically or based on outcomes, rather than via a specific measure, and by a focus on safety by design as well as takedown of illegal content. This issue seems to have been a result of the approach taken to the sort of measures recommended. See also our discussion on the measures in the codes of practice in section 6, below.

This approach is likely to significantly limit the likelihood that there will be much material change in the online safety of users when these first codes of practice are published. Indeed, as we suggest above, it could potentially lead to a rowing back of some measures already deployed by services because they do not need to continue to resource them in order to comply with the codes. See, for example, this from the Atlantic Council's report on "scaling trust": "Until investments in reactive and proactive T&S are established as a requirement for doing business or a de facto generator of long-term value, the incentives structures necessary to ensure better, safer online spaces will continue to fail users—and societies." p 34

In this context, we were concerned to hear an Ofcom Principal describe, on a recent webinar addressed to businesses, how Ofcom's evidence threshold was in effect a bar to them codifying measures which are already accepted by regulated companies as "good practice" and how voluntary principles were all that they could rely on in many areas as a result.

> "Voluntary principles are already in place across a number of harms that a number of us have helped to formulate over the years .. and actually, to be candid, for quite a while some of those voluntary principles are going to go further than we're going to be able to go on the codes until we're able to catch up … **It's going to be easier to recommend something as a voluntary principle than it is to have to meet the bar of evidence to codify that in a code of practice**. So there will be some time where voluntary principles go further until we catch up .. a lot of those voluntary principles contain some really good practice things about what companies can be doing." (our emphasis) (WE Communications webinar: Navigating Tech Regulation in the Wake of the Online Safety Act – 31 January 2024; this extract is at 36 minutes in)

A further point that has been omitted entirely from consideration is that absence of evidence of a proposition is not proof that that proposition is not true. We also note that where there is

presumptive harm, especially harm which is serious in nature and wide reaching – as has been clearly evidenced by Vol 2 – that both Parliament in its debate and the overarching duty of care principle would dictate a more precautionary approach. Ofcom's position here is therefore not what would have been anticipated:

> "Recognising that we are developing a new and novel set of regulations for a sector without previous direct regulation of this kind, and that our existing evidence base is currently limited in some areas, these first Codes represent a basis on which to build, through both subsequent iterations of our Codes and our upcoming consultation on the Protection of Children. (Vol 4 11.14)"

Evidence

In previous work for Carnegie UK which set out the initial proposal for basing online harms regulation on a duty of care approach, Professor Lorna Woods and William Perrin set out the merits of the precautionary principle – already established within regulatory practice – as a means to address the risk of harm in areas of fast-moving innovation, where the evidence base may not nascent. We quote the relevant extract from that work here.

> "One of the recurrent arguments put forward for not regulating social media and other online companies is that they are unique or special: a complex, fast-moving area where traditional regulatory approaches will be blunt instruments that stifle innovation and require platform operators to take on the role of police and/or censors. Another is that the technology is so new, sufficient evidence has not yet been gathered to provide a reliable foundation for legislation; where there is a body of evidence of harm, in most cases the best it can do is prove a correlation between social media use and the identified harm, but not causation.
>
> We believe that the traditional approach of not regulating innovative technologies needs to be balanced with acting where there is good evidence of harm. The precautionary principle provides a framework for potentially hazardous commercial activity to proceed relatively safely and acts as a bulwark against short term political attempts to ban things in the face of moral panic. Rapidly-propagating social media and messaging services, subject to waves of fashion amongst young people in particular, are an especial challenge for legislators and regulators. The harms are multiple, and may be context- or platform- specific, while the speed of their proliferation makes it difficult for policymakers to amass the usual standard of long-term objective evidence to support the case for regulatory interventions. The software that drives social media and

messaging services is updated frequently, often more than once a day. Facebook for instance runs a 'quasi-continuous [software] release cycle' to its webservers. The vast majority of changes are invisible to most users. Tweaks to the software that companies use to decide which content to present to users may not be discernible. Features visible to users change regularly. External researchers cannot access sufficient information about the user experience on a service to perform long term research on service use and harm.

Evidencing harm in this unstable and opaque environment is challenging, traditional long-term randomised control trials to observe the effect of aspects of the service on users or others are nearly impossible without deep cooperation from a service provider. Nonetheless there is substantial indicative evidence of harm both from advocacy groups and more disinterested parties. OFCOM and ICO demonstrated that a basic survey approach can give high level indications of harm as understood by users. So, how do regulation and economic activity proceed in the face of indicative harm but where scientific certainty cannot be achieved in the time frame available for decision making?

This is not the first time the government has been called to act robustly on possible threats to public health before scientific certainty has been reached. After the many public health and science controversies of the 1990s, the UK government's Interdepartmental Liaison Group on Risk Assessment (ILGRA) published a fully worked-up version of the precautionary principle for UK decision makers. 'The precautionary principle should be applied when, on the basis of the best scientific advice available in the time-frame for decision-making: there is good reason to believe that harmful effects may occur to human, animal or plant health, or to the environment; and the level of scientific uncertainty about the consequences or likelihoods is such that risk cannot be assessed with sufficient confidence to inform decision-making.'

The ILGRA document advises regulators on how to act when early evidence of harm to the public is apparent, but before unequivocal scientific advice has had time to emerge, with a particular focus on novel harms. ILGRA's work focuses on allowing economic activity that might be harmful to proceed 'at risk', rather than a more simplistic, but often short-term politically attractive approach of prohibition. The ILGRA's work is still current and hosted by the Health and Safety Executive (HSE), underpinning risk-based regulation of the sort we propose. We believe that – by looking at the evidence in relation to screen use, internet use generally and social media use in particular – there is in relation to social media "good reason to believe that harmful effects may occur to human[s]" despite the uncertainties surrounding causation and risk. On this basis we

> propose that it is appropriate if not necessary to regulate and the following sets out our proposed approach." (Woods and Perrin, Online Harms: a statutory duty of care and regulator, Carnegie UK 2019; pp10-11)

Where there is evidence of harm – which Ofcom ably demonstrates in volume 2 – a proportionate regulatory response is to require measures to proceed in such a "precautionary" way, via well-established industrial processes such as product safety testing and risk-assessed design principles.

There is plenty of evidence from recent court filings and whistle-blower material that the big platforms, Meta in particular, have not done this despite internal evidence on the harmful design of their products and the decisions that would/should be taken to mitigate that. While Ofcom may not feel that it has – at present – evidence to support the recommendation of specific measures for all in-scope services to mitigate these harms, it is very likely that the biggest companies do but have chosen not to develop, test or deploy these measures. (Indeed, as far back as 2017, one of Facebook's co-founders, Sean Parker, admitted that they knew when developing the site that the objective was "How do we consume as much of your time and conscious attention as possible?" It was this mindset that led to the creation of features such as the "like" button that would give users "a little dopamine hit" to encourage them to upload more content. It's a social-validation feedback loop … exactly the kind of thing that a hacker like myself would come up with, because you're exploiting a vulnerability in human psychology." (Reported in the Guardian)

If the codes (as we discuss above), do not compel companies to comply with anything beyond the specific measures recommended therein, then there is no regulatory imperative and therefore no consequence for those services if they don't.

This underlines the importance of having an upfront catch-all measure in the codes on illegal content that requires companies to act on the knowledge they may already have about the harmful design effects of their products, notwithstanding the need also to adopt the evidence-based measures that Ofcom includes in the rest of the codes. (See section 1 above; and reiterated below.)

For example, the New Mexico attorney-general findings point to Meta's knowledge of the fact that users looking for images of children on their platforms – for example, on gymnastics – can be directed by their algorithms to videos sexualising children. It goes on: "The article notes that the problem of algorithm-served, sexual content was known to Meta. Meta conducted a review in connection with launching its Reels product, and the Journal reports that Vaishnavi J, Meta's

former head of youth policy, described the safety review's recommendation as: "'Either we ramp up our content detection capabilities, or we don't recommend any minor content,' meaning any videos of children" (p104)

Meta's awareness of how their platforms encouraged and enabled the discovery of suicide content goes back at least as far as 2019, when discussion on how to handle media responses to the Molly Russell case included the following (as summarised in the New Mexico AG filings):

> "Although the coroner's inquest took several years, Meta employees were acutely aware of the lack of safeguards built into Instagram and expressed their concerns in emails following the Guardian's outreach to Meta for comments on Ms. Russell's death. In a January 26, 2019 email thread addressing Meta's response to a forthcoming media story profiling "30 families of suicide victims accusing Instagram of killing their children," one Meta employee wrote: **"We are defending the status quo when the status quo is clearly unacceptable to media, many impacted families, and when revealed in press, will be unacceptable to the wider public."** Recipients of the thread included Zuckerberg, Sandberg, and Mosseri. Another Meta employee responded to echo the theme that Instagram protocols were insufficient: "our present policies and public stance on teenage self harm and suicide are so difficult to explain publicly that our current response looks convoluted and evasive . . . The fact that we have age limits which are unenforced (unenforceable?) and that there are, as I understand it, important differences in the stringency of our policies on IG vs Blue App [Facebook] makes it difficult to claim we are doing all we can." Sandberg eventually chimed in, asking whether Meta could improve its policies or whether it was a question of enforcement and confirmed "We can definitely say that we need to improve our enforcement of our policies." (p17) (our emphasis)

Note that the respondents were indeed describing their lack of action as "defending the status quo", something which – by taking the "best practice" evidence at face value, Ofcom is now at risk of baking into the regulatory obligations from the outset of the regime.

In his recent evidence to Congress, Meta whistleblower Arturo Bejar said: "Meta's current approach to these issues only addresses a fraction of a percent of the harm people experience on the platform. In recent years, repeated examples of harm that has been enabled by Meta and other companies has come to light, through whistleblowing, outside research studies, and many stories of distressing experiences people have there. Whenever such reports emerge, Meta's response is to talk about 'prevalence', and its investment in moderation and policy, as if that was the only relevant issue. But there is a material gap between their narrow definition of prevalence and the actual distressing experiences that are enabled by Meta's products.

However, managers including Meta CEO Mark Zuckerberg do not seem to seek to understand or actually address the harms being discussed. Instead, they minimize or downplay published findings, and even sometimes the results of their own research. They also try to obfuscate the situation by quoting statistics that are irrelevant to the issues at hand."

Also, on metrics, the New Mexico Attorney General case talks about how "the prevalence metric consistently underestimated the amount of problematic and illicit content displayed on Facebook. The prevalence metric contradicted the findings of Meta's own BEEF study, which showed a much greater "prevalence" of bad experiences involving illicit, questionable or violative conduct on Meta's platforms. This shows discrepancies between the BEEF study (reported instances only over the prior 7 days) and the corresponding Community Standards Enforcement Report (which reports "prevalence" over a longer period of time)"

In the complaint by the Utah Attorney-General against TikTok, it states: "A parent looking at TikTok's website might feel reassured by TikTok's increasing rates of moderation success for metrics like "proactive removal" rate. But these metrics share a common flaw: they compare violations caught quickly against violations caught slowly. TikTok's published metrics do not reflect at all how much violative content is not caught and therefore remains a danger to children."

The Utah materials go on to say, in relation to how Terms of Service are drafted: "Further, content promoting or normalizing risky sexual behaviors to teens, including sexual acts like choking, sex without a condom (known on the app as "breeding kink"), sex work, and relationships between children and older adult men is a grey area that appears to fall between TikTok's policy gaps; it is not a performance of "sexual activity," does not expose nudity, and does not constitute a "seductive performance.""

TikTok's public representations that it "addressed" filter bubbles amount to minor tweaks that have proven ineffective. For instance, TikTok allows users to "refresh" their feeds if "recommendations no longer feel relevant" or do not "provide enough topical variety," and it does not recommend "two videos in a row made by the same creator or that use the same sound." These modest changes have not stopped the app from recommending increasingly despairing messages, adult themes, and other dangerous content that violate TikTok's policies."

A recent case study on the impact of user tools on content recommended showed that user tools had little impact on content recommended; this is not the first such study.

Recommendation

We believe that, based on the analysis above, the addition of the proposed additional measures – as set out in section 1 above, with suggested wording, would address the problems we have identified. This approach avoids the risk of Ofcom effectively requiring something of companies that is ineffective and inefficient and is in line with the "precautionary principle" approach to regulation in other sectors where there are safety risks.

We include the wording we propose here again for completeness.

> "Design of functionalities, algorithms and other features
>
> *Product testing*
>
> For all services, suitable and sufficient product testing should be carried out during the design and development of functionalities, algorithms and other features to identify whether those features are likely to contribute to the risk of harm arising from illegal content on the service.
>
> The results of this product testing should form a core input to all services risk assessments.
>
> *Mitigating measures*
>
> For all services, measures to respond to the risks identified in the risk assessment should be taken, including but not limited to, providing extra tools and functionalities, by redesigning the features associate with the risks, by limiting access to them where appropriate or where the risk of harm is sufficiently severe by withdrawing the function, algorithm or other feature.
>
> Decisions taken on mitigating measures, as part of the product design process or as a response to issues arising from the risk assessment, should be recorded. (Note: this would be included in the record keeping duties under section 23 (u2U) and section 34 (search).)
>
> *Monitoring and measurement*
>
> All services should develop appropriate metrics to measure the effectiveness of the mitigating measures taken in reducing the risk of harm identified in the risk assessment. These measures should feed back into the risk assessment."

Issue 4: the approach to proportionality

Issue

Ofcom's approach to proportionality is primarily economic: to avoid imposing costs on companies. While the OSA requires regulated services take a "proportionate" approach to fulfilling their duties, and recognises that the size and capacity of the provider is relevant, the Act also specifies that levels of risk and nature and severity of harm are relevant. This focus on costs and resources to tech companies is not balanced by a parallel consideration of the cost and resource associated with the prevalence of harms to users (for example, on the criminal justice system or on delivering support services for victims) and the wider impacts on society (particularly, for example, in relation to women and girls and minority groups, or on elections and the democratic process). The assumption in the proportionality analysis that "small" means "less harm" due to less reach is also an issue, particularly given that it downplays the severe harm that can occur to minoritised groups on targeted, small sites - which we discuss further below. We look below in section 7 at how the principle of proportionality plays into Ofcom's differentiated approach to small and large companies.

What the Act says

There are 53 references to "proportionate" within the Act. While the Act defines proportionality (in relation to safety duties), Ofcom has not expressly stated how it is approaching the required balancing text.

Section 10 sets out the safety duties for all services as follows:

> (2) A duty, in relation to a service, **to take or use proportionate measures** relating to the design or operation of the service to—
>
> > (a) prevent individuals from encountering priority illegal content by means of the service,
> >
> > (b) effectively mitigate and manage the risk of the service being used for the commission or facilitation of a priority offence, as identified in the most recent illegal content risk assessment of the service, and
> >
> > (c) effectively mitigate and manage the risks of harm to individuals, as identified in the most recent illegal content risk assessment of the service (see section 9(5)(g)).

(3) A duty to operate a service **using proportionate systems and processes** designed to—

(a) minimise the length of time for which any priority illegal content is present;

(b) where the provider is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way, swiftly take down such content.

(4) The duties set out in subsections (2) and (3) apply across all areas of a service, including the way it is designed, operated and used as well as content present on the service, and (among other things) require the provider of a service to take or use measures in the following areas, **if it is proportionate** to do so—

(a) regulatory compliance and risk management arrangements,

(b) design of functionalities, algorithms and other features,

(c) policies on terms of use,

(d) policies on user access to the service or to particular content present on the service, including blocking users from accessing the service or particular content,

(e) content moderation, including taking down content,

(f) functionalities allowing users to control the content they encounter,

(g) user support measures, and

(h) staff policies and practices.

Section 10 (10) sets out the interpretation:

(10) In determining what is proportionate for the purposes of this section, the following factors, in particular, are relevant—

(a) all the findings of the most recent illegal content risk assessment (including as to levels of risk and as to nature, and severity, of potential harm to individuals), and

(b)the size and capacity of the provider of a service.

A comparable approach to proportionality is found in the analogous provisions for search services in section 27.

In Schedule 4, which sets out details on how Ofcom should approach the codes of practice, it says:

2 (c) the measures described in the code of practice must be proportionate and technically feasible: measures that are proportionate or technically feasible for providers of a certain size or capacity, or for services of a certain kind or size, may not be proportionate or technically feasible for providers of a different size or capacity or for services of a different kind or size; (NB this does not mention cost in relation to proportionality)

2 (d) then makes a specific reference to proportionality in relation to the risk of harm:

"the measures described in the code of practice that apply in relation to Part 3 services of various kinds and sizes must be proportionate to OFCOM's assessment (under section 98) of the risk of harm presented by services of that kind or size."

It is our assessment that the Act, as drafted, does not direct Ofcom to take costs into account as the main driver of whether measures are proportionate or not but to make a judgement as to whether the recommendation of the measures itself is proportionate based on the kind or size of a service and the likely level of risk that those services pose, according to the functionalities that are identified in the risk assessment and also to weigh that against the severity of the harms also identified in the risk assessment (including the recognition that some of those harms might constitute an interference with individuals' human rights).

Parliamentary debate

In the Lords Committee stage debate on 2 May, Lord Parkinson – the Government Minister – gave the following reassurances in relation to the child safety duties which have broader application (indeed he emphasised this) to all duties that fall to companies regulated under the new regime:

> "The provisions in the Bill on proportionality are important to ensure that the requirements in the child-safety duties are tailored to the size and capacity of providers. It is also essential that measures in codes of practice are technically feasible. This will ensure that the regulatory framework as a whole is workable for service providers and enforceable by Ofcom. **I reassure your Lordships that the smaller providers or providers with less capacity are still required to meet the child safety duties where their services pose a risk to children. They will need to put in place sufficiently stringent systems and processes that reflect the level of risk on their services, and will need to make sure that these systems and processes achieve the required outcomes of the child safety duty.** …

> The passage of the Bill should be taken as a clear message to providers that they need to begin preparing for regulation now—indeed, many are**. Responsible providers should already be factoring in regulatory compliance as part of their business costs**. Ofcom will continue to work with providers to ensure that the transition to the new regulatory framework will be as smooth as possible." (Hansard 2 May col 1485)

Ofcom's proposals

We have set out a lot of material in section 7, below, in relation to the judgements on "proportionality" that lead to differential obligations being placed on small and large services and do not propose to repeat them here.

The following extracts are relevant here to demonstrate the over-emphasis on costs as a means by which to judge proportionality.

For example, on risk assessments:

> Vol 3, 9.66L "In addition, our proposed methodology is intended to be flexible depending on service's risk levels, size and resources in order to minimise the cost burden. We intend that it could be integrated into existing risk management practices to improve the effectiveness of online safety risk assessments and minimise additional costs"
>
> Vol 4, p4 "We consider larger services will tend to be better able to bear the costs of the more onerous measures than smaller services."

The choice of the word "onerous" – which appears 20 times in volume 4 – is, we would contend, a subjective, value judgement. Dictionary.com provides this definition: *"burdensome, oppressive, or troublesome; causing hardship: onerous duties; having or involving obligations or responsibilities, especially legal ones, that outweigh the advantages: onerous agreement*". It is an inappropriate choice of word by a regulator charged with implementing a regime that is about reducing harm to individuals (including human rights obligations) and not about preserving the profitability of companies.

We note that Ofcom is also using it in direct communications with businesses, such as an email on 19 February 2024 providing - as a response to an FAQ framed as "I don't agree with your proposed measures or think they will disproportionately affect my business. What should I do?" - the reassurance that "we've designed our measures to be proportionate: generally, we've only proposed the most onerous ones for the largest and riskiest services." We would contend

here that when Ofcom says the measures are "proportionate", in fact it means "light-touch" e.g. the opposite of "onerous". This does not align with the reassurances given by Government Ministers to Parliament in relation to the illegal content duties and their application to all businesses.

In relation to specific measures, Ofcom argues the following:

> "We consider it can be prudent to exempt smaller services from incurring those costs (where appropriate provided they are not high risk), as there will often be significant uncertainty in any assessment of benefits and costs, and **we want to reduce the possibility of imposing financially damaging costs on businesses when the magnitude of benefits expected to result from the measure is uncertain**." (11.53)
>
> On measure 2 (developing content policies), which only applies to large and multi-risk services: "Services that do not currently have internal content policies would incur the costs of developing them. This could take a small number of weeks of full-time work and involve legal, regulatory, as well as different ICT staff, and online safety/ harms experts. In some cases, services may use external experts which could increase costs. Agreeing new policies may also take up senior management's time which would add to the upfront costs. For most services we expect these costs to be in the thousands of pounds, although larger/riskier services may require more complex content policies which may increase costs. In addition there may be some small ongoing costs to ensure these policies remain up to date over time." (12.88)

This has the danger of suggesting that the riskier the business is – and therefore the greater the potential compliance costs – the less likely it is that Ofcom will find the possibility of costs to be proportionate.

> "We are not proposing to recommend this measure for smaller and lower risk services. We consider the benefits of an internal content moderation policies are likely to be materially smaller for services which are neither large nor face material risks. **They are unlikely to face large volumes of content they need to assess.** So even though the costs of this measure are low, we do not propose to recommend it for such services." (our emphasis) (12.98)

There seems to be no evidence underpinning this assessment as to quantity of content and given the threshold for small services, seems unlikely to be the case.

The severity of the crime and the costs to society (quantified at c£2.bn in the "underestimate" provided in the Government's Impact Assessment) are significant. Yet Ofcom's consideration of the merits of CSAM measures are also weighed up against the costs to business – without considering the extent of the harms to the individuals nor the costs to society to eradicate this sort of crime and to provide support to affected individuals:

> "The level of detail and complexity in the comparison of costs and benefits is greater for some measures than others. This sometimes reflects the availability of information. It can also reflect where a more detailed assessment is more likely to impact our recommendations, and when it can affect which services we recommend measures for. This is especially the case for some of the measures we recommend to reduce grooming and the hash matching measure we recommend to reduce CSAM, where we carefully consider whether to recommend the measures for smaller services". (Vol 4, 11.32)

Elsewhere, while inaccurate flagging of content does have freedom of expression and privacy concerns which the companies have raised with Ofcom, this comes second in their analysis relating to the costs of using keyword lists to ensure that content is "not incorrectly removed":

> "Evidence from industry indicates that keyword lists are used alongside human review, and although this can help to ensure that content is not incorrectly removed from a service, it is likely to have significant cost implications for services and may also have freedom of expression and privacy implications if content is incorrectly flagged as being CSAM." (Vol 4, 14.318)

While there are legitimate freedom of expression and privacy concerns, Ofcom does not consider the values that are being protected here (which might also be human rights), especially given the possibility of procedural safeguards found elsewhere in the Act to protect against over-takedown. The balance found here is particularly concerning given how Meta reacted (or failed to react) to some instances of CSAM on its platforms. This is from the New Mexico court filing: "Meta does not notify predators who post CSAM or other sexualized images or make inappropriate contacts with or comments to minors that their conduct violates community standards ***because the company does not want to run the risk of offending users.*** As a result, Meta forgoes the opportunity to let predators know that it is monitoring their activities, which would deter them from abusing the platforms" (p95; our emphasis).

Does the focus on costs suggest that if a company does not have risk management processes in place, the costs of implementing it are not justifiable? If there is a "flexible", proportionate approach to minimise cost burden, then it is an inescapable fact that some companies should

incur more costs if they are starting from a lower base. If their services are risk assessed and found to be unsafe, then the costs of rectifying this (again, emphasising that this is about illegal content) should not be weighed up against the necessity to address the lack of safety. Arguing against this – on the basis of proportionality – does not account for the significance of harm, the impact on users and the costs to society. We note that the quantification of costs to society in the Government's Impact Assessment – while referenced a few times Volume 2 – is not used to counterbalance Ofcom's own assessment of what the likely costs are to services in Volume 4.

Evidence

The Government's 2022 Impact Assessment (IA) quantified the cost to society of a number of illegal and other harms (including CSEA, hate crime, drugs, modern slavery and cyberstalking) and estimated that these added up to £5 billion/year. The IA went on to say that "these estimates are likely to underestimate the full extent of online harms for several reasons:

- It has only been possible to quantify the cost of a subset of all online harms in scope: there are a number of harms that are encountered by a significant number of adults and children in the UK, but for which there is no evidence on which to make an estimate of their cost. These include encouraging terrorism and radicalisation online, which 5% of adults and 6% of children in the UK have encountered, and encouraging self-harm, which 5% of adults and 10% of children have encountered.
- For those harms that have been quantified, a conservative approach has been undertaken. For example, for illegal harms analysis is based on the number of recorded offences with an online element, which is likely to understate the true prevalence (as some crimes will go unreported - although this is adjusted in part by the use of multipliers where appropriate)
- Crimes may feature an online element but not be flagged as online: currently, whether a crime is recorded as having an online element is reliant upon police recording practices and how police forces apply the online flag. This, again, will reduce the reported prevalence of a given harm, and lead to an underestimate of its cost." (p80)

Against this backdrop, the IA – as we noted above – focused on the "safety by design" measures that businesses, especially small ones, could start adopting to get ahead of the regulation**:** **"**While per business costs are expected to be higher for medium and large businesses, it is important to consider the possibility that some in-scope SMBs will have limited resources for compliance. To minimise burdens on SMBs, it will be vital for Ofcom to work with businesses and to ensure both requirements and enforcement are proportionate to the risk of harm and resources available to businesses. Proportionality in the context of effective safety measures must be balanced against the risk of harmful content being displaced to smaller and less well-

equipped platforms. The government and Ofcom will work with SMBs to ensure that steps taken are effective in both reducing harms and minimising compliance costs. The government's Safety by Design framework and guidance is targeted at SMBs to help them design in user-safety to their online services and products from the start thereby minimising compliance costs." (p90)

At p94, the importance of building in this safety by design approach was emphasised again in relation to "SMBs": "The Government is also developing a Safety by Design framework targeted at SMBs that will support businesses in adopting a "Safety by Design" approach, helping them design in user-safety to their online services and products. This work will produce practical online guidance tailored to SMBs. The framework will support SMBs to prepare for the introduction of the duty of care."

In addition, the Impact Assessment – in relation to a specific small and medium sized business concern around "More discretion for smaller businesses to meet regulatory requirements* (e.g. extended transition period or temporary exemption)" said that these measures were not required because:

> "This was not considered separately as the duty of care approach already builds in significant discretion for businesses to decide how to meet regulatory requirements… **businesses will not face prescriptive requirements, but will be expected to assess their level of risk and put in place proportionate measures to address this**. Laying of codes will undergo consultation and IAs and will be staggered allowing time for SMBs to comply with individual codes, as opposed to a specific date in which the whole regime comes into force at once." (p93)

Ofcom has, in these codes, done the opposite of this: taken the discretion away from companies, set out prescriptive requirements based on their own assessment of what is proportionate regardless of the level of risk rather than leaving this to regulated services – of whatever size – to justify their own measures based on their own risk assessment and evidence.

Moreover, the primacy of Ofcom's assessment about costs - there is little recognition that there are legitimate costs in running a business, but, as we note above, Ofcom takes the companies' current investment as the acceptable baseline - so the worse a company currently is the more it could argue that there's an increase in costs and that therefore the regulation is disproportionate. And at one point Ofcom even says for companies that are taking some measures, requiring them to continue to take those measures could incur a cost if they wanted to remove the feature. [see eg 22.31].

There has been one very prominent recent example of what happens when companies take operational, governance and resourcing decisions based on costs rather than on safety and the risk of harm. The Australian e-Safety Commissioner recently reported on information provided to her office by X/Twitter via a transparency report including the decision to cut staff working on safety globally, which demonstrates what happens when costs rather than risks are the primary driver of company decision-making: " X Corp. said Twitter/X's global trust and safety staff have been reduced by a third, including an 80 per cent reduction in the number of safety engineers, since the company was acquired in October 2022. The company also said the number of moderators it directly employs on the platform have been reduced by more than half, while the number of global public policy staff have also been reduced by almost 80 per cent."

What is notable in this report are the findings of the impacts on the platform's safety. In the same period since the acquisition by Elon Musk:

- there had been a 20% slowing in the median time to respond to user reports about Tweets and a 75% slowing in the median time to respond to direct messages. safety notes that prompt action on user reports is particularly important given that Twitter solely relies on user reports to identify hateful conduct in direct messages.
- As of May 2023, X Corp. reported that no tests were conducted on Twitter recommender systems to reduce risk of amplification of hateful conduct. However, X Corp. stated no individual accounts are artificially amplified, and that its enforcement policies apply to Twitter Blue accounts in the same way as other accounts.
- As of May 2023, automated tools specifically designed to detect volumetric attacks or "pile-ons" in breach of Twitter's targeted harassment policy were not used on Twitter.
- As of May 2023, URLs linking to websites dedicated to harmful content are not blocked on Twitter.
- From 25 November 2022 (the date it was announced)1 to 31 May 2023, 6,103 previously banned accounts were reinstated by Twitter, which eSafety understands relates to accounts in Australia. Of these, 194 accounts were reinstated that were previously suspended for hateful conduct violations. X Corp. stated that Twitter did not place reinstated accounts under additional scrutiny.

Recommendation

Based on the Parliamentary debates, Government statements and the Government's own impact assessment, we would argue that Ofcom's interpretation of what is "proportionate" is not appropriate; the issue of costs has weighed too heavily, and express consideration of the nature and severity of harms as required by sections 10(10)(a) and 27(10)(a) has not taken place. Measures Ofcom has discounted on the basis of costs should be reconsidered. Moreover, we refer back to the recommendation we make in section 1 for additional measures relating to product safety testing and safety by design to be added to the draft codes, which would place the responsibility on services (of all sizes) to take measures that are proportionate to them to address the risk of harm that is identified in their risk assessment.

We include the wording we propose again for completeness.

> "Design of functionalities, algorithms and other features
>
> *Product testing*
>
> For all services, suitable and sufficient product testing should be carried out during the design and development of functionalities, algorithms and other features to identify whether those features are likely to contribute to the risk of harm arising from illegal content on the service.
>
> The results of this product testing should form a core input to all services risk assessments.
>
> *Mitigating measures*
>
> For all services, measures to respond to the risks identified in the risk assessment should be taken, including but not limited to, providing extra tools and functionalities, by redesigning the features associate with the risks, by limiting access to them where appropriate or where the risk of harm is sufficiently severe by withdrawing the function, algorithm or other feature.
>
> Decisions taken on mitigating measures, as part of the product design process or as a response to issues arising from the risk assessment, should be recorded. (Note: this would be included in the record keeping duties under section 23 (u2U) and section 34 (search).)
>
> *Monitoring and measurement*
>
> All services should develop appropriate metrics to measure the effectiveness of the mitigating measures taken in reducing the risk of harm identified in the risk assessment.

These measures should feed back into the risk assessment."

We would also recommend that Ofcom should consider a requirement that there is no rolling-back on allocation of resources/measures introduced by companies that contribute to their compliance with the duties – for example, to avoid a situation such as that which has arisen at X/Twitter.

Issue 5: The approach to human rights

We also attach the text of this section separately as a standalone PDF.

Issue

The OSA directs Ofcom to consider freedom of expression (Art 10 ECHR) and privacy (Article 8 ECHR), but these are not the only relevant rights – as indeed Ofcom notes.

All the rights protected by the Convention should be considered when considering the impact of the regime – or the lack of it. So, as well as the qualified rights of freedom of expression (Article 8 ECHR), the right to private life (Article 11 ECHR) and rights noted by Ofcom – e.g. the right to association (Article 11 ECHR) – we should consider other rights including the unqualified rights – the right to life (Article 2 ECHR), freedom from torture and inhuman and degrading treatment (Article 4 ECHR) as well as the prohibition on slavery and forced labour (e.g people trafficking) (Article 4 ECHR). Note also that rights can include positive obligations as well as an obligation to refrain from action; a public body can infringe human rights by failing to protect as well as by interfering itself in an individual's rights.

Article 14 ECHR constitutes the requirement for people not to be discriminated against in the enjoyment of their rights; all people (and not just users of a particular service) should be considered. This reflects the general principle of human rights that all people's rights should be treated equally – and indeed that the starting point is that no right – for example, freedom of expression – has automatic priority over another. It also means that the European Court has adopted a specific methodology for balancing rights of equal weight (see e.g. *Perinçek v. Switzerland* (27510/08) [GC] 15 October 2015, para 198; *Axel Springer AG v. Germany* (39954/08) [GC] 7 February 2012, paras 83-84 on the balance between articles 8 and 10) rather than its typical approach where a qualified right may suffer an interference in the public interest but that interference must be limited. This difference in methodology reaffirms the significance of seeing all the rights in issue when carrying out balancing exercises. A failure to carry out a proper balance by national authorities has itself led to a finding of a violation of the procedural aspects of the relevant right. – the precise factors taken into account in the balance will vary depending on the underlying facts in a case and the rights involved.)

Note also that Article 17 prohibits the abuse of rights so that "any remark directed against the Convention's underlying values would be removed from the protection of Article 10 by Article 17" (*Seurot v France* (57383/00), decision 18 May 2004). While this applies only to a narrow sub-set of speech, it is nonetheless a factor that should form part of the balancing exercise

where relevant. Areas where Article 17 might be relevant include threats to the democratic order (*Schimanek v Austria* (32307/96), dec 1 February 2000); racial hatred (*Glimmerveen and Hagenbeek v NL* (8348/78 8406/78), dec 11 October 1979); holocaust denial (*Garaudy v France* (65831/01), dec 24 June 2003); religious (*Belkacen v Belgium* (34367/14), dec 27 June 2017) or ethnic (*Ivanoc v Russia* (35222/04), dec 20 February 2007) hate; hatred based on sexual orientation; incitement to violence and support for terrorist activity (*Roj TV A/S v Denmark* (24683/14), dec 18 April 2018). The Court has not considered CSAM material but it is submitted that it, likewise, would fall outside the protection of Article 10.

What the Act says

Section 22 sets out duties with regard to freedom of expression and privacy, and says that all user-to-user services:

> (2) When deciding on, and implementing, safety measures and policies, a duty to have particular regard to the importance of protecting users' right to freedom of expression within the law.
>
> (3) When deciding on, and implementing, safety measures and policies, a duty to have particular regard to the importance of protecting users from a breach of any statutory provision or rule of law concerning privacy that is relevant to the use or operation of a user-to-user service (including, but not limited to, any such provision or rule concerning the processing of personal data).

An analogous provision (section 33) applies in relation to search services.

As a public body, Ofcom falls within section 6 of the Human Rights Act which specifies that "[i]t is unlawful for a public authority to act in a way which is incompatible with a Convention rights".

Ofcom's proposals

The concern here is that Ofcom's approach, as set out in its illegal harms consultation, considers only the rights of users (as speakers) and has principally focused on their freedom of expression. In doing so, it has not really considered the nature of the speech (which the Convention court does take into account), nor provided evidence that speech in some instances would be chilled – it has rather hypothesised a rather theoretical concern. (In *Wille v Liechtenstein* (28396/95) [GC], 28 October 1999, the Court suggested that interferences could take a wide range of forms "formality, condition, restriction or penalty", but this does not seem

to match what Ofcom is suggesting. In *Metis Yayıncılık Limited Şirketi v Turkey* (4751/07), decision 20th June 2017, the Court suggested that there should be some substance to any cause leading to a claim of chilling effect where swiftly terminated criminal proceedings were not deemed to have a chilling effect; see also *Schweizerische Radio- und Fernsehgesellschaft and Others v. Switzerland* (68995/13), decision 12 November 2019, para 72. It has not considered the rights of other users and non-users that require steps to be taken against rights infringing harms – and where the infringement of a right has been recognised in the judgments of the European Court, or the opinion of UN Special Rapporteurs. This means that any balancing exercise is skewed towards not taking action for fear of inconveniencing users (who could well be infringing the rights of others) and companies.

We set out a number of examples below taken from various sections of the Illegal Harms consultation to demonstrate our concern. These intersect with our concerns about the proportionality judgements that underpin the overall approach in the consultation which we cover in section 2, above.

Ofcom on prioritising rights of users over rights of intended victims with regard to content moderation:

- "Content moderation is an area in which the steps taken by services as a consequence of the Act may have a significant impact on the rights of individuals and entities - in particular, to freedom of expression under Article 10 ECHR and to privacy under Article 8 of the European Convention on Human Rights ('ECHR')." (Vol 4, 12.57)

Ofcom on applying a proportionality test to including measures in the codes

- "to include a measure in the Codes, we need to assess that the measure is proportionate (with reference to both the risk presented by a service, and its size, kind and capacity) and does not unduly interfere with users' rights to freedom of expression and privacy." (Vol 4 11.22) – **no mention of the rights of others – whether this be their freedom of expression and privacy or other aspects of Article 8, let alone Articles 2, 3 or 4.**

Ofcom on recommending cumulative risk scoring systems

- "We consider that cumulative risk scoring systems could provide various benefits for tackling illegal harms such as fraud, drugs and weapons offences, child sexual exploitation and abuse, terrorism, and unlawful immigration. We recognise however that there is significant complexity involved in these systems, and that there could be adverse impacts on user privacy or freedom of expression if the operation of the system were to result in inappropriate action being taken against content or user accounts. We

have limited evidence on this at present. As a result, we are not proposing to include a recommendation that services use cumulative risk scoring systems in our Codes of Practice at this time" (Vol 4, 14.322) – **adverse impact on freedom of expression and privacy trumps "various benefits" for tackling illegal harms but does not consider the need to protect fundamental rights**

Ofcom's interpretation of the "chilling effect"

- "In addition, there could be a risk of a more general 'chilling effect' if users were to avoid use of services which have implemented a more effective content moderation process as a result of this option. (Vol 4, 13.52)Potential interference with users' freedom of expression arises insofar as content is taken down on the basis of a false positive match for CSAM or of a match for content that is not CSAM and has been wrongly included in the hash database. In addition, there could be a risk of a more general 'chilling effect' if users were to avoid use of services which have implemented hash matching in accordance with our option." (Vol 4, 14.87)
- "Potential interference with users' freedom of expression arises insofar as content detected by services deploying keyword detection technology in accordance with this option does not amount to a priority offence regarding articles for use in frauds, but is wrongly taken down on the basis that it does. There could also be a risk of a more general 'chilling effect' if users were to avoid use of services which have implemented keyword detection technology in accordance with this option." (Vol 4. 14.281)
- "We recognise that these user support measures may have a limited chilling effect on the rights to freedom of expression and freedom of association in that they would briefly delay children from disabling defaults and may result in children being less likely to do so (preserving the existing restrictions on their rights outlined in paragraph 18.65 above). The measures may also result in children being less likely to establish new connections or communicate with new users online." (Vol 4, 18.135)
- **Chilling effect here is to dissuade or inconvenience users from using services that act robustly on illegal harms including CSAM and fraud. There is no mention of "chilling effect" in relation to impact of individual users on others.**

Ofcom on balancing freedom of expression rights with recommending measures for strikes or blocking of accounts

- "Although blocking and strikes may be a way of tackling illegal content, there are also concerns about the use of these systems on lawful speech. Preventing a user from accessing a service means removing their ability to impart and receive information and to associate with others on that service. It therefore represents, for the duration of the block and in respect of that service, a significant interference with that user's freedom

of expression and association. The impact also extends to other users, who will be unable to receive information shared by the blocked user on the service in question. Restricting access to certain functionalities as part of a strikes system may also interfere with user rights, for example if the user is prevented from posting content on the service." (Vol 4 21.39) - **no consideration of rights protected through such blocking; or the value ascribed to the speech in the blocked account as regards both the speakers' rights and those receiving the information.**

- "Our proposed recommendation around strikes and blocking in this consultation relates to proscribed groups. We are inviting further evidence from stakeholders to be able to explore broadening this in future work; in particular, we are aiming to explore a recommendation around user blocking relating to CSAM early next year. We are particularly interested in the human rights implications, how services manage the risk of false positives and evidence as to the effectiveness of such measures." (Vol 4 11,15) - **measures for CSAM blocking not recommended in the first codes as a result despite the impact on children and likely interference with children's Article 8 and 3 rights and possibly also Articles 2 and 4.**

Evidence

1. The Silencing Effect of Abuse – Article 10 ECHR

As Ofcom has recognised, women and other minoritised groups receive a disproportionate amount of abuse– abuse here can take various forms from direct threats to gendered or racist misinformation and the use of deepfakes to undermine and harass – to name but a few. (For example, in Volume 2's discussion of the nature of the risks arising from the harassment, stalking threats and abuse offences, Ofcom notes: "Reviewing Ofcom and third-party evidence indicates that these offences disproportionately affect certain identity groups – most notably women – alongside other intersecting risk factors; the impact on those individuals can be severe." (para 6E:15)) Yet it was established for at least 5 years that "*online gender-based abuse and violence assaults basic principles of equality under international law and freedom of expression*". (Press Release UN experts urge States and companies to address online gender-based abuse but warn against censorship) Dubravka Šimonović, the UN Special Rapporteur, in 2018 highlighted the importance of applying a human rights-based approach to online violence against women (see UN Human Rights Council. Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective. June 2018. 38th session) and it has been recognised that women in particular are being targeted, especially those in public life. (Naser Al Wasmi, "UN High Commissioner for Human Rights says women are being silenced" The National, 28 September,

2018). Ofcom's own risk analysis finds that "Experiencing abuse and harassment can have a silencing effect, making victims and survivors feel unsafe in expressing themselves on social media services. Human rights organisation Amnesty International found that 24% of women experiencing online abuse and harassment said they stopped posting their opinions on certain issues. A study from 2016 found that 27% of US internet users censor their own online posts for fear of being harassed. This silencing effect can harm victims and survivors' careers. A study of women journalists found that those facing abuse and harassment reported making themselves less visible (38%), missing work (11%), leaving their jobs (4%), with some deciding to abandon journalism altogether (2%)." (Volume 2, 6E:23)

As the UN Special Rapporteur on Freedom of expression emphasises, there should be no trade-off between the right to be safe and the right to speak. (UN, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan (A/78/288)).

This point can be made in relation to other minoritised groups – and those with intersectional identities suffer particularly. In short, the failure to provide a safe environment in which to express themselves – which the European Court of Human Rights recognises is part of the positive obligations under Article 10 (*Özgür Gündem v. Turkey* (23144/93), 16 May 2000; *Dink v Turkey* (2668/07) judgment 14 September 2010; *Khadija Ismayilova v. Azerbaijan* (65286/13 and 57270/14) 10 January 2019) constitutes an infringement of the victims' expression rights, as well as those who share relevant characteristics with them. The point about freedom of expression is particularly important for those in public life, but the underlying facts in any given case may also implicate Article 8 and have an even wider impact.

This point about shared characteristics is also important. Of course, men receive abuse online too, but it seems more addressed to ideas (and thus could be categorised as an extreme form of debate) whereas women seem to be targeted for their characteristics, (Marjim Nadim and Audun Fladmoe, 'Silencing Women? Gender and Online Harassment' (2021) 39(2) *Social Science Computer Review* 245) which is pure abuse, and which does not receive high protection under the Convention if it receives any at all (See eg Norwood v UK (2004) App No 23131/03 on racist abuse and the impact of Article 17 ECHR) By contrast, as discussed below, negative stereotyping of a group, when it reaches a certain level, is capable of impacting on the group's sense of identity and the feelings of self-worth and self-confidence of its members. It is in this sense that it can be seen as affecting their "private life" within the meaning of Article 8(1). Moreover, the approach to dealing with misogynistic trolling in particular in imposing obligation on the victim itself contributes to an environment in which victims are not taken seriously and rape culture (through symbolic violence) continues. (This can also be seen in media portrayals

of trolling: Karen Lumsden and Heather Morgan "Media framing of trolling and online abuse: silencing strategies, symbolic violence, and victim blaming" (2107) 17(6) *Feminist Media Studies*) The impacts, while clearly affecting speech, are deep, wide-ranging and often misunderstood and undervalued (Azmina Dhrodia, 'Social media and the silencing effect: why misogyny online is a human rights issue - A survey of women in eight countries shows a clear trend' *The New Statesman*, 23 November 2017) – especially when the violation of rights is not recognised and what is going on is characterised vaguely as harmful.

2. CSAEM/Grooming of Children – Article 8 and 3

Article 8 is not just about the confidentiality of communications. The text of Article 8 covers four groups of interests: private life, family life, home and correspondence, each of which has been interpreted broadly. As for Article 10, in addition to protecting interference with these rights by public authorities, there are positive obligations to ensure that Article 8 rights are respected even as between private parties. Positive obligations are particularly significant when the interests at stake involve "fundamental values" or "essential aspects" of private life, and the Court looks for deterrence and effective protection. So, granting an amnesty for the perpetrator of sexual assault constituted a breach of Article 8 (and also Article 3) – a point that should be borne in mind when understanding severity of harm and appropriate balances in terms of impact on with service providers or the user speaking. (*EG v Moldova* (37882/13), judgment 13 April 2021)

As regards the relevance of Article 8 to CSAEM offences, *KU v Finland* specified that an advert of a sexual nature placed in the name of a 12-year-old boy on an Internet dating site, leaving him open to approach by paedophiles, was indisputably within Article 8 which covers the physical and moral integrity of a person. Here the Court emphasises the potential threat to the boy's physical and mental and moral welfare, as well as the boy's vulnerability because of his age. This then was a grave threat: "sexual abuse is unquestionably an abhorrent type of wrongdoing, with debilitating effects on its victims. Children and other vulnerable individuals are entitled to State protection, in the form of effective deterrence, from such grave types of interference with essential aspects of their private lives". (*KU v Finland* (2872/02), judgment 2 December 2008, para 46) In this instance, although there were laws in place they were ineffective and the Finnish Government had failed to "put in place a system to protect child victims from being exposed as targets for paedophiliac approaches via the Internet". (*KU* para 48) In *Söderman*, (*Söderman v Sweden* (5786/08) [GC], judgment 12 November 2012) the step-father of a 14 year old covertly videoed her underdressing before showering; the film was subsequently destroyed without anyone seeing it. He was subsequently acquitted because the

act of filming was not illegal in itself. Again, this fell within Article 8 and the State's obligation to protect the physical and psychological integrity of an individual from other persons – especially where that person is a child. Rape and sexual abuse of a child implicate fundamental values and essential aspects of private life, and aggravating factors include the offence taking place in the child's home where the child should feel safe. So, "in respect of less serious acts between individuals, which may violate psychological integrity, the obligation of the State under Article 8 to maintain and apply in practice an adequate legal framework affording protection …" that must be "sufficient". (*Söderman*, para 85, 91) So this is not just the legislative framework that is considered, but also the *implementation* of that framework.

KU was dealt with under Article 8; it did not involve physical assault of the child. There is some suggestion that verbal abuse without physical violence would fall within Article 8 rather than Article 3. (*Association Accept and Others v Romania* (19237/16), judgment 1 June 2021; *Fo.O v Croatia* (29555/13), judgment 22 April 2021) These more serious assaults may trigger Article 3 or even Article 2 (both of which are unqualified rights). Here there are also positive obligations on the state to protect the personal integrity of a child, with breaches of the right being found where state procedures are ineffective (*X v Bulgaria* (22457/16) [GC], judgment 2 February 2021; *Z and Others v UK* (29392/95) [GC], judgment 10 May 2001).

3 VAWG-related offences – Articles 8 and 3

A similar analysis can be made in relation to a whole range of offences related to violence against women. Coercive control, domestic violence and similar behaviours likewise infringe the survivors' Article 8 and, in some instances, Article 3 rights. (See eg *Vuckovic v. Croatia* 12 December 2023) While many of the cases involve a State's failure to protect the victim of domestic violence against physical injuries and psychological damage, there are cases involving digital tools. In *Volodina v Russia (No 2)*, the claimant claimed breaches of her Convention rights arising from the State's failure to take action in respect of cyber-harassment: her former partner had used her name, personal details and intimate photographs to create fake social media profiles, that he had planted a GPS tracker in her handbag, that he had sent her death threats via social media. In this case the Court found a violation of Article 8. Image-based sexual abuse likewise engages article 8: *Ismayilova*. (*Ismayililova*, para 116) More generally, the Court has recognised 'cyber-bullying' as an aspect of violence against women and girls and that it could take on a variety of forms, including cyber breaches of privacy, intrusion into the victim's computer and the capture, sharing and manipulation of data and images, including private data. In this case, *Buturugă v Romania,* the Court found a violation of Articles 3 and 8.

These offences also contravene the Istanbul Convention and the UN Convention on the Elimination of Discrimination Against Women. They should be seen as infringing fundamental rights, infringements in respect of which the State has positive obligations.

Ofcom's risk profile analysis sets out some of the evidence around coercive control offences:

> "To put the risks of harm into context, a survey by Women's Aid of victims and survivors of domestic abuse in 2013 found that 45% had experienced abuse online during their relationship. Research by Refuge conducted in 2022 found that 82% of victims and survivors of tech abuse had experienced harassment or stalking on social media, 41% had experienced threats of violence and 29% had experienced intimate image abuse. However, getting up-to-date measurement of CCB, including the specific prevalence of online CCB, is challenging. It is not only a relatively new offence, but also overlaps with other offences (see in 6G.10) (Vol 2: 6G:14).
>
> "Low incidences of reporting further complicate the measurement of CCB; like related offences, CCB is consistently under-reported. Refuge found that half of victims and survivors (49%) said they told nobody about the abuse, with only 13% of women reporting the abuse to the social media platform where the abuse happened. Only one in ten victims and survivors (10%) reported it to the police. One in five victims and survivors (18%) did not tell anyone because they were not sure how to report the abuse. Evidence also suggests that women may not identify as victims and survivors if asked directly; responses are higher if women are given specific examples of abuse to relate to. (6G:15) … Research by Refuge also shows that women fear for their physical safety following online CCB. Almost one in five (17%) said they felt afraid of being attacked or subjected to physical violence following tech abuse. Fifteen per cent felt their physical safety was more at risk, 5% felt more at risk of 'honour'-based violence and 12% felt afraid to leave the house. (6G.26) Some victims and survivors are left feeling uncomfortable online. 711 If victims and survivors disengage from online services because of CCB, this can have significant adverse effects including isolating them from family, friends, professional and social networks (thereby reducing their ability to access support). Being inaccessible online can in fact heighten abuse or amplify the risk of physical contact to enable the perpetration of abuse. (6G.28)

Ofcom's balancing and proportionality assessments should take this into account; so far the analysis has not recognised this.

The President of the Court of Human Rights recently noted that, "the victims of domestic and gender-based violence are not born vulnerable. They are rendered vulnerable, on their journey from girl to womanhood, by the imbalanced social structures into which they are born, by the law and by law-makers, and by attitudes and patterns of behaviour in their regard which are ignored, permitted or endorsed by society, including the State." She suggests that the focus of the Court "must remain the actions and omissions of State authorities" and suggests the key question "were the applicants accorded equal and sufficient protection before the law?" (Siofra O'Leary, Speech on the Opening of the Judicial Year 2024, Strasbourg, 26 January 2024 p5)

4. Trafficking – Article 4

Article 4 prohibits "slavery or servitude" in para (1), while Article 4(2) prohibits forced or compulsory labour. The Court has distinguished between these terms. (*S.M v Croatia* (60561/14) [GC], judgment 25 June 2020; *Rantsev v Cyprus and Russia* (25965/04), judgment 7 January 2010; *Siliadin v France* (73316/01), judgment 26 July 2005) In *S.M v Croatia*, the Court clarified that "forced or compulsory labour" covers serious exploitation, for example forced prostitution, irrespective of whether it is linked to trafficking. Trafficking in human beings, by its very nature, is based on the exercise of powers attaching to the right of ownership. It threatens the human dignity and other fundamental freedoms of its victims. While the prohibitions in Article 4(1) arguably relate to more serious infractions of fundamental rights, both Article 4(1) and Article 4(2) constitute rights that should be considered as such in any balancing of interests. Note also the recommendations from Council of Europe Group of Experts of Action against Trafficking in Human Beings (GRETA) in this regard: here.

Recommendation

Ofcom should review its recommendations in the light of their obligations, taking into account the weight of the rights violations as against companies' costs in particular.

Issue 6: disconnect between risk analysis and the recommended mitigation measures

Issue

Ofcom says in Volume 2 - "The Causes and Impacts of Online Harm" that it "presents our assessment of the causes and impacts of illegal online harms based on the evidence that we have gathered over the past three years. The analysis we set out here forms part of our duty under the Act to assess the factors that can cause a risk of harm to individuals on a service. We expect services to have reference to it when they carry out their own risk assessments."

Ofcom's assessment focuses on the over 130 priority offences defined in the Act (but not non-priority offences), grouped in Volume 2 under 15 broad kinds of illegal harm. Ofcom notes in its introduction to Volume 2 that "The impact of the harms we have looked at can be extremely severe. It is not limited to the online world but can also profoundly affect people's lives offline" and that "the kinds of illegal harm we have looked at occur on services of all types. Services as diverse as social media services, dating services, marketplaces and listings services, search services, adult services, and file-storage and file-sharing services are all used to disseminate some of the types of harmful content we have looked at in this volume. **Bad actors use both large and small services to spread illegal content**, although the way in which they use large services sometimes differs from the way in which they use small services." (our emphasis)

It also notes that "certain 'functionalities' stand out as posing particular risks", picking out in its introduction the following:

- End-to-end encryption
- Pseudonymity and anonymity
- Livestreaming
- Recommender systems

The introduction goes on: "We expect services to think about these risk factors when doing their risk assessments (see Volume 3). As we explain in Volume 4, **we have designed a number of the measures in our Codes of Practice to target high-risk service types and functionalities** … The role of the new online safety regulations is not to restrict or prohibit the use of such functionalities, but rather to get services to put in place safeguards which allow users to enjoy the benefits they bring while managing the risks appropriately."

Volume 2 is a commendable standalone document within the suite of documents that make up the illegal harms consultation. It brings together a vast amount of evidence as to how the illegal offences covered by the Act are prevalent online and is analytical and thorough in identifying the functionalities that contribute to this prevalence and/or risk of harm to individuals. Many of these functionalities are vectors for multiple harms.

However, this assessment does not flow through to the mitigation measures set out in the Codes of Practice (Annex 7) (for users to user services) and Annex 8 for search, which focus primarily on content takedown and measures to deal, ex-post, with illegal content once it has been identified. The rules-based nature of the Codes - specifying specific recommended measures rather than describing desired outcomes - and the fact that these are designed as a "safe harbour" (eg if companies follow the measures they will be judged to have complied with their duties under the Act*), means that there is no incentive for companies to implement mitigating measures beyond those described in the codes. This is the case even if their risk assessment has flagged that their service poses particular risks from other functionalities (arising from design choices) and despite the fact that the risk assessment notes the need for voluntary actions over and above what is set out in the codes. The Atlantic Council make this point: "if compliance replaces problem-solving, it establishes a ceiling for harm reduction, rather than a floor founded in user and societal protection" (p 36).

(*The "safe harbour" provision is described here: Services that choose to implement the measures we recommended in our Codes of Practice will be treated as complying with the relevant duty. This means that Ofcom will not take enforcement action against them for breach of that duty if those measures have been implemented. Service providers may seek to comply with a relevant duty in another way, but the Act provides that, in doing so, they must have regard to the importance of protecting users' right to freedom of expression within the law, and to the importance of protecting users from breaches of relevant privacy laws. Where providers do take alternative measures, they must keep a record of what they have done and explain how they think the relevant safety duties have been met. (Volume 4, para 11.7) This reflects s 49.)

Furthermore, smaller companies are in many instances exempt from implementing particular mitigating measures due to Ofcom's proportionality analysis.

We have produced a supporting document [annex A] to illustrate where the gaps between the analysis of harm and the recommended mitigations of it lie, along with a summary "at a glance" table. We have also recently published a blog discussing the choices made in relation to the codes of practice and compliance, which we also draw from below.

What the Act says

In Section 10(4), the Act says that the duty in 10(2) (to "take proportionate measures relating to the design and operation of the service" to "prevent individuals from encountering priority illegal content", to "effectively mitigate and manage the risk of the service being used for the commission or facilitation of a priority offence" and to "effectively mitigate and manage the risks of harm to individuals) and the duty in 10 (3) (to "operate a service using proportionate systems and processes designed to" "minimise the length of time for which any priority illegal content is present" and, when alerted, "swiftly take down such content") apply across all areas of a service "*including the way it is designed, operated as well as used*" and "*require the provider to take or use measures in the following areas, if it is proportionate to do so*":

> (a) regulatory compliance and risk management arrangements,
>
> (b) design of functionalities, algorithms and other features,
>
> (c) policies on terms of use,
>
> (d) policies on user access to the service or to particular content present on the service, including blocking users from accessing the service or particular content,
>
> (e) content moderation, including taking down content,
>
> (f) functionalities allowing users to control the content they encounter,
>
> (g) user support measures, and
>
> (h) staff policies and practices. (Section 10(4))

Section 236(1) of the Online Safety Act then describes "measures" as follows

> "any reference to a measure includes a reference to any system or process relevant to the operation of an internet service or any step or action which may be taken by a provider of an internet service to comply with duties or requirements under this Act"

In addition, Schedule 4 of the OSA sets out the approaches that Ofcom must take to drawing up the codes of practice. Under the General Principles, it says:

> (d) the measures described in the code of practice that apply in relation to Part 3 services of various kinds and sizes must be proportionate to OFCOM's assessment (under section 98) of the risk of harm presented by services of that kind or size.

Schedule 4 also includes the following at section 3: OFCOM must ensure that measures described in codes of practice are compatible with pursuit of the online safety objectives, which we have extracted at (page/para) above. As well as setting out a number of objectives relating to systems and processes in section 3(a), the objectives specify at 3(b):

> (b) 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.

Schedule 4 requires that the recommendations be clear and precise, but this does not mean that the service providers should have no freedom of choice.

Finally, Schedule 4 also requires that Ofcom ensure that (9(1)) Codes of practice that describe measures recommended for the purpose of compliance with a duty set out in section 10(2) or (3) (illegal content) **must include measures in each of the areas of a service listed in section 10(4)** (our emphasis).

As we can see above, 10(4) includes at (b) design of functionalities, algorithms and other features, all of which – as we set out below – are lacking measures in this first iteration of the codes.

The significance of the Codes is seen in section 49, which envisages two ways in which in-scope providers can comply with their relevant statutory duties: (a) compliance through recommended measures; and (b) compliance through alternative measures, but with caveats. Section 49 states that a service 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 which are recommended for the purpose of compliance with the duty in question".

This means that services that service providers which choose to implement measures recommended to them for the kinds of illegal harms and their size or level of risk indicated in the regulator's Codes will be deemed as compliant with the relevant duty and Ofcom will not take enforcement action for breach of that relevant duty against those services. The level and nature of Ofcom's recommendations are therefore significant for the level of safety provided to users and the extent to which the Act's objectives are achieved.

In the event of identifying potential risks in services that are not adequately addressed by the existing Codes, and where transparency measures prove ineffective, Ofcom has the authority to

update and enhance the Codes (see sections 47(1) and 48 of the Act) - a point which Ofcom recognises when it notes that the development of the Codes will be an iterative process. This, of course, has the disadvantage of introducing further delays to the effective implementation of the regime.

Schedule 4 provides further requirements about the measures to be included in any codes, as we discuss below.

Parliamentary debate

In Lords Committee stage day 1, the Government Minister Lord Parkinson said: "Through their duties of care, all platforms will be required proactively to identify and manage risk factors associated with their services in order to ensure both that users do not encounter illegal content and that children are protected from harmful content. To achieve this, they will need to design their services to reduce the risk of harmful content or activity occurring and take swift action if it does". (Column 725)

At Lord Committee stage day 3, in response to a debate on the nature of cumulative harm, Lord Parkinson said: The Bill will address cumulative risk where it is the result of a combination of high-risk functionality, such as live streaming, or rewards in service by way of payment or non-financial reward. This will initially be identified through Ofcom's sector risk assessments, and Ofcom's risk profiles and risk assessment guidance will reflect where a combination of risk in functionalities such as these can drive up the risk of harm to children. Service providers will have to take Ofcom's risk profiles into account in their own risk assessments for content which is illegal or harmful to children. **The actions that companies will be required to take under their risk assessment duties in the Bill and the safety measures they will be required to put in place to manage the services risk will consider this bigger-picture risk profile**. (Lords Committee stage 27 April 2023 Column 1385) (our emphasis)

Later in Lords Committee stage, when challenged by Baroness Morgan as to why the Government would not concede on a code of practice for women and girls, Lord Parkinson set out a number of reasons why the existing codes would be sufficient in this regard. He also made replied directly to Morgan's claim that the Bill "misses out the specific course of conduct that offences in this area can have" and referred to (then) clause 9 re services needing to mitigate and manage the risk of being used for the commission or facilitation of an offence.

Parkinson said: "This would capture patterns of behaviour. In addition, Schedule 7 contains several course of conduct offences, including controlling and coercive behaviour, and

harassment. The codes will set out how companies must tackle these offences where this content contributes to a course of conduct that might lead to these offences." (O)

Ofcom's proposals

In the light of the above, it is worth noting that there is a substantial disconnect between the tone and approach set out in Ofcom's "Approach" document and the detail that is buried in the subsequent volumes of the consultation.

In the Approach document, Ofcom says: "The Act is clear: first and foremost, the onus sits with service providers themselves, to properly assess the risks their users may encounter, and decide what specific steps they need to take, in proportion to the size of the risk, and the resources and capabilities available to them" (p4) Yet, when you reach Volume 4, the "safe harbour" of the codes for companies that follow them is set out: "Services that choose to implement the measures we recommended in our Codes of Practice will be treated as complying with the relevant duty". While this follows the terms of the Act, significantly Ofcom has in the main interpreted "measures described" as requiring very specific recommendations to which proportionality and costs criteria have to be applied on an individual basis before they can be "recommended for the purpose of compliance". Ofcom is pre-assessing proportionality here to limit the scope of the measures recommended, rather than allowing services to make their own assessments.

We submit that this approach is not required by the Act and does not reflect Parliamentary intention. One implication of section 236(1) in this context is that the obligations to take or use measures – notably those set out in non-exhaustive lists under sections 10(4), 12(8) for user-to-user as well as 27(4), 29(4) for search services - are not limited to specific types of technology but extend to processes as well. Ofcom's recommended measures, mapped against the relevant duties in indexes of recommended measures - see Annex 7, pp 6-10 (U2U) and Annex 8, pp 6-9 (search) – include not only tech, e.g., hash matching for CSEAM and keyword detection regarding articles for use in frauds, but also some process-driven recommendations, like internal monitoring and assurance, dedicated reporting channels etc.

A requirement for an obligation to be clear and precise (Schedule 4, para 2b) does not mean that a service provider should no choice or discretion in responding to the obligation; rather what it means is that the service provider should be able to understand the nature of the requirement. Ofcom is not precluded from imposing process requirements and offering illustrative examples of good or best practices when making recommendations of a procedural nature.

Indeed, it is arguable that Ofcom could make more use of objective-focussed process obligations to cover gaps in mitigations that are currently found in the recommended measures. There are many instances where a functionality has been found to be problematic in Vol 2 and for the purposes of the risk register, but where Vol 4 finds the evidence of those solutions not to be specific enough to justify making a specific technical recommendation.

An approach based on broader process-based obligations orientated towards the Act's objectives could also be within the scope of Section 49(1) which would allowed a much more flexible orientation towards user safety while still satisfying the requirements for clarity and precision and allowing for proportionality of response.

As we set out in section 3, throughout the consultation document, Ofcom makes its own judgements – without qualification – about a) what evidence it deems to be acceptable to support the inclusion of measures in the codes of practice (we talk further about evidence thresholds in section 3, above); and b) what measures it deems proportionate for services to implement to mitigate the harms they may have already identified in their risk assessment. While there is some methodology set out in Volume 2 about what evidence they have accepted for the purpose of the risk register, for Volume 4 (the codes) there is no equivalent. This is a different issue from when the threshold has been reached - and why.

The wording of the Act, however, does not imply that this is for Ofcom to judge – rather that it is for providers to "take or use measures … if it is proportionate to do so" (s 10 (4)).

Ofcom's judgement on this leads to a rules-based, prescriptive, de minimis approach to safety, which does not take into account the fact that the Act itself says the illegal content duties apply across all areas of the service "*including the way it is designed, operated as well as used*" and that the duties "*require the provider to take or use measures" in areas, including "regulatory compliance and risk management arrangements", "design of functionalities, algorithms and other features".* On the impact of proportionality, we refer to Section 4.

There are multiple references to the fact that the codes are iterative (eg Vol 3 8.16), the approach is novel (Vol 4 11.15) and the evidence base is incomplete (Vol 4 11.16) leading to Ofcom's description that the draft "first Codes aim to capture existing good practice within industry and set clear expectations on raising standards of user protection, especially for services whose existing systems are patchy or inadequate. Each proposed measure has been impact assessed, considering harm reduction, effectiveness, cost and the impact on rights." (Chapter 11, Introduction p3)

We understand that Ofcom is taking a cautious approach, that it is reliant on evidence and that its proportionality assessment is stringent. However, there is a fundamental choice that has been made here about the approach to the codes that does not fit with the legislative intent: the regime was supposed to be principles based or risk based. While Schedule 4, para 1(a) does require Ofcom to "consider the appropriateness of provisions of the code of practice to different kinds and sizes of Part 3 services and to providers of differing sizes and services", it does not have to pre-judge all the measures it recommends on that basis nor is it required to set down specific rules. While there are expectations that obligations should be clear (and not impose unnecessary obligations on service providers) this does not mean more general obligations cannot be imposed. Indeed, as Lord Parkinson remarked;

> Ofcom's guidance and codes of practice will set out how they can comply with their duties, in a way that I hope is even clearer than the Explanatory Notes to the Bill, **but certainly allowing for companies to have a conversation and ask for areas of clarification**, if that is still needed. (our emphasis) (Lords Committee stage 25 April 2023)

It is reasonable as the regulator to place an expectation on the companies to respond to outcome-defined obligations.

Ofcom's Economic Director, Tania Van Den Brande set out the problems with a rules based approach in 2021:

> "..rules are at a greater risk of leading to undesirable effects if a given conduct can be harmful, neutral or beneficial depending on the circumstances of the market or the characteristics of the firm they apply to. ... Rules can also become outdated in highly dynamic markets."

Despite the amount of evidence Ofcom has collected on the nature of harm, its decision to follow a rules-based model of recommendations has significantly limited the likelihood that companies will take a risk-based approach to mitigation. Furthermore, the rigid rules-based approach then requires Ofcom to decide, based on its proportionality assessment, that it should exempt smaller services from following those rules – rather than specifying an outcome or a principle and judging whether the regulated service has acted proportionately in its response. We discuss the issue relating to small companies further in section 7; but deciding whether or not to apply code of practice measures to all companies, based on Ofcom's own assessment of the "onerous" impact they might have on their profitability, is entirely inconsistent with Ministerial expectations that the illegal content duties would apply to all regulated services, regardless of size – with the proportionality test being for companies to

judge and account for to Ofcom, rather than Ofcom making that decision for them upfront. (For example, in the volume 3 discussion on automated content moderation systems: "if we based our recommendations on the limited evidence we do have, which would be drawn from the practices of larger, mainly social media services, we could drive smaller services or non-social media services to adopt practices or seek to achieve outcomes that are not appropriate for their services. This may also result in more onerous expectations on smaller services which may not have the resources or need to match solutions of larger services, potentially undermining their ability to operate, with implications for competition and innovation more widely" (Vol 3. 12.31))

Evidence

We set out our evidence on this disconnect between the harms identified and the measures proposed to address them in annex A, which is attached to this submission as a PDF and which can be found on our website here.

Recommendation

We refer Ofcom back to our recommendation in section one which sets out additional measures to be added to the draft codes of practice which require companies to take mitigating measures based on the risks arising from their services' "functionality, algorithms and features" that they have identified in their risk assessment. This would provide a "catch-all" approach to improving the effectiveness of the illegal content duties in this first iteration, rather than waiting for further evidence to be provided and assessed to support the later insertion of specific individual measures.

We include the wording we propose again for completeness.

> "Design of functionalities, algorithms and other features
>
> *Product testing*
>
> For all services, suitable and sufficient product testing should be carried out during the design and development of functionalities, algorithms and other features to identify whether those features are likely to contribute to the risk of harm arising from illegal content on the service.
>
> The results of this product testing should form a core input to all services risk assessments.

Mitigating measures

For all services, measures to respond to the risks identified in the risk assessment should be taken, including but not limited to, providing extra tools and functionalities, by redesigning the features associate with the risks, by limiting access to them where appropriate or where the risk of harm is sufficiently severe by withdrawing the function, algorithm or other feature.

Decisions taken on mitigating measures, as part of the product design process or as a response to issues arising from the risk assessment, should be recorded. (Note: this would be included in the record keeping duties under section 23 (u2U) and section 34 (search).)

Monitoring and measurement

All services should develop appropriate metrics to measure the effectiveness of the mitigating measures taken in reducing the risk of harm identified in the risk assessment. These measures should feed back into the risk assessment."

Issue 7: small vs large companies makes size rather than risk the primary aspect

Issue

While the illegal content duties apply to all regulated services (regardless of size), Ofcom's recommended measures in the codes of practice do not apply equally to all of them. Instead, they are differentiated according to size and then differentiated further based on the services' own risk assessments. Ofcom's tear sheet sets out "at a glance" its proposals and who they apply to. Ofcom has recently published an explainer which, in part, responds to concerns that had been raised with them during the consultation process about their approach and which stresses (again) the iterative nature of the codes. As with their chosen approach to mitigating measures, which we set out in section 6, we are concerned that this means a "lowest-common denominator" baseline for the codes when they come into force – and one which in many areas may even risk weakening existing protections.

We also do not think that their approach to proportionality and size is justified by the legislative framework nor reflects the intention of Parliament. It also runs counter to the approach – agreed late in the Online Safety Bill's passage – to designating "category 1" services by taking account of both size OR functionality/risk – eg, bringing small, high-harm platforms into scope of the duties, on which Ofcom will be consulting at a later stage. We discussed proportionality with reference to cost above in section 4.

What the Act says

Section 7(2) says All providers of regulated user-to-user services must comply with the following duties in relation to each such service which they provide—

(a)the duties about illegal content risk assessments set out in section 9,

(b)the duties about illegal content set out in section 10(2) to (8)

Section 9 sets out the risk assessment duties (which we discuss further below/above – cross referencing etc)

Section 10 sets out the safety duties for user-to-user services (section 27 sets out similar, but not exactly comparable duties for search). Section 10(4) for user-to-user services, which we discuss above, sets out the types of measures providers may be "required" to take "if it is proportionate to do so". Section 10(10) for user-to-user and section 27(10) for search define "what is proportionate for the purposes of this section", stating that "the following factors, in particular, are relevant—

(a)all the findings of the most recent illegal content risk assessment (including as to levels of risk and as to nature, and severity, of potential harm to individuals), and

(b)the size and capacity of the provider of a service.

Section 91 also inserts into Section 3 of the Communications Act 2003 new duties on Ofcom, including:

"(4A) …. OFCOM must have regard to such of the following as appear to them to be relevant in the circumstances—

(a)the risk of harm to citizens presented by regulated services;

(b)the need for a higher level of protection for children than for adults;

(c)the need for it to be clear to providers of regulated services how they may comply with their duties set out in Chapter 2, 3, 4 or 5 of Part 3, Chapter 1, 3 or 4 of Part 4, or Part 5 of the Online Safety Act 2023;

(d)the need to exercise their functions so as to secure that providers of regulated services may comply with such duties by taking measures, or using measures, systems or processes, which are (where relevant) proportionate to—

(i)the size or capacity of the provider in question, and

(ii)the level of risk of harm presented by the service in question, and the severity of the potential harm;

Parliamentary debate

Throughout the development of the Bill, Government Ministers were at pains to stress that all platforms would be covered by the illegal content duties. Here, for example, is former DCMS Minister Chris Philp at the Second Reading of the Bill in the Commons in April 2022: "all platforms, regardless of size, are in scope with regard to content that is illegal and to content that is harmful to children. (Hansard link here) When the Bill had its Second Reading in the Lords the following February, Lord Parkinson said in his opening statement: "All companies in scope will be required to tackle criminal content and activity online. If it is illegal offline; it is illegal online. All in-scope platforms and search services will need to consider in risk assessments the likelihood of illegal content or activity taking place on their site and put in place proportionate systems and processes to mitigate those risks. Companies will also have to take proactive measures against priority offences. This means platforms will be required to take proportionate steps to prevent people from encountering such content." (Hansard 1 February 2023 col 687)

As we can see from the duties in the Act above, there is much stress on "proportionate" measures – which Government Ministers, in Parliament, were also at pains to emphasise when challenged on the number of businesses that were potentially within scope of the legislation.

For example, Lord Parkinson – in response to an amendment proposed by Baroness Fox, to exempt small services – said the following at Lords Committee stage:

> "My Lords, I am sympathetic to arguments that we must avoid imposing disproportionate burdens on regulated services, but I cannot accept the amendments tabled by the noble Baroness, Lady Fox, and others …. The current scope of the Bill reflects evidence of where harm is manifested online. There is clear evidence that smaller services can pose a significant risk of harm from illegal content, as well as to children, as the noble Baroness, Lady Kidron, rightly echoed. Moreover, harmful content and activity often range across a number of services. While illegal content or activity may originate on larger platforms, offenders often seek to move to smaller platforms with less effective systems for tackling criminal activity in order to circumvent those protections. Exempting smaller services from regulation would likely accelerate that process, resulting in illegal content being displaced on to smaller services, putting users at risk.
>
> … The Bill has been designed to avoid disproportionate or unnecessary burdens on smaller services … Ofcom's guidance and codes of practice will set out how they can comply with their duties, in a way that I hope is even clearer than the Explanatory Notes to the Bill, but certainly allowing for companies to have a conversation and ask for areas of clarification, if that is still needed. They will ensure that low-risk services do not have to undertake unnecessary measures if they do not pose a risk of harm to their users."

Despite that recognition, it is also clear that proportionality was not intended as a vehicle to undercut protection; rather it acknowledged the need to recognise the risk of harm posed by the service.

There is another important part of the Parliamentary and legislative context to take into account here. At a late stage in the Parliamentary passage of the Online Safety Bill, Baroness Nicky Morgan won a hard-fought concession from the Government that the threshold for category 1 services (to whom the so-called "Triple Shield" duties would apply) would not be set based on size AND functionality but could include size OR functionality or a combination of both – in effect, accepting

that a large size was not the only determinant of risk. In introducing the amendment at Commons Consideration, DSIT Minister Paul Scully said:

> "The Government are grateful to Baroness Morgan of Cotes and my right hon. and learned Friend the Member for Kenilworth and Southam (Sir Jeremy Wright), who like many in the House have steadfastly campaigned on the issue of small but risky platforms. We have accepted an amendment to the Bill that changes the rules for establishing the conditions that determine which services will be designated as category 1 or category 2B services and thus have additional duties. In making the regulations used to determine which services are category 1 or category 2B, the Secretary of State will now have the discretion to decide whether to set a threshold based on the number of users or the functionalities offered, or both factors. Previously, the Secretary of State was required to set the threshold based on a combination of both factors. It is still the expectation that only the most high risk user-to-user services will be designated as category 1 services." (Hansard 12 September 2023 Col 807)

This amendment is now included in the Act at Schedule 11, Section 1 (4): Regulations under this paragraph must specify the way or ways in which the relevant threshold conditions may be met, and that may be by meeting the conditions in any specified combination, subject to the rule that—

> (a)in relation to the Category 1 threshold conditions and the Category 2B threshold conditions, at least one specified condition about number of users or functionality must be met, and
>
> (b)in relation to the Category 2A threshold conditions, at least one specified condition about number of users must be met.

The implementation of the categorisation thresholds is not in scope for Ofcom's consultation at present. We set out its significance in this blog post. However, this earlier statement from Lord Parkinson at Lords Report stage on 19 July, where he was still resisting conceding on Baroness Morgan's amendment, is instructive as to the lack of size differentiation intended for companies complying with the illegal harms duties:

> "I will say more clearly that small companies can pose significant harm to users—I have said it before and I am happy to say it again—which is why there is no exemption for small companies… All services, regardless of size, will be required to take action against illegal content, and to protect children if they are likely to be accessed by children. This is a proportionate regime that seeks to protect small but excellent platforms from overbearing regulation. **However, I want to be clear that a small platform that is a font**

of illegal content cannot use the excuse of its size as an excuse for not dealing with it." *(our emphasis)*

We see below that – in its draft proposals – Ofcom is indeed, from the outset of the regulatory regime, giving small companies many excuses for not dealing with illegal content.

Ofcom's proposals

Despite the very strong commitments from the Government, Ofcom is exempting small and/or single risk services from many of the measures in the codes on the grounds of proportionality and cost. This compounds the fact that these services are also in effect let off carrying out a robust risk assessment: if they don't assess their own risk adequately (meaning risks might be under assessed resulting in a lower risk classification for Ofcom's framework), and they also don't have to comply with all the measures in the codes, the small-but-risky services will not be required to address the illegal content duties effectively.

Ofcom sets out its rationale for this in Volume 4 as follows:

- 11.16: "Many of the measures we propose are for large services. This is often because we do not yet have enough information on the potential costs and benefits to know whether the measures are proportionate for smaller services at this point. As our understanding develops, it may be appropriate in future iterations of the Codes to expand the range of services for which some measures are recommended."
- 11.50 "We propose that many Codes measures are only recommended for 'large services'. We often apply these recommendations to services that have also identified high or medium risk of certain illegal harms".
- 11.51 "**We propose to define 'large service' as a service with a number of monthly UK users that exceeds 7 million …** "
- 11.51 "Part of the reason for recommending some measures only for large services is that the benefits of measures are likely to be greater for such services, because more users will be protected by the measure … Another reason for restricting measures to large services is that for some proposed measures the costs for services may be significant, and those costs could have a material effect on the operation of non-large services."
- 11.53: "We assume that in general large services have more resources available to undertake measures than smaller services. **We consider it can be prudent to exempt smaller services from incurring those costs (where appropriate provided they are not high risk), as there will often be significant uncertainty in any assessment of benefits and costs, and we want to reduce the possibility of imposing financially damaging costs on businesses when the magnitude of benefits expected to result from the**

measure is uncertain. Also, in some cases, imposing costly measures on smaller services could reduce their ability to operate and compete effectively in certain online markets. A lack of competition and innovation can be very costly for society and needs to be considered against the scale of potential benefits from any measure." (our emphasis)

This position was further consolidated in its recent explainer document: "In the first version of the codes, we have recommended measures only to large services where we do not yet know whether it is proportionate to extend a measure to smaller services. **This can be because of uncertainty on whether a measure is effective enough to reduce harm materially on smaller services, or whether the costs to these businesses or the inconvenience to their users might be disproportionate to the harm the measure can address.** As our understanding develops, we might recommend measures for a wider range of services." (our emphasis)

In all this the human suffering and the long-terms costs to society are taken as an acceptable price for companies operating profitably.

There are further decisions made by Ofcom throughout the consultation documents that are not presented for consultation but rather framed as a logical, evidence-based fait accompli. For example, the list of regulated services in volume 1 is presented as a final list but very likely omits important sites of severe harms, such as suicide sites, revenge porn collector sites and sites focused on hate and extremism.

Ofcom does not present a definition of small companies and uses as a definition of "large" an equivalent to the DSA definition VLOPs – 7 million monthly users in the UK (vol 4, 11.51). That is 10% of the UK population and might be the sort of numerical threshold relevant when considering Category A services, which imposes further duties. It is less clear this very high threshold is appropriate for implementing measures in relation to illegality. Additionally, reliance on a numerical perspective is problematic. Using either profitability or the size of the user base to define risk of harm excludes from mitigating action the types of harm that minority or intersectional groups might experience from smaller sites that are designed to target them.

Elsewhere, Ofcom's justification for a differential obligation between small and large companies seems based on what they do already (e.g. large companies do more already) and the impact of the harmful consequence. This is a quantitative assessment of harm - how many people are harmed, not how badly they are hurt, and therefore is not well framed to assess the impact of small, single issue services.

Ofcom do introduce a "floor" relating to the most harmful, illegal, egregious activity – such as CSAM – where smaller services with a specific risk are required to follow measures in the code relating to these risks. But then the next level up in terms of requirements on small services is for those that are "multi-risk" – which means that to be required to take those specific measures (enhanced user control measures, for example), companies need to be at medium or high risk of "at least two" of a limited list of "kinds of illegal harm".

Ofcom makes this assumption – without evidence – that single-risk sites cause less harm and therefore consequential requirements to take mitigating measures will also have less benefit:

> "We intend these measures to apply to services that face significant risks for illegal harms in general. There is a question over what it means for a service to have such risks. One option would be to recommend these measures to services that have identified as medium or high risk of at least one kind of illegal harm. **However, where services only identify a risk of a single kind of illegal harm, the benefits of these measures to address all harms will be lower.** This is partly because if services have only identified a single area of risk, the extent of harm will tend to be lower compared to if they have identified a range of kinds of offence where they are high risk.

Moreover, it assumes – again without evidence – that single-risk sites will know all about that risk and will be taking measures to address it.

> "It is also partly because many of these measures are about enabling services to have a good understanding of their risks and of the content moderation policies needed to address those risks. **If a service was only of medium or high risk for a single kind of illegal harm, the risk is more likely to be well understood across the organisation, such as the risk of fraud for some marketplace services.** This tends to mean the benefits of these measures in terms of improving understanding and consistency of approach are smaller than if there were multiple areas of risk. The case for the measures to address all harms being proportionate therefore tends to be stronger if we only apply them to services that have identified multiple kinds of illegal harm." (Vol 4, 11.44)

As a result of these judgements, we see that the only measures in the illegal content code of practice (annex 7) that apply to all U2U services, regardless of risk or size, are:

- 3B – named person accountable to most senior governance body
- 4A – content moderation systems or processes designed to take down illegal content swiftly
- 5 – 8 of the 10 measures reporting to reporting and complaints
- 6A&B – terms of service measure

- 10A – removal of terrorist accounts

The only measures in the illegal code of practice (annex 8) that apply to all search services, regardless of risk or size are:

- 3B – named person accountable to most senior governance body
- 4A – systems and processes designed so that illegal content is deprioritised
- 4G – URLs identified as CSAM are deindexed
- 5 – 8 of the 10 measures reporting to reporting and complaints
- 6A&B – publicly available statements

This internal logic that flows from Ofcom's differentiation of services ("large" or "multi-risk" vs "small" or "single risk") then further compounds the application of what would otherwise seem to be proportionate measures for dealing with illegal content, beyond CSAM. For example, measure 2 (**setting internal content policies having regard to at least the findings of their risk assessment and any evidence of emerging harms on their service)** only applies to services which are large or multi-risk, which means that the subsidiary measures sitting underneath this only apply to services that meet those criteria too. This further limits the potential impact of measures that Ofcom provides evidence will be effective and exempts smaller or single-risk services (even when that risk and the resulting harm is serious) from adopting these approaches based on the likely cost.

Ofcom, in para 12.114-115 says in relation to setting performance targets for content moderation measures: "we consider that services that follow this measure are more likely to operate effective content moderation systems. As we have shown, the evidence suggests that effective content moderation plays a hugely important role in mitigating the risk of harm to users meaning the measure would have important benefits. As with measure 2, these benefits will be greatest for services that are either large or multi-risk … We are not proposing to recommend this measure for smaller and lower risk services, because it is less clear the benefits are great enough given the lower volume of content such services need to assess."

In Volume Vol 4, 12.171, this then leads to the following judgement about resourcing content moderation functions effectively: "We are not at this point proposing extending the proposal to services that are not large and are not multi-risk. The amount and diversity of content such services need to moderate is likely to be materially lower and the benefits would therefore be materially smaller, making it questionable whether the potentially substantial costs of the measure were always justified for such services. **Moreover, this measure is predicated on services having the internal content policies of our proposed Measure 2 above and the**

performance targets we propose in Measure 3, so it makes sense for this measure to apply to the same set of services as those proposed measures are recommended for." (our emphasis)

Ofcom sets out in Volume 4 (11.47) that "We are required under the Act to consider the impact of our proposed measures on small and micro businesses". This requirement is at section 93(4) which amends the Communications Act:

> After subsection (4) insert—
>
> > "(4A) An assessment under subsection (3)(a) that relates to a proposal mentioned in subsection (2A) must include an assessment of the likely impact of implementing the proposal on small businesses and micro businesses.
> >
> > (4B) An assessment under subsection (3)(a) that relates to a proposal to do anything else for the purposes of, or in connection with, the carrying out of OFCOM's online safety functions (within the meaning of section 235 of the Online Safety Act 2023) must, so far as the proposal relates to such functions, include an assessment of the likely impact of implementing the proposal on small businesses and micro businesses.)

In Ofcom's explanation of how the costs of compliance with the safety duties may affect multi-risk small or micro-businesses, it sets out a clear (if slightly apologetic in tone) justification of the measures which may cause some sites to cease operating: "To give a sense of the scale of the costs, we are aware of a small service which needed to increase spending for online safety by several £100,000s per annum to deal with problematic content on its service relating to more than one harm area, where some of this material was illegal. This cost is principally driven by content moderation costs, and the number of (human) moderators engaged. This suggests that the costs that some small and micro business will need to incur could be substantial. **Some small and micro businesses with significant risks may struggle to resource the recommendations we propose for them.** Services would have the option of not following our Codes and describing how they have met their duties under the Act in another way, but they would still need to meet their safety duties in the Act. **That such services need to incur costs if they are not already undertaking suitable measures is inevitable given the significant and important new safety duties that the Act places on them. It is even possible that some such services may cease to operate in the UK, or cease to operate at all if the UK is an important market for them**." (Volume 4 23.21)

Ofcom however offers no justification, despite compelling evidence of harm, as to why the same compliance should not be required from small, single-risk targeted small or microbusinesses, notwithstanding the necessity for a greater overall focus on safety by design across all services regardless of size and the Government's assessment that safety by design is a

more cost effective route for such services. Despite the evidence amassed in the risk profile section, this doesn't dock into assessments on proportionality. If a small platform has one significant area of risk like fraud, why would they not be expected to undertake the same range of risk assessments as the larger platforms?

Evidence

The "number of monthly users" threshold is set so high as to likely exclude a number of prominent sites, including Fortnite (average UK monthly users 4 million – 5% of the global average of 80m) and Roblox (average UK monthly users) 3.4m. Both sites have been linked with illegal activity in the past, including money laundering and grooming while the role that gaming sites can play in radicalisation and extremism has been well documented. Roblox is currently facing a class action brought by parents in California relating to grooming and sexual content.

There is increasing evidence of the direct offline harm caused by dedicated, single-risk sites. For example:

- groupings of providers that do not have a distinct legal form or are shell companies and therefore can reconstitute themselves as different sorts of legal entities with different URLs or websites (eg marketplaces for suicide methods that are repeatedly taken down and re-emerge, evading regulatory intervention; here and here);
- small sites that have a single purpose that is extremely harmful to some groups, often with targeting of individuals - eg revenge porn collector sites (for example, here and here);
- dedicated hate and extremism sites, such as those researched in relation to inceldom by CCDH here and covered in this Parliamentary submission; far-right ideologies investigated by Hope Not Hate here and here; and extremism in this ISD report.

In relation to the concern about small suicide sites and message forums that sit behind URLs, the ICO has had to cope with some of this in the UK with cold calling companies going into insolvency the moment the ICO goes after them with regulatory measures (in the ICO's case mainly fines) but then the person behind the company pops up again with another company and carries on doing the same thing. You could have a forum that then changes its name slightly but has the same people behind it. Who is the provider (see s 226(3) on this) and more specifically can Ofcom keep a track of them? The enforcement plan does not seem to consider this issue (and that of 'refusenik' sites) in general.

The differential requirements relating to even core expectations such as content moderation is surprising given how central this function is to the duties in the Act – and how its under-resourcing in

even the largest platforms has been evidenced to cause harm. For example, the need to ensure that content moderation is sufficiently resourced was acknowledged – but not actioned – by Meta in relation to a number of harmful areas of content in 2020: "Internal documents make clear that, despite all of these internal studies demonstrating harm from its platforms, Meta refused to devote sufficient resources in order to address the problems, despite its public statements to the contrary. An August 2020 email cited a "severe lack of capacity for restricted content," including "nudity, graphic violence, child safety and SSI content review:" (New Mexico Attorney-General court filings)

Recent research by [redacted] has shown how moderation is badly enforced on Snapchat and the impact this has on the scale and prevalence of illegal content seen by children: "The children who took part in this research assumed that Snapchat was a less moderated space than other social media platforms, based on the fact that this is the platform where they see higher proportions of violent and sexual graphic content shared. Adults working with these children, for example youth workers interviewed for the project, had similar assumptions because they rarely, if ever, witnessed or heard about users being banned, or content being blocked or taken down. Because most content disappears after being viewed, or after 24 hours, several children said they and their peers perceived Snapchat to be a "safer" place for people to share illicit or illegal content. Some children were aware that it was possible that they or other users could report content to Snapchat, but none said they had ever done so. Some said that they would be very reluctant or even fearful of reporting content on Snapchat, because they could potentially face repercussions if other people in their local area found out they had been a "snake" or a grass".

We also refer to the extracts from the X/Twitter Australian transparency reports covered in section 4, above.

Recommendation

We would emphasise that this consultation focuses on illegal content – including CSAM, terrorism and material relating to a series of criminal offences that cause immense harm with potentially long-lasting repercussions to victims. We therefore see no justification in the decontextualised differentiation Ofcom has chosen to make between (very) large services and everything else. We recommend that Ofcom review its definition of proportionality to ensure that all services, regardless of size, are required to take measures that will address the risks they have identified in their risk assessment if they correspond to one or more of the risks set out in the risk register. We also recommend that Ofcom remove the differentiation based on size that it has applied to the specific measures recommended in the codes of practice and require services instead to decide on – and justify to Ofcom – whether their adoption of these measures is proportionate to the risks posed by their services.

We include the wording we propose in section one, above, for addition to the draft codes again for completeness here, as we recommend that this applies to all services regardless of size.

> "Design of functionalities, algorithms and other features
>
> *Product testing*
>
> For all services, suitable and sufficient product testing should be carried out during the design and development of functionalities, algorithms and other features to identify whether those features are likely to contribute to the risk of harm arising from illegal content on the service.
>
> The results of this product testing should form a core input to all services risk assessments.
>
> *Mitigating measures*
>
> For all services, measures to respond to the risks identified in the risk assessment should be taken, including but not limited to, providing extra tools and functionalities, by redesigning the features associate with the risks, by limiting access to them where appropriate or where the risk of harm is sufficiently severe by withdrawing the function, algorithm or other feature.
>
> Decisions taken on mitigating measures, as part of the product design process or as a response to issues arising from the risk assessment, should be recorded. (Note: this would be included in the record keeping duties under section 23 (u2U) and section 34 (search).)
>
> *Monitoring and measurement*
>
> All services should develop appropriate metrics to measure the effectiveness of the mitigating measures taken in reducing the risk of harm identified in the risk assessment. These measures should feed back into the risk assessment."

Issue 8: governance and risk assessment

Issue

In Ofcom's Approach document, improved governance is named as one of their key priority outcomes:

> "Specifically, we anticipate implementation of the Act will ensure people in the UK are safer online by delivering four outcomes (Figure 1):
>
> - stronger safety governance in online firms" (p7)

Yet, there seems to be a significant reliance in Ofcom's proposals on what platforms are already doing in terms of what they assess might be possible and/or should be recommended. It is not clear that Ofcom has determined that what these platforms are doing is a) effective; and b) enough to deliver their duties under the OSA. This links to the burden of proof point we set out above in section 3. We are also concerned that the risk assessment process is not effectively orientated towards outcomes (ie increased safety).

In the draft Guidance (annex 5), some of the outcomes for the steps in the risk assessment seem to go to process (eg you will have read this document) rather than objectives of the process (identified relevant risks)? Again, it is predicated (along with governance proposals) on the basis that companies are doing this already and therefore won't need to incur more costs.

Governance structures, along with robust risk assessment processes, are fundamental to influencing product design choices with a view to reducing the risk of harm. So, Ofcom's proposals here are crucial to the overall effectiveness of the Online Safety Act regime.

What the Act says

The risk assessment duties are at section 9 for User to User and section 26 for Search. Regulated services are required to carry out a "suitable and sufficient" illegal content risk assessment, keep it up to date and redo it "Before making any significant change to any aspect of a service's design or operation." For User-to-User services, section 9 (4), requires that the risk assessment to take into account "the risk profile that relates to services of that kind" —

> (a) the user base;
>
> (b) the level of risk of individuals who are users of the service encountering the following by means of the service—
>
> (i) each kind of priority illegal content (with each kind separately assessed), and

(ii) other illegal content,

taking into account (in particular) algorithms used by the service, and how easily, quickly and widely content may be disseminated by means of the service;

(c) the level of risk of the service being used for the commission or facilitation of a priority offence;

(d) the level of risk of harm to individuals presented by illegal content of different kinds or by the use of the service for the commission or facilitation of a priority offence;

(e) the level of risk of functionalities of the service facilitating the presence or dissemination of illegal content or the use of the service for the commission or facilitation of a priority offence, identifying and assessing those functionalities that present higher levels of risk;

(f) the different ways in which the service is used, and the impact of such use on the level of risk of harm that might be suffered by individuals;

(g) the nature, and severity, of the harm that might be suffered by individuals from the matters identified in accordance with paragraphs (b) to (f);

(h) how the design and operation of the service (including the business model, governance, use of proactive technology, measures to promote users' media literacy and safe use of the service, and other systems and processes) may reduce or increase the risks identified.

A smaller set of factors are included at section 26(4) for search.

Parliamentary debate

The prominence of the risk assessments in the Government's intentions for the regulatory regime are seen in, for example, Lord Parkinson's statement at Lords Report on 6 July 2023:

> "That is why the legislation takes a systems and processes approach to tackling the risk of harm. User-to-user and search service providers will have to undertake comprehensive mandatory risk assessments of their services and consider how factors such as the design and operation of a service and its features and functionalities may increase the risk of harm to children. Providers must then put in place measures to manage and mitigate these risks, as well as systems and processes to prevent and protect children from encountering the categories of harmful content." (Hansard 6 July 2023 col 1384)

Also, "the list of functionalities in the Bill is non-exhaustive. There may be other functionalities which could cause harm to users and which services will need to consider as part of their risk assessment duties. For example, if a provider's risk assessment identifies that there are functionalities which risk causing significant harm to an appreciable number of children on its service, the Bill will require the provider to put in place measures to mitigate and manage that risk. (Hansard 6 July col 1382)

Elsewhere, in part of a debate on end-to-end encryption, Lord Parkinson referred to the fact that "companies will need to undertake risk assessments, including consideration of risks arising from the design of their services, before taking proportionate steps to mitigate and manage these risks. Where relevant, assessing the risks arising from end-to-end encryption will be an integral part of this process". He went on to say that the risk assessment process used in "almost every other industry" and said that "it is right that we expect technology companies to take user safety into account when designing their products and services" (Col 1320).

Ofcom's proposals

Ofcom says: "Governance and accountability underpin the way that a service manages risk and ensures that efforts to mitigate them are effective. We consider that these processes are essential components of a well-functioning system of organisational scrutiny, checks and balances, and transparency around risk management activities. Effective governance and accountability processes should be effective in tackling all priority illegal harms." (Volume 3 8.13)

In Volume 3 of Ofcom's consultation, there are frequent references to the responses received from industry to their 2022 call for evidence and they cite, in many instances, evidence provided to them by large and prominent platforms – including X – that shows they are "already" doing much of what is required by the Act.

For example*: "*Responses to our 2022 Illegal Harms Call for Evidence demonstrated that several online services already have arrangements whereby they have a dedicated accountable staff member for regulatory compliance with online safety outcomes. This included Mojeek, Google, Trustpilot, X and Glassdoor, which all described overall ownership for online safety compliance at a senior manager level.*"* (para 8.58)

It is promising that many big platforms can point to existing governance structures (and there are likely to be plenty of platforms and smaller services who won't be able to do this). But as yet neither we nor Ofcom know the extent to which these existing governance structures are working effectively. For example, Facebook has run into trouble in the past with investors about

its oversight structures for risk. This is not a reason to discard what is there of course, but equally Ofcom should not assume that it is sufficient.

In Volume 4 of Ofcom's consultation documents, which focuses on the codes of practice, they include: "*Governance and accountability arrangements around the management of online safety risks, including senior management visibility of and accountability for key risks" and go on to "propose to recommend that all services establish clear accountability for compliance with their illegal content safety duty, complaints and risk assessment obligations, with additional expectations on large and multi-risk services*."

So how is Ofcom going to assess whether the structures, policies and accountability processes that already exist are sufficient? How will they measure their effectiveness? Is it enough for companies just to say they are doing it? How will they reassess the baseline? There are no examples of how improvements will be measured – either in risk assessment or mitigation (codes). It is also not clear what the difference is between accountability, responsibility and identifiability in relation to governance and senior management roles. (See Eg Volume 3 8.48, 8.52, 8.53, 8.54, 8.83.) Nor is it clear what the written statements of responsibilities will achieve, when Ofcom is primarily citing current practice. E.g. "Several services suggested in their responses to our 2022 Illegal Harms Call for Evidence that they already specify responsibilities for senior members of staff in relation to online safety and risk management". (8.68) As noted, there are concerns about the current levels of practice in even the large service providers.

Ofcom cites examples of risk assessment best practice, but these are largely focused on reputational risks and external risks to the company, not product safety and design risks created by their own products and services. See, eg, Table 9.1, 9.44 "comprehensive risks faced by an organisation".

Despite the Act specifying that both functionalities and the "design and operation of the service (including the business model, governance, use of proactive technology, measures to promote users' media literacy and safe use of the service, and other systems and processes) may reduce or increase the risks identified", any consideration of design in missing from Ofcom's proposed risk assessment process:

> "In our draft detailed guidance on methodology, we have proposed a process which reflects these four steps: i) understand the harms; ii) assess the risks; iii) decide measures, implement and record; and iv) report, review and update the risk assessment. We also include key common concepts from best practice which align to the risk assessment duties, such as: a) Assessing risk through a matrix of likelihood and

impact; b) Assigning a risk level for each harm; and c) Considering residual risk after mitigating measures have been applied." (Vol 3, 9.52)

The proposals for governance oversight are retrospective – reviewing the process of risk management retrospectively (what the company is going to do to mitigate the risks as they arise) rather than engaging in prospective analysis, looking at results from a risk assessment of the design and safety of their service and the risks of harm that may arise from it and putting mitigating measures upfront. For example:

> "Regular review of risk management and regulatory compliance by a governance body is required for appropriate oversight over internal controls. Evidence supporting this principle can be found in corporate governance good practice principles and codes. It will be important for governance bodies within services to have a full understanding of risks as identified in an illegal content risk assessment, measures that a service has put in place to mitigate and manage those risks, and how a service intends to deal with developing areas of risk. This requires that governance bodies are made aware of relevant information regarding risk management in a service (provided, for example, by internal assurance functions) and have appropriate reporting lines with senior management." (Vol 3, 8.25 and 8.26)

It is also not clear how this suggestion in the risk assessment provisions that the providers measure impact of what they do fits in with the "safe harbour" approach that Ofcom has constructed. Ofcom's overview document says services do not have to assess risk of every possible offence occurring on service - but if they have evidence, they should consider this:

> "Services do however need to assess the risk of harm from relevant non-priority offences appearing on the service … this does not mean assessing the risk of every possible individual offence that is not a priority offence occurring on your service. However, if you have evidence or reason to believe that other types of illegal harm that are not listed as priority offences in the Act are likely to occur on your service, then you should consider those in your risk assessment." (Summary document 2.33)

But the "tick-box", rigid nature of the measures set out in Volume 4 doesn't lend itself to a flexible, iterative approach to risk management that would take the service's own experience to deliver continuous improvement. See section 6, above. There is also the question whether – given that Ofcom hasn't really dealt with required mitigations for non-priority crimes - a service which totally ignores them in its risk assessment or which takes required mitigation measures but only so far as specified by Ofcom can just leave non-priority content alone. We would

assume not, but this further emphasises the need to ensure that measures in the codes apply to the general regulatory obligations that apply across the board.

The risk assessment also contains the same assumptions that bigger size equates to bigger risk purely based on the quantification of users (see also section 7, above): "As part of the risk level table, we also provide draft guidance on the effect of a service's user numbers on its level of risk**. In general, all else being equal, the more users a service has, the more users can be affected by illegal content and the greater the impact of any illegal content.** We have therefore proposed that services which reach certain user numbers should consider the potential impact of harm to be medium or high." (Vol 3, 9.59)

This focus on quantity of people affected rather than the seriousness of the harm contradicts Ofcom's own evidence in Volume 2 (6F.31) in relation to the hate offences: "However, there is evidence that niche online services can contain far more abuse, including hateful activity, than mainstream services, despite these services attracting far fewer users".

There is some acknowledgement in the risk assessment guidance that "in some instances the number of users may be a weak indicator of risk level. They need to be considered alongside other risk factors. It is possible for a large service to be low risk, and for a small service to be high risk, depending on the specific circumstances of each service" (Vol 3, 9.62) But this then does not flow through to the recommended measures in the codes which, as we have seen above in section 7, do not apply to all platforms regardless of size.

Evidence

We have attached to this submission at annex F a paper prepared by Peter Hanley and Gretchen Peters that argues for Ofcom to shift its approach to a "product assured safety management" approach which would "encourage safety rather respond to risk, and stop problems before the emerge rather than cleaning them up afterwards". This builds on their expertise and experience in other sectors and is in line with the principles that underpin the UK's Health and Safety at Work Act 1974.

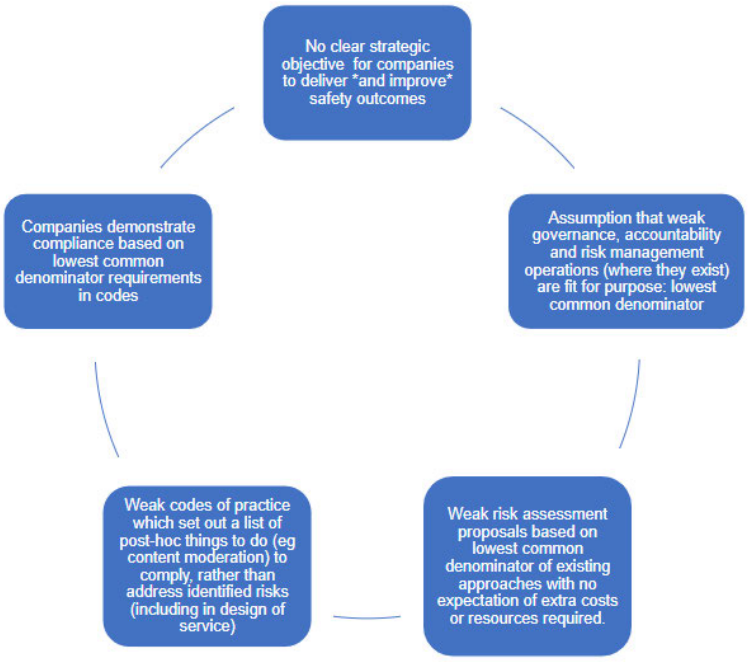
We published a blog on Ofcom's approach to governance in the light of a Wired interview with Del Harvey - the former head of Trust and Safety at Twitter (now X). In it, Harvey talks about some of the things that concerned her during her time in her role. She gives the example of trying to escalate within the company the potential threat from a DM she had received suggesting that Twitter's offices should be bombed: there was no route within the company to do this for such tweets. Harvey says:

> *"It was the same issue that it always has been and always will be, which is resourcing. I made requests in 2010 for functionalities that did not get implemented, in many instances, till a decade-plus later."*

She also gives the following example: "*Multiple account detection and returning accounts. If you're a multiple-time violator, how do we make sure you stop? Without going down this weird path of, "Well, we aren't sure if this is the best use of resources, so instead, we will do nothing in that realm and instead come up with a new product feature." Because it was growth at all costs, and safety eventually*."

Finally, and crucially, she says: "*When trust and safety is going well, no one thinks about it or talks about it. And when trust and safety is going poorly, it's usually something that leadership wants to blame on policies.* ***Quite frankly, policies are going to be a Band-Aid if your product isn't being designed in a way that actually doesn't encourage abuse****. You've got to plan there, guys*." [*emphasis added*]

Ofcom has baked in something of a "band-aid" approach to the codes of practice in Volume 4, which implies that most costs are related to moderation resources *not* investing in safety by design, risk assessment, etc. (Again, see analysis above in section 6 and 7.) There's therefore a circular weakness built into the proposals: "as is" (weak, ineffective) governance structures - loose risk assessment expectations - weak codes (based on weak risk assessments and assumption that moderation is all that's required) = compliance which then perpetuates the status quo. (See diagram below.)



We would urge Ofcom to take a step back in the light of this submission and use the evidence that we have provided along with submissions from the organisations we work closely with, to review their proposals and ask "Are we happy with what these big companies are doing already?" and "what is realistic to expect from the smaller ones"?

The most prominent recent example of a failure in risk assessment and mitigation is related to Facebook's failure to take adequate measures to assess the human rights implications of its activities in Myanmar and address the risks of extremism and real-world violence that flowed from that. The report commissioned by Facebook is here. Sadly, that is not the only example and, while Meta has commissioned a number of reports, there have been criticisms about them. It is also unclear the extent to which Meta has taken these reports on board. There has been less publicity around the impact assessments and reports on human rights carried out by other tech companies – though not all have done this in any event. We include references here:

- Commentary on Meta's Human Rights Annual Reports here and here.
- Meta's response to another report re Palestine which gave rise to some concerns here and here; and a report on hate speech in India
- Meta's submission to the UN - which has some references in it:
- Meta's HRIA re Sri Lanka

- Wikimedia Foundation HRIA
- Details on Google: here, here and here
- Microsoft report

There are plenty of existing frameworks for rights-based risk assessments that Ofcom can use to improve its approach and methodology. Professor Lorna Woods, under the auspices of Carnegie UK, developed a four-stage model for risk assessment and mitigation on social media platforms that draws on best practice processes through a code-based approach. We would refer Ofcom to her Model Code of Practice as evidence but also provide here extracts from the Ad Hoc Advice to the United Nations Special Rapporteur on Minority Issues which focus on risk assessment. (pp 7-11) This advice was a precursor to the advice to inform the development of his guidance on hate speech as a precursor to developing the Model Code.

> "There is a wealth of high-level guidance on risk assessment that social media companies do not appear to be following. (See Sanjana Hattotuwa, "Making Facebook's New Human Rights Policy Real", Institute for Human Rights and Business 20 April 2021).
>
> Social media companies coming to risk assessment for the first time should evaluate its existing risk management practices and processes, practices in relation to human rights impact assessments generally, and data protection/ privacy impact assessments to evaluate any gap or tensions in those practices and processes and ensure that there is appropriate. Particular attention should be paid to reliance on techniques driven by machine learning and artificial intelligence and the well-known questions around the design and deployment of ML/AI46. (see Council of Europe 'Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems The risk assessment process should be based on data and, where available, research, rather than a hopeful expectation that bad stuff is not happening or, if it is, that it is not the problem of the social media provider. It involves the recognition that the use of technology, including AI, does not in and of itself necessarily ensure human flourishing. (See UNESCO First Draft of the Recommendation on the Ethics of Artificial Intelligence). It should cover an assessment of actual and potential impacts. This involves gathering data in a systemic manner as to what is happening on the service (e.g. what sorts of user complaints are coming, how are they dealt with), as well as the results of any testing on the product (see below), to understand the nature of the problem, as well as its scale, context and triggers and to acknowledge that information, not bury it.

For example, hate speech tends to spike for 24-48 hours after key national or international events such as a terror attack, and then rapidly fall. (See Matthew Williams and Mishcon de Reya, Hatred Behind the Screens: A Report on the Rise of Online Hate Speech). Systems should be responsive to foreseeable public events (e.g. major sporting championships), and the due diligence process and mitigations should reflect this. Companies should also bear in mind wider industry experience (e.g. whether certain features – for example live streaming – are particularly risky) and good practice. Where human rights are involved in risk assessment and risk management, their special nature should be recognised, as the OECD due diligence guidance recognises. Companies should respect the need for diversity and inclusion in a risk assessment process so that issues – especially those which particularly affect minorities – are not overlooked or under-valued. This may be particularly relevant when products designed for operation in one state are then deployed in others.

Recommendation

While Ofcom has carried out an extensive review of the literature on risk assessment, we would recommend that further advice is sought on the many experts available who understand how best to carry this out – particularly with regard to product safety testing – in sectors that have a similar obligation with regard to the safe design and operation of their products and services. We also suggest – as per the recommendation in section 1 above - that product testing should be a mandatory part of the risk assessment process, even if discretion is given to services on the way in which they undertake this.

We include the wording we propose in section one, above, again for completeness here, given that it addresses many of the points we make above in relation to risk assessments.

"Design of functionalities, algorithms and other features

Product testing

For all services, suitable and sufficient product testing should be carried out during the design and development of functionalities, algorithms and other features to identify whether those features are likely to contribute to the risk of harm arising from illegal content on the service.

The results of this product testing should form a core input to all services risk assessments.

Mitigating measures

For all services, measures to respond to the risks identified in the risk assessment should be taken, including but not limited to, providing extra tools and functionalities, by redesigning the features associate with the risks, by limiting access to them where appropriate or where the risk of harm is sufficiently severe by withdrawing the function, algorithm or other feature.

Decisions taken on mitigating measures, as part of the product design process or as a response to issues arising from the risk assessment, should be recorded. (Note: this would be included in the record keeping duties under section 23 (u2U) and section 34 (search).)

Monitoring and measurement

All services should develop appropriate metrics to measure the effectiveness of the mitigating measures taken in reducing the risk of harm identified in the risk assessment. These measures should feed back into the risk assessment."

Issue 9: Violence Against Women and Girls (VAWG)

Issue

There are a number of new criminal offences proposed that address online VAWG, which are welcome. But the impact of all the strategic and policy decisions taken by Ofcom above will do little to shift the dial in terms of their overall safety online. Indeed the UN Special Rapporteur on Violence Against Women and Girls who, at the time of writing, has just finished a visit to the UK, said:

> "While the enactment of the Online Safety Act is a welcome development, gaps remain, specifically around the issues of violence within the pornography industry, the influence of pornography on individual and societal attitudes towards VAWG and the impact of legal pornography on perpetration of child sexual abuse, both online and offline. We need to move away from companies self-regulating towards a legally enforced duty of care on tech companies across the distribution chain to ensure that they have adequate infrastructure to prevent tech abuse and to support survivors." (UNSR Summary of Preliminary Findings after visit to UK – 21 February 2024)

Until the Government conceded on Baroness Morgan's amendment in the latter stages of the Bill's Parliamentary passage, the Government promised that the new offences would go a long way to improving protections for women and girls and that a separate code of practice was unnecessary. The opposite is true – and Ofcom's guidance on VAWG, which was the Government's concession, will not be consulted on for at least another year.

Further, as evidenced in Glitch's Digital Misogynoir Report, Black women continue to be disproportionately impacted by online abuse, and the online abuse directed towards Black women is interconnected with other forms of hate online, like antisemitism, Islamophobia and transphobia. While the OSA accounts for intersectionality, it remains to be seen how those vulnerable to harm because of their intersectional identities will be protected; nor is it clear how Ofcom plans to develop and implement frameworks for ensuring Black women – and many other multiply-marginalised communities – do not fall through regulatory and legal gaps.

More detail on these concerns is provided in a detailed joint submission from 15 organisations in the VAWG sector, which we support. We would also draw Ofcom's attention to the submission from Professor Clare McGlynn, from Durham University, which looks in detail at the consultation's proposals in relation specifically to the intimate image abuse, cyberflashing and extreme pornography offences.

Issue 10: Gaps and other consultation issues

In this section we cover a number of issues emerging from the consultation, including gaps in the proposals that either have not been acknowledged by Ofcom or have been acknowledged but could be (partially) filled and some points to make on the process.

Gaps in protections

Priority and non-priority offences

- **Missing offences: we mention these in section 2 above on the illegal content judgements but to reiterate here** Section 127(1) Communications Act, s 1 Malicious Communications Act and obscenity are missing from the list of harms - these will perform a mopping up role (eg relating to racist abuse of footballers; videos of executions; abuse related to tragedies etc). Detailed guidance given by Ofcom is only on priority harms and a limited selection of non-priority offences (seemingly because they are new). Section 127 of Communications Act: offence of "sending a message or other matter that is grossly offensive or of an indecent, obscene or menacing character". Note that OSA Schedule 6 refers to s 2 Obscene Publications Act (but in reference to children only). The fact that it is mentioned there means complete disregard later is the more noticeable. CPS guidance on that is here and, on relevance of section 127 and other communications offences, see here.

- **Neither the governance approach nor the guidance on illegal content focus on non-priority offences more broadly**: does this mean that all non-priority offences are effectively excluded from the duties? In the approach document, Ofcom says: "Services do however need to assess the risk of harm from relevant non-priority offences appearing on the service … this does not mean assessing the risk of every possible individual offence that is not a priority offence occurring on your service. However, if you have evidence or reason to believe that other types of illegal harm that are not listed as priority offences in the Act are likely to occur on your service, then you should consider those in your risk assessment." (Summary document Vol 1 2.33)

  It then goes on to say that "Effective governance and accountability processes should be effective in tackling all priority illegal harms" (Vol 3 8,13) Then, in Annex 10 Ofcom says at para A1: 30 "In recognition of the quantity and complexity of offences which could be included within the scope of the definition of 'other' offences, Ofcom has chosen to provide specific guidance on 'other' offences where they have been created by the Online Safety Act and do not wholly overlap with any priority offences." (This is limited to epilepsy trolling,

self-harm, cyberflashing, false communications, threatening communications – these are the new offences introduced by the Act, not necessarily a complete list of those most likely to be relevant under non-priority offences). We do not think this is what is intended by the Act: while there are some distinctions between priority and non-priority offences, and 'other illegal content' is dealt with together, it does not in principle exclude categories of illegal content (see e.g. s. 9(5)(d) which seems to expect consideration across the board).

- **Animal protection**: while these offences were added at a late stage in the Bill's passage (references here), Ofcom's position potentially means that service providers have no obligations with regard to this priority offence until such time as Ofcom decide to include it in their risk register.

> "At a fairly late stage in its consideration of the Bill which became the Online Safety Act, the offence in section 4(1) of the Animal Welfare Act 2006 (unnecessary suffering of an animal) was added to the list of priority offences. We will consult in due course on how we propose to include that offence in our Register." (Vol 2, 5.21)

Exclusions from risk assessments and codes

- **Supply chain**: there are very limited references to risks of supply chain/third party involvement despite recognition that many services will rely on third party software (or moderation services) in their business. For example:

> "Requiring services to have measures to mitigate and manage illegal content risks audited by an independent third-party; d) Requiring due diligence of third-party contractors or providers of services involved in the mitigation and management of illegal content risks to assure their approaches lead to good online safety outcomes". (Vol 2, 8.97)

> "If they have automated technology at all it is likely to be trained by a third-party (i.e. 'off-the-shelf' tools), rather than bespoke and/or specially trained automated technology." (Vol 3, 12.22)

> "We understand that third-party entities support perceptual hash matching, and it forms the basis of many in-house solutions developed by larger service providers. Some services discuss their use of perceptual hash matching technology and solutions publicly, such as through transparency reporting". (Vol 3, 14.50)

- **External events**: the consultation flags "the propensity for external events to lead to a significant increase in demand for content moderation" (measure 5 of the codes). But this aspect of content moderation resources and practice does not take account of terrorism, only "expected" or "scheduled" events. (Vol 4, 13.129)

- **Search** - clicking through thumbnails to harmful content is identified in risk profile document in a few places but then in the codes, there is no mention of a "one-click" limit**.**

  For example, at 6U38: "Service design may in some instances facilitate the risk of illegal content being encountered and shared and therefore increase the risks of harm to users on U2U or search services. Offence-specific risks of harm associated with service design are outlined in different chapters of this Register, and the most prominent examples are in chapter 6D: Encouraging or assisting suicide or serious self-harm and chapter 6L: Extreme pornography. Such examples relate to how vulnerable users may be recommended content that is increasingly harmful and potentially illegal. Similarly, users may be led to illegal content within a few clicks from their query on a search service (for further information, see chapter 6T on risks of harm to individuals on search services)." (Volume 2, para 2.29)

  6U.50 "Further information as to how services can implement service design effectively on search services, and mitigate the risks described here, can be found in the Codes of Practice"

  "It is important to recognise that content is to be treated as 'encountered via' search results where it is encountered as a consequence of interacting with results (for example by clicking on them). This means that search content includes content on a webpage that can be accessed by interacting with search results. The safety duties, and the measures we recommend for the purposes of complying with them below, should be considered in this context." (Volume 4, 13.5)

  Evidence recently demonstrated how deepfake porn was found just one click away via Google and Bing and Ofcom's own recent research has found similar with regard to self-harm content (research commissioned to inform the child safety code but which has direct relevance to design choices relating to illegal content too.) Harm may also be indirect. This also may be a particular issue for landing pages or review sites which make the route to illegal content clear; adverts for/discussion of tools (e.g. nudification apps) which are then used for illegal purposes.

- **Nudification apps**: these are not illegal but there is no mention of these (or related harm) in volume 2; it is hard to think of a legitimate use for them.

- **GenAI and metaverse:** there are references to these in paras 3.60 and 3.61 in the summary document in the section on "future trends and developments", with Ofcom reflecting that the illustrative examples we've given above only reflect the sectors as they stand now, and we expect their features and models to evolve over time, alongside other developments and new entrants to the landscape. In order to regulate effectively we will therefore need to scan the horizon for new developments and we expect our approach to regulation will evolve over time as things change.
  However, the Government during the passage of the Bill was keen to emphasise how the approach was "technology neutral" and harms arising from these new technologies would be covered if it was user-to-user in nature. See, for example, Lord Parkinson in the Lords Committee stage debate on 25 May:

  > "The Bill has been designed to be technology-neutral in order to capture new services that may arise in this rapidly evolving sector. It confers duties on any service that enables users to interact with each other, as well as search services, meaning that any new internet service that enables user interaction will be caught by it … the Bill is designed to regulate providers of user-to-user services, regardless of the specific technologies they use to deliver their service, including virtual reality and augmented reality content. This is because any service that allows its users to encounter content generated, uploaded or shared by other users is in scope unless exempt. "Content" is defined very broadly in Clause 207(1) as
  >
  > "anything communicated by means of an internet service".
  >
  > This includes virtual or augmented reality. The Bill's duties therefore cover all user-generated content present on the service, regardless of the form this content takes, including virtual reality and augmented reality content. **To state it plainly: platforms that allow such content—for example, the metaverse—are firmly in scope of the Bill.**" (Hansard 25 May col 1010)

There is plenty of evidence already of harm from both technologies in the here and now which could be reflected in Volumes 2, 3 and 4. There was a particularly graphic debate in the Lords at Committee stage of the Online Safety Bill (indeed, so graphic that a group of school children were ushered out of the public gallery) on the sexual abuse of children within VR environments. And there have been numerous recent reports: see for example,

the IET report on harm arising in virtual spaces; the NSPCC's detailed report on "Child Safeguarding and Immersive Technologies" and the recent news report of a virtual gang-rape of an under-16 in the metaverse. On Gen AI, Europol reported last year on its exploitation by criminals and Taylor Swift has recently been a very high-profile victim of deepfake porn. Fraud is another area of concern. Yet Ofcom gives no timescales for how they are going to respond to this in future iterations of the codes and again, without the "catch-all" measure we recommend above, there is no obligation on services to take steps to address these harms in order to comply with their regulatory duties.

- **Equality Act** – Ofcom has an obligation with regards to the Welsh language (and includes a question on this in the consultation form) But while there is a brief mention of Welsh language posts in Annex 13, in relation to content moderation and resourcing, there is no equivalent obligation re minority languages in the UK. This is surprising given that Ofcom has insight into some of the challenges here from its broadcasting role, including the tensions between different communities within the UK; for example, here and here.

  Annex 13 at A13.8: "More generally, we are proposing that services should have regard to the needs of their user base in considering what languages are needed for their content moderation, complaints handling, terms of service and publicly available statements. To this extent, we consider our proposals are likely to have positive effects or increased positive effects on opportunities to use Welsh and treating Welsh no less favourably than English."

  It is worth nothing here the New Mexico attorney general finding re Meta and CSAM: "As with images, Meta's identification and blocking of terms associated with trafficking and CSAM are too narrow and rigid, and easily evaded, and do not adequately screen communications and terms in Spanish or other languages" (page 96) Also this report looks at the role of Facebook and Telegram in allowing incitement and online hate to spread in countries where a lack of moderators in the local languages was a factor.

Iterative approach

- **Speed vs comprehensiveness:** this has been used by Ofcom officials in a number of discussions as a reason for not covering particular aspects (eg things added late in the passage of the Bill) with a promise that the illegal content codes will be a "first iteration" and will be revised. This is understandable, especially given that Ofcom has only recently been given information gathering powers. Even accepting this, there are still questions. What are the timescales for this? What evidence will be needed for these next iterations? The risk here is that the regime gets embedded in this "lowest common denominator" form and watered down, via company lobbying, judicial review actions etc, from there, rather than being built on stronger foundations and continuously improved. Moreover, as

we have argued, there are some issues arising out of Ofcom's interpretation of the Act which skew the conclusions drawn. This would not necessarily be revised in subsequent iterations of codes.

Barriers to effective engagement

- **Approach to the consultation:** the size of the consultation and the resources required to engage with it in a meaningful way is a barrier to the engagement of a huge swathe of civil society, where resources and capacity is limited. While Ofcom, as they frequently point out, have provided a very succinct and clear summary of their proposals, the detail is in the 1700-odd pages and – as we have set out above – is often woven into the fabric of the multiple volumes and annexes. Moreover, there does not seem to have been a mechanism set up with appropriate safeguards and protections for individual survivors, victims or those with lived experience of the harms set out in this consultation to contribute in a way that could enable Ofcom to judge the effectiveness of their proposals against the reality of harm and its impact on users.

February 2024