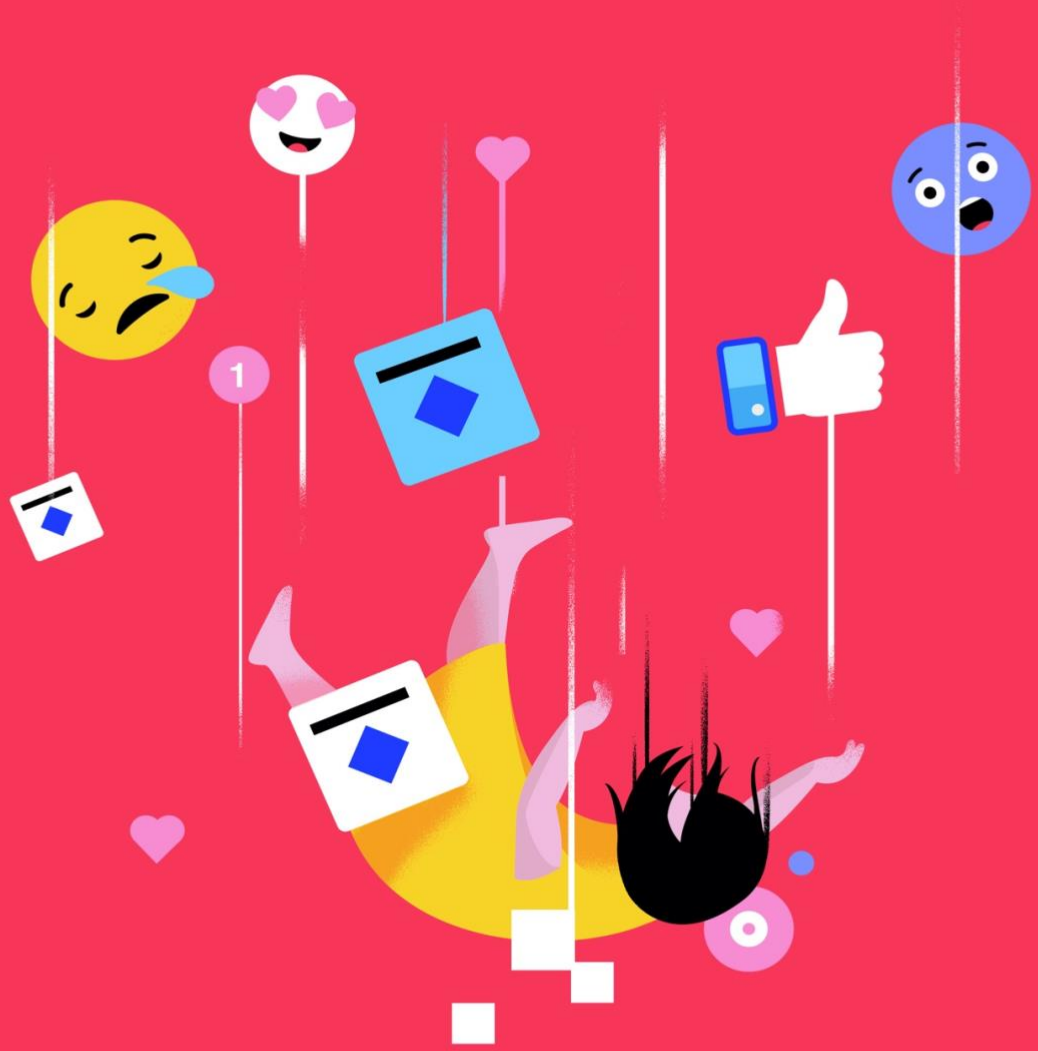
April 2021

Ambitions for the Online Safety Bill



5RIGHTS FOUNDATION

About 5Rights Foundation
Building the Digital World Young People Deserve

5Rights develops new policy, creates innovative frameworks, develops technical standards, publishes research, challenges received narratives and ensures that children's rights and needs are recognised and prioritised in the digital world.

Our focus is on implementable change and our work is cited and used widely around the world. Our co-operation with governments, intergovernmental institutions, professional associations, academics, businesses, and children is focused on making change, so that digital products and services can impact positively on the lived experiences of young people.

5Rights works on behalf of and with children and young people under 18, our observations are reserved entirely for this age group though many of our comments are relevant to other user groups. Our mission is to build the digital world young people deserve.

Contents

Preface 4

Executive summary 5

Chapter 1: Scope 7

- Services out of scope 8
- The tier system 12
- Services likely to be accessed by children 13

Chapter 2: Types of harm 15

- Definition of harm 15
- Harms out of scope 20

Chapter 3: Duties on regulated companies 25

- Child Risk Assessments 26
- Safety by design 27
- Age assurance 28
- Moderation, reporting and transparency 29
- Terms and conditions 32

Chapter 4: Duties and powers of the regulator 34

- Enforcement 34
- Digital and media literacy 37
- Upholding children's rights 38
- Timing and the wider regulatory landscape 39

Conclusion 42

Preface

The problems children face from the digital world are systemic. They are not restricted to technical bugs or bad actors but are also present in the features and architecture of the products and services that make up the world they inhabit.

Children rely on digital products and tools to access education, health, entertainment, civic engagement and to manage their relationships with family and friends. They are impacted by technology that they cannot see, do not control and may not be aware of. From facial recognition in schools and the use of algorithms for exam grading to predictive analytics used in policing, the digital world has an immeasurable impact on children.

Into this picture comes the Online Safety Bill, first announced by the UK government in 2017 and set out in full in December 2020. The stated purpose of the Online Safety Bill is to prevent harm. Yet this is undermined by exemptions and caveats that offer a complicated regulatory landscape with places for companies to hide. The responsibilities envisioned in the concept of a duty of care have been limited by restricting the scope. The promise of a systemic approach that tackles harm at its root has been replaced by a focus on content and on what is already illegal, shying away from the features and processes that optimise for profit over children's safety.

In its response, the government defines harm as extreme and individual, when in reality, many harms of the digital world are incremental, repetitive, cumulative, and impact on individuals, groups and on society more broadly. There is a worrying lack of clarity about the status of the proposed measures, with a complex set of duties, codes, and provisions that may or may not have the full force of law.

The UK government was the first to identify the need for a systemic, precautionary and risk-based approach to the digital world. The Online Safety Bill is an opportunity to fulfil that vision and make the digital world meet the expectations of society, uphold our laws, realise our rights and treat children with the respect for their wellbeing and safety that they deserve.

This report builds on our previous reports *Towards an Internet Safety Strategy (2019)* and *Building the Digital World that Young People Deserve (2020).* It is complemented by our recent report on age assurance*, But how do they know it is a child? (2021).*

I wish the Online Safety Bill well as it makes it journey through Parliament, but while there is much to welcome, there is still some way to go before it becomes worthy of its name.

Baroness Beeban Kidron, Chair, 5Rights Foundation

Executive summary

The arrival of the Online Safety Bill is very welcome, and so too is the government's determination to protect children. But the scope, definition of harm and enforcement regime all require further consideration and clarification.

- **Children front and centre.** The government has placed particular emphasis on protecting children online, but in restricting the scope, has left them at risk of known harms. The scope of the Bill should include all digital technology that impacts on children, not only services that host user-generated content or facilitate interaction between users.
- **Duty of care.** The new duty of care model requires regulated companies to take proactive measures to protect their users. To drive the necessary culture change across the sector, the duty of care must be applied to a wider group of companies and cover a broader set of harms.
- **Loopholes in the regulation.** The criteria used to define scope and differentiated expectations on companies sets the stage for a complex regulatory system. It risks companies remodelling their services to avoid regulation instead of taking on the responsibility to make them safer.
- **Small is not safe.** The government has made carve-outs for smaller services in an attempt to reduce the regulatory burden. Small services need greater support to comply with regulation, not permission to harm.
- **Futureproofing the new regime.** The Bill must be futureproof to account for new technologies and to allow the regulator to address risk and harm, wherever it may emerge in the future.
- **Taking a proportionate approach.** A fair regime can be achieved by a proportionate approach to enforcement, as per the Regulators' Code, not by exemptions or narrowing scope.
- **Definition of harm.** The threshold of harm sets a high bar that will fail to capture many of the less obvious harms children experience online. The Bill's definition of harm must account for harms that build over time and those that impact groups of children and society itself, not only the most immediate cases of harm to individuals.
- **Accounting for all known harms.** The parameters defining scope and the exclusion of some harms will leave significant risks to young people unaddressed. Where harms are left out of scope, the government should set out a time-bound plan to address them in alternative legislation.
- **Moderation and reporting.** Moderation, reporting and transparency are important features for user safety, but the concept of a duty of care must be rooted in preventative measures that identify and mitigate risks before harm occurs.
- **Safety by design.** The safety by design framework must carry statutory weight as the primary mechanism by which companies can identify and mitigate the risks

posed by their services and fulfil the duty of care. It must be applicable to all products and services likely to be accessed by children.

- **Automated systems.** The emphasis on harmful content must not result in other risks being overlooked, particularly those embedded in automated systems. Automated systems and algorithms must be firmly within the scope of the Bill and the regulator must be given sufficient oversight and enforcement powers to regulate and audit them.
- **Age assurance.** Tools that identify, restrict or protect children (including age assurance and parental controls) must be privacy-preserving and subject to common standards that are enforced. The Bill must mandate child-centred design, rather than focus on keeping children out of spaces to which they have a right to access.
- **Digital and media literacy.** Empowering children to be responsible actors in the digital world is fundamental, but neither they, nor their parents, should be made responsible for mitigating risks that need to be addressed at the level of system design.
- **Enforcing regulation.** The efficacy of the regulation will depend on how it is enforced and the independence, capacity, and resources of the regulator. The Codes of Practice will only have efficacy if they are mandated and enforced.
- **Timing.** The government must commit to drawing up Codes of Practice and frameworks alongside the drafting of the Bill, so they are ready to implement as soon as the Bill has passed.
- **Recognising children's rights.** The Bill must recognise and adopt existing provisions for children in the digital environment. As a signatory to the United Nations Convention on the Rights of the Child, the government should ensure that the Online Safety Bill formally recognises General comment No. 25 on children's rights in relation to the digital environment.[1]

The government must make good on its promise to protect children online by applying the duty of care to all digital services that impact on children, wherever they are and whatever the harm.

[1] General comment No. 25 (2021) on children's rights in relation to the digital environment.

Chapter 1: Scope

The government has restricted the scope of the new legislation to apply only to services that host user-generated content or facilitate interaction between users. Products or services that do not fall into these categories or which have explicit exemption will not need to comply with the new regulations. The parameters used to define scope will mean a huge number of companies will not be subject to the new duty of care - in all, less than 3% of UK businesses will fall within scope.[2]

- User-generated content is defined as image, text or audio content that is produced or shared by users of an online service. User interaction is defined as any public or private online interaction between service users. For both definitions, a 'user' refers to any individual or organisation (private or public) that shares content on an online service.
- Social media sites, instant messaging services, video sharing platforms, online forums and some video games will be in scope, as well as search engines.
- Under a two-tier system, there will be differentiated expectations on companies in scope depending on their size and functionality.
- A small number of "high risk, high reach" category 1 services will need to address legal but harmful content for adults.
- The majority of services will be in category 2 and will be subject to less stringent rules. Services "likely to be accessed" by children, in both categories, will have to meet additional requirements.[3]

Restricting services in scope to those that host user-generated content or facilitate interaction between users misunderstands the nature of the digital world. User journeys involve hundreds of interactions with multiple services and products. For example, a child may watch sponsored content from an influencer on a social media service, advertising a new game that does not host user-generated content or facilitate interaction between users. The child is then taken from a service within the scope of the Bill to one outside of it. As noted by the NSPCC, the proposals do not go far enough to address the cross-platform nature of risk.[4]

[2] "Fewer than 3% of UK businesses will be in scope. We will focus on the biggest, highest risk online companies where most illegal and harmful activity is taking place." See Joint Ministerial foreword, Online Harms White Paper: Full government response to the consultation, Department for Digital, Culture, Media & Sport and Home Office, December 2020.

[3] "The framework will deliver a higher level of protection for children than for adults." Online Harms White Paper: Full government response to the consultation, Department for Digital, Culture, Media & Sport and Home Office, December 2020.

[4] Delivering a Duty of Care, An assessment of the Government's proposals against the NSPCC's six tests for the Online Safety Bill, NSPCC, March 2021.

In the four or more years since the online harms regime was first conceived, the digital world and the risks it creates for children have evolved considerably. In restricting the scope, the government will create potential scenarios in which the regulator will be powerless to prevent both known and emerging harms, and under the current proposals, the Bill will not account sufficiently for new technologies and new risks.

Services out of scope

> "The legislation will exempt many low-risk businesses with limited functionality… the online harms regulatory framework has been designed to reduce the burden on UK business by focussing on the areas that present the greatest risk of harm."

The government repeatedly cites the burden of regulation on business to justify exemptions and does not acknowledge that regulation can in fact *support* fledgling companies as they look to scale-up and become commercially viable. Introducing exemptions on this basis also risks incentivising companies to redesign their services to avoid being in scope of regulation, instead of redesigning their services to make them safer. For example, under the proposals as currently set out, a commercial pornography website would only need to disable its commenting functions to be out of scope.[5]

This approach is inconsistent with safety regulation in other industries. For example, Food safety regulations "apply to anything from a hot dog van to a five-star restaurant, from a village hall where food is prepared to a large supermarket, or to a vending machine."[6] Similarly, all employers, regardless of the size or nature of the workplace, have a duty to protect the health, safety and welfare of employees under the Health and Safety at Work Act.[7] It should follow that online safety regulation applies to any and all online services, irrespective of nature or size. Children wish to be protected from all harm, whether from user-generated material or commercial content, by a tech giant or by start-up, in school or at home.

Proportionality is a requirement of all regulatory regimes and the regulator must take a proportionate view of the risks posed by companies in scope. Unnecessary regulatory burden can be avoided by additional support, proportionate mitigation strategies and enforcement activity from the regulator, as set out in the Regulators' Code.[8]

[5] This is based on the pornography website showing only 'publisher' generated rather than 'user' generated content.

[6] The Food Safety (General Food Hygiene) Regulation, 1995.

[7] Health and Safety at Work Act, 1974.

[8] Regulators' Code, Office for Product Safety and Standards, April 2014.

Exemption of services managed by education institutions

The stated reason for the exemption of online services managed by education institutions is "to avoid unnecessarily adding to any online safeguarding regulatory or inspection frameworks (or similar processes) already in place."[9] This does not reflect the reality of life in schools. Teachers and pupils use a variety of apps and services in the classroom and for homework that are not always subject to school procurement and safeguarding regimes.

It would be unnecessarily high-handed for the government or schools to ban the use of such services, many of which enliven and support learning, but at the same time, pupils and parents need to feel safe in the knowledge that these products and services are subject to online safety regulation. Many organisations supporting online education have invested heavily in robust access control and safeguarding systems and guidance to improve online safety in education settings.[10] These efforts will be undermined if the products and services used by schools have risk engineered into their systems.

This exemption is also inconsistent with Ofsted's analysis of remote education, which acknowledges the additional safeguarding challenges created by the shift to online learning.[11] It is also inconsistent with the views and needs of parents and teachers.

> **"Remote learning [platforms'] terms and conditions are difficult to understand, so it shouldn't be on parents or teachers to try to regulate. The government should be making standards for this to be better and have safer design." - Primary school teacher, London.**

By characterising online harm in education settings as a safeguarding issue, the government has failed to consider that many EdTech providers do not offer fit for purpose security mechanisms on learning platforms or provide sufficient protections for children's data. It also outsources the cost of mitigating harm onto the education sector, which spends increasing amounts of money and hundreds of school hours on safeguarding issues that could be avoided.

---

[9] Paragraph 1.6, Online Harms White Paper: Full government response to the consultation, Department for Digital, Culture, Media & Sport and Home Office, December 2020.

[10] For example, Jisc and the London Grid for Learning (covering 3,500 schools in London and elsewhere) offer an array of training packages, downloadable guides, and first-line support for online/remote education.

[11] "When it comes to remote education, considerations around protecting the safety of students, particularly in online environments, posed a central challenge since March 2020. Many schools we spoke to stated safeguarding as a key consideration when making decisions about which platforms and digital tools to use for their remote offers. In particular, there were concerns about cameras. Some schools opted to disable cameras in the interests of safety. Others felt face-to-face contact online was fundamental to the well-being of staff and pupils and so made a risk-assessed decision to keep them on based on this. A number of schools made reference to having additional staff present for live lessons. Other safeguarding considerations mentioned less frequently in the interview sample were around data protection (one school leader worried about the sharing of pupil data with 'big tech' companies) and keeping screen time at a healthy level." Remote Education Research, Ofsted, January 2021.

Recent concerns about safeguarding in school speak to a wider concern about the availability of pornography and the normalising of certain behaviours present on many products and services. In many ways, schools are at the front line of some of the most insidious harms of the digital world. To exclude online services managed by education institutions from a proactive duty of care leaves schools and their pupils more vulnerable to these harms. Importantly, it threatens to undermine a school's status as a place in which children can learn safely.

Cases of 'zoombombing' — when video conferencing sessions are disrupted by malicious actors — have been widely reported by schools throughout the coronavirus pandemic. Footage from such cases emerged on platforms such as YouTube and TikTok.[12] Not only do these incidents cause huge disruption to learning but can be highly distressing for both students and teachers. In some cases, children have been exposed to child sexual abuse material (CSAM) and pornography as a result of these intrusions.

Google Docs, part of Google for Education's suite of resources, uses collaboration features including a live chat function. The chat is deleted after users close out of the document, meaning a record of what is said is not kept. More than 60,000 cases of bullying through the Google Doc chat function have been detected and reported by a parental monitoring tool.[13]

In bringing forward an Online Safety Bill, the government has recognised that individual users — particularly children and parents — are in an asymmetric power imbalance with the demands and commercial interests of tech companies. By exempting EdTech, the government has failed to extend this acknowledgement to teachers who cannot and should not be expected to mitigate the negative impacts of the digital services on which schools and children rely.

Exemption of business services

Services that play "a functional role in enabling online activity" and business-to-business services will be out of scope. Some of these services, including file-transfer services, corporate email services, cyber lockers and webhosts are known to be used to distribute child sexual abuse material (CSAM), but despite this acknowledgement, business services have been given total exemption from the regulation.

[12] Were You Zoom-Bombed? Video of It May Now Be on YouTube, TikTok for All to See, PC Mag, April 2020.

[13] Google Docs for Kids: What Parents Should Know, Bark, November 2018.

The statistics on the production of CSAM are staggering,[14] and the UK has one of the highest numbers of users seeking to access CSAM in the world. In April 2020 alone, the Internet Watch Foundation blocked and filtered at least 8.8 million attempts by UK internet users to access videos and images of children experiencing sexual abuse.[15]

Child sexual abuse and exploitation is illegal and already companies must act when issued with a takedown notice.[16] But there is evidence that most companies do not know the extent to which their services can be used to host CSAM.[17] In the time it takes to issue and follow a takedown notice, considerable harm has already occurred.

Requiring services which otherwise pose a low risk of harm to children to ensure, at a minimum, that their service is not being used to host known CSAM, would make a measurable difference to children's safety online. By excluding business services from the new proactive measures to protect children set out in the proposed duty of care will deny the most vulnerable children protection at scale. If this exemption makes its way into the Bill, the government must set out the mechanisms through which it intends to tackle the spread of CSAM on these services.

> Cloudflare is a business-to-business service providing web infrastructure and security tools. The company has acknowledged the threat of "hacks" or the ability for "malicious employees" to store CSAM on company/internal servers.[18] In 2019, Cloudflare made over 1,000 reports to the National Centre for Missing & Exploited Children (NCMEC), the US centre for child sexual abuse material detection.

Exemption of news publishers

Content from news publishers is out of scope, including user comments relating to published content. This means the websites of newspapers and other broadcasters will not be subject to the requirements of the regulation. Freedom of the press is synonymous with the values of liberal democracy and all citizens, including children, have a right to access a variety of information.[19]

---

[14] The National Crime Agency identified at least 300,000 individuals posing a sexual threat to children in the UK in 2020 and warned of a spike in online CSA offending during the coronavirus pandemic. 320 of the UK's most dangerous child sex offenders among 4,760 arrested since first coronavirus lockdown, National Crime Agency, January 2021.

[15] Watchdog reveals 8.8m attempts to access online child abuse in April, The Guardian, May 2020.

[16] When notified about the presence of CSAM on their services via a takedown notice, companies are legally obliged to remove the content. Takedown Notices, Internet Watch Foundation, date accessed 8 April 2021.

[17] The NetClean Report, NetClean, 2016.

[18] Announcing the CSAM Scanning Tool, Free for All Cloudflare Customers, The Cloudflare Blog, December 2019.

[19] Article 17 of the UNCRC states that every child has the right to reliable information from a variety of sources, and governments should encourage the media to provide information that children can understand. Governments must help protect children from materials that could harm them.

Some news sites and user comments on published pieces are known vectors for the spread of misinformation. A misinformation tracker from NewsGuard estimates that of over 400 news sites in the UK, US, Germany, France and Italy involved in disseminating fake news about vaccines, at least 16 are based in the UK.[20]

The final definition of 'news publisher' must consider the application of editorial standards[21] of the media broadcasters and news publishers in question, to ensure that this exemption does not create a backdoor for misinformation.

The tier system

> "Services will be designated as either category 1 or 2, according to the size of their audience and the functionalities they offer, such as the ability to share content widely or contact users anonymously. All services likely to be accessed by children, whether category 1 or 2, will need to provide additional protections. Category 1 will include the most "high risk, high reach" services, and only these services will have to address legal but harmful content and activity accessed by adults. Ofcom will provide guidance on the thresholds for category 1 services, and will be able to add and remove services to the list."

The tier system complicates the requirements for companies in scope and may push risky activity to services with fewer regulatory obligations. A service with a smaller number of users can still cause significant harm and a company with a small turnover or workforce can still reach a vast number of users (see example below). Categorising by size creates the opportunity for companies to restructure to avoid regulation rather than to prevent harm.

The full response lacks detail on how category 1 and category 2 services will interface, and does not explain how it would address, for example, a user of a category 1 service sharing a link to content on a category 2 service with lesser regulatory requirements. Fixing a problem in one area, on a particular set of services, will not fix the ecosystem. The tier system also introduces another layer for potential dispute if a company took legal action against its categorisation, potentially adding years to the time it takes to resolve safety issues.

The response is either unclear or silent on how the tier system will account for:

- **'Popular by surprise' services** that start out with a small number of users and limited functionality, before quickly gaining a large user-base over a short period of time.[22] Ofcom will have to ensure that new services which present a high level of

[20] Coronavirus Misinformation Tracking Center, NewsGuard, date accessed 8 April 2021.

[21] For example, adherence to the Editor's Code.

[22] Expanding the debate about content moderation: scholarly research agendas for the coming policy debates, Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T. Roberts, Aram Sinnreich and Sarah Myers West, Internet Policy Review, Volume 9, Issue 4, October 2020.

risk are subject to the requisite regulatory requirements before reaching the threshold of a category 1 service. For example, the subscription-based service OnlyFans, where users can pay for sexually explicit content, had only 120,000 users in 2019. By December 2020, the service had more than 90 million users and over one million content creators.[23]

- **Small services that carry risk** such as the video-sharing platform Clapper, which has under 500,000 downloads on the Google Play store. Despite a minimum user age of 17, the service's weak age assurance means a child can log into Clapper via their Google account, even if they are underage. The service is known to harbour misinformation and its terms of service explicitly state that it "cannot ensure the prompt removal of objectionable material as it is transmitted or after it has been posted."[24]
- **Start-ups and SMEs** looking to attract investment and scale their products and services. As these businesses scale up or add functionality, they may be forced to redesign their services or modify their operations to meet the different regulatory requirements. Small businesses can scale more quickly and cheaply if user safety and child-centred design are baked in from the start. Australia's eSafety Commission have recently published an Investors Checklist to encourage investors and venture capitalists to consider a start-up's capacity and commitment to user safety as part of their investment criteria.[25]

Services likely to be accessed by children

The proposals recognise that children have different needs to adults, so all services in scope will be expected to assess whether children are "likely to access" their services — that is, the possibility of a child accessing it is more probable than not.[26] However, in setting the parameters of the scope to include only those services that host user-generated content or facilitate user interaction, there will be services accessed by children that do not have to comply with the duty of care, creating a completely avoidable situation in which children are offered no protection by the Online Safety Bill.

Children's experiences are not limited to products and services *directed at them or even accessed by them.* There are many services and situations that impact on children without their direct participation, for example, facial recognition technology in public places, predictive policing technology, technology used by local councils to assess need or algorithms in exam grading. Technologies that engage children without their

[23] Jobless, Selling Nudes Online and Still Struggling, The New York Times, January 2021.

[24] Terms of Service – Clapper, Clapper iOS app, last accessed 17 February 2021.

[25] "Safety measures should be considered from the very start of the design and development process, rather than being bolted on after users have already experienced online harm. Managing these risks up front can help to mitigate significant costs to reputation, brand values and business performance." Investor resources, eSafety Commissioner, date accessed 8 April 2021.

[26] When are services likely to be accessed by children, Age appropriate design: a code of practice for online services, 2020.

participation often affect them in ways they may not know. The restriction of scope to include only those services with user generated content or that facilitate interaction between users flies in the face of the real needs of children. The government has yet to suggest how it will manage the multiple situations that are out of scope.

Chapter 2: Types of harm

The full government response to the Online Harms White Paper acknowledges the need to take a systems-based approach to addressing risk, but then focusses on addressing harmful content, rather than the design features and operating processes that create risk. The definition of harm sets an unreasonably high bar and overlooks the vulnerabilities of groups of children who may experience harm differently or groups that may be impacted collectively.

We welcome the government's actions to prevent illegal content and activity online, specifically child sexual abuse and terrorist material. While content of this nature requires urgent enforcement and a far greater response from the tech sector, the pathways to illegal activity and the design features that facilitate grooming and radicalisation are under-represented in government plans.

A limited number of priority categories of harmful content is to be set out in secondary legislation. These will include:

- criminal offences, such as child sexual exploitation and abuse, terrorism, hate crimes and the sale of illegal drugs and weapons
- harmful content and activity affecting children, such as pornography or violent content
- harmful content and activity that is legal when accessed by adults, but which may be harmful to them, such as abuse and content about eating disorders, self-harm or suicide.

These 'priority' categories should be set out in primary legislation to avoid any delay in companies making their services safer for children. Meanwhile, a systemic approach to tackling risk that includes a more holistic definition of harm must be put forward.

Definition of harm

> Harmful online content and activity is defined as that which "gives rise to a reasonably foreseeable risk of a significant adverse physical or psychological impact on individuals." Content that has a "minor impact" will not be in scope of regulation.

The narrow definition of harm does not account for content and activity that alone may not cause significant or immediate harm, but in combination, can have a serious impact on children. By focusing on acute harms, the legislation will fail to address risks that develop over time, often with equally damaging effects. For example, if a child is constantly exposed to violent pornography, content that encourages extreme diets or

that vilifies women and girls, the immediate effects may not give rise to “significant adverse physical or psychological impact” but may over time have a negative impact on that child’s self-esteem, body image and relationships.[27] Importantly, some children (both individually and in groups) are more susceptible to certain types of harmful material and activity than others. For example, 45% of girls say social media makes them feel they have to look or act a certain way, compared with 29% of boys.[28]

Facebook founder Mark Zuckerberg told the BBC in May 2020 that Facebook had and would remove any content likely to result in "immediate and imminent harm" to users.[29]

He went on to suggest that that while COVID-19 misinformation is a matter of immediate harm, vaccine misinformation is simply a matter of personal belief. This fails to recognise that vaccine misinformation can seriously impact the health of children – resulting in an upsurge of measles[30] with the associated risk of blindness, infertility and occasionally death.

The antivax movements also laid the groundwork for belief in COVID-19 misinformation, which has cost the lives of many.

By February 2021, Facebook had banned vaccine misinformation.[31] As part of this policy, Facebook removes all posts (including user-generated and paid advertisements) with false claims about vaccines. However, posts containing vaccine misinformation continue to spread on the website, some gaining over 12,000 interactions before being taken down.[32]

The impact of a toxic information diet, recommended, ranked and nudged to children at a scale cannot be underestimated: from risky health or sexual behaviours and eroded trust in institutions, to polarised or extreme political views.[33] It is inappropriate to suggest such widespread targeting of children with harmful material, suggestions and nudges can be dealt with by “empowering users” to engage critically with digital

[27] Using more social media has been linked to children and young people feeling less satisfied with their bodies. In a survey conducted by the Mental Health Foundation, 40% of young people (26% of boys and 54% of girls) said that images on social media have caused them to worry in relation to their body image. Body image in childhood, Mental Health Foundation, May 2019.

[28] Reclaiming the internet for girls, Plan International UK, 2017.

[29] Facebook's Zuckerberg defends actions on virus misinformation, BBC News, May 2020.

[30] Anti-Vaccine Decision-Making and Measles Resurgence in the United States, Olivia Benecke and Sarah Elizabeth DeYoung, Global Pediatric Health, July 2019.

[31] Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19, Facebook Newsroom, February 2021.

[32] Facebook says it’s taking on Covid disinformation. So what’s all this?, WIRED, March 2021.

[33] Services do not categorise content by ‘type’, source or credibility, but promote and prioritise content based primarily on its popularity – the ‘likes’ or ‘shares’ it has amassed.

services and products. This *responsibilises*[34] children and is an unrealistic assessment of the asymmetry of power between children and the technology they are using. Crucially, it overlooks the fundamental role that algorithms and automated systems play in creating harm. The final definition of harm must account for automated systems, and the insidious and incremental risks that accumulate to create serious harm.

Missing in the government's full response is the notion of cumulative harm. Very often a child encounters a series of problems which accumulate to create a toxic or risky environment. We draw the government's attention to our risky by design[35] work and in the case study below, show how a number of different harms are present in just one activity – playing a video game. By excluding some types of harm, the government is setting itself on a path where regulation will be needed elsewhere to address the problems that are not within the scope of the Bill.

Multiple harms in video games

93% of children in the UK play video games.[36] Video games that enable interaction between users online will be in scope of regulation, but the exemptions and narrow definitions of harm in the proposed regulatory framework will leave the most prevalent risks to young people to go unaddressed. Below are some of the harms that children routinely encounter in video games.

Inconsistent age classification

All games available to download from the Apple App Store and Google Play Store are age rated. However, age ratings are inconsistent across the two app stores and in some cases, are different from the minimum age of play stated in the privacy policies of the games themselves. This can result in children downloading and playing games that contain violent content, gambling-style features, in-game purchases or advertising that are not suitable for their age.

The top 6 free games on the Apple App Store are listed below, with age ratings across both app stores and the minimum age requirements set out in each game's terms of service.[37]

[34] The Rights of the Child in the Digital Environment: From Empowerment to De-Responsibilisation, Professor Dr Eva Lievens, Freedom Security Privacy | The Future of Childhood in the Digital World, 5Rights Foundation, 2020.

[35] Risky by design, 5Rights Foundation.

[36] Lifting the Lid on Loot-Boxes, Chance-Based Purchases in Video Games and the Convergence of Gaming and Gambling, University of Plymouth, GambleAware and University of Wolverhampton, March 2021.

[37] List obtained on February 3rd 2021.

Game	Apple App Store rating	Google Play Store rating	Minimum age stated in game's privacy policy
ABC Runner	4+	PEGI 3 (All age groups)	16+
Guess Their Answer	9+	PEGI 3 (All age groups)	16+
High Heels!	4+	PEGI 3 (All age groups)	16+
Blob Runner 3D	4+	PEGI 3 (All age groups)	16+
Tricky Track 3D	12+	PEGI 3 (All age groups)	16+
Project Makeover	4+	PEGI 3 (All age groups)	16+

Mis/disinformation

Misinformation can be shared in online games via chat functions (both voice and text) or through game-play and gaming imagery.[38] For example, research from the Anti-Defamation League found that 17% of online gamers have been exposed to topics suggesting that 'Coronavirus should be called Kung Flu/ Wu Flu/ Chinese virus', 12% were exposed to conspiracy theories that Antifa were behind George Floyd protests, 10% were exposed to discussions suggesting that the Holocaust is greatly exaggerated, and 9% were exposed to White supremacy content.[39] It also found that 8% were exposed to disinformation about vaccines.[40] Few gaming services have community guidelines that address misinformation[41] and under the new regime, only misinformation that has a "significant adverse psychological or physical impact" will need to be addressed by services in scope.

False advertising

5Rights research has found that some popular games advertised as free to play and appropriate for children, use a 'pay-to-progress' model in which players are forced to spend large amounts of money or watch in-game advertising to obtain virtual currency and continue game play.[42]

[38] What to Do When Your Video Game Gets Co-opted by Neo-Nazis, OneZero, May 2020.

[39] Free to Play? Hate Harassment and Positive Social Experiences in Online Games 2020, Anti-Defamation League, November 2020.

[40] Free to Play? Hate, Harassment, and Positive Social Experiences in Online Games, Anti-Defamation League, July 2019.

[41] Disinformation will come for animal crossing, Slate, September 2020.

[42] For example, Final Fantasy XV is rated as suitable for 10+ on the Google Play Store and 12+ on the Apple App store and has more than 10 million downloads on the Google Play store. Several recent reviews for the game comment on its pay-to-progress model, which make the game unplayable unless players are willing to spend a lot of money on the game.

The Advertising Standards Authority (ASA) have recently banned misleading advertisements for the games 'Homescapes' and 'Gardenscapes'[43] because they did not accurately represent the content of the games.

Financial and consumer harms

Gaming sites can put children at risk of financial harm through the presence of micro-transactions, loot boxes, and other in-app purchases. It is estimated between 25% and 40% of UK children who play online games have made a loot box purchase.[44] Children as young as four are spending money online[45] and 5Rights research has shown that 80% of the top 50 'free' apps deemed suitable for children aged 5 and under on the Apple UK App store contain in-app purchases. Additionally, 1 in 10 children report making in-app purchases accidentally.[46] Cases of children acquiring huge debts from micro-transactions in video games are illustrated below.

- Sonic Forces, 6-year-old, $16000[47]
- FIFA, under 17, £3000[48]
- Minecraft and Roblox, 8-year-old, £3120[49]
- Tiny Monsters, 6-year-old, £2000[50]
- Roblox, 8-year-old, £1450[51]
- The Battle Cars, Minecraft and Among us, 7-year-old, £1200[52]
- Zombie Takeover, 6-year-old, £950[53]

Individual versus collective harm

The definition of harm does not extend to collective harms felt by groups, rather than individuals. The government rightly acknowledges in the full response that intimidation and abuse in public life can stop talented people, particularly women and those from

[43] Homescapes and Gardenscapes ads banned as misleading, BBC News, October 2020.

[44] Lifting the Lid on Loot-Boxes, Chance-Based Purchases in Video Games and the Convergence of Gaming and Gambling, University of Plymouth, GambleAware and University of Wolverhampton, March 2021.

[45] Children as young as four are spending money online, The Telegraph, April 2021.

[46] Young People Losing Millions to Addictive Gaming – REPORT, Safer Online Gambling Group, August 2019.

[47] This 6-year-old racked up $16k on mom's credit card playing video games, New York Post, December 2020.

[48] Loot boxes: I blew my university savings gaming on Fifa, BBC, July 2020.

[49] Single mother-of-two, 40, tells of horror after her eight-year-old son racked up a massive £3,000 bill on her credit card in just three weeks by buying Xbox add-ons, Daily Mail, September 2019.

[50] Schoolboy, 6, runs up £2,000 credit card bill playing Tiny Monsters app on grandfather's iPad, Daily Mail, September 2012.

[51] Mum's warning as girl, eight, runs up £1,450 bill on Roblox in three days, Cambridgeshire Live, January 2020.

[52] Boy, 7, accidentally spends £1,200 on online games, including £800 on virtual cat food, Mirror, February 2021.

[53] Boy of six blows £950 of parents' money playing iPad cartoon game, Mirror, April 2012.

minority backgrounds, from going into high profile roles in journalism or to stand for public office.[54] While reform of the Communications Offences (currently under review by the Law Commission) may bring additional protections, the government must also look beyond criminal law and the actions of individuals to the design features and operating processes of services that not only allow but actively facilitate online abuse that impacts on individuals or groups with specific characteristics.

Despite the government's assertion that collective and democratic harms will be addressed by the Bill,[55] the current definition of harm leaves much disinformation out of scope. Disinformation that is designed to subvert the electoral process and algorithms that manipulate the information ecosystem present very real threats to democracy, but only in cases where disinformation and misinformation present a "significant threat to public safety, public health or national security", will Ofcom have the power to act, which is arguably too late.

The response references the Defending Democracy Programme, which is designed to protect the UK democratic processes from interference from cyber, personnel and physical threats. Its stated aim "to encourage respect for open, fair and safe democratic participation and promote fact-based and open discourse, including online"[56] can be much more easily achieved if the definition of harm in the Online Safety Bill includes cumulative harm, and specific measures to address collective and democratic harms, including disinformation and online abuse which affects minority groups, gendered or religious misinformation and groups with intersecting identities.

Harms out of scope

As stated, the Online Safety Bill should introduce a proportionate duty of care across the sector, whilst supporting business to comply with regulation. Where other legislation may be better suited for specific harms, for example pile-on harassment,[57] the government should set out a plan to introduce time-bound measures to address these harms outside of the Bill. The Bill will have failed to fulfil its purpose if, after the many years it has taken for it to become law, we are left with a list of known harms that remain unaccounted for.

Algorithms

The response does acknowledge the role of algorithms, but only in the context of promoting illegal content. This does not account for the fundamental role they play in

[54] Public sector, Human Rights Channel, date accessed 8 April 2021.

[55] On 16 March 2021, Darren Jones MP asked the Prime Minister if he could confirm that the Online Safety Bill will contain sufficient powers to tackle collective online harms, including threats to our democracy, to which the Prime Minister responded 'Yes'.

[56] Introduction, HM Government – written evidence (DAD0034).

[57] The Law Commission is reviewing whether new offences are necessary to deal with harmful online behaviour such as abusive messages, cyber-flashing, pile-on harassment, and the malicious sharing of information known to be false. The government has said it will consider if the Law Commission's final recommendations should be brought into law under the Online Safety legislation.

what children see, do and are exposed to in the digital world, including the spread and consumption of legal but harmful material. The design of algorithms and automated systems — the main drivers of user experience — are also notably absent in the outline of the proposed safety by design framework.[58]

Algorithms recommend what children might like and nudge them to make certain choices, including who to befriend and what information to believe. They are designed to extend user engagement, and in so doing, maximise commercial opportunities. The business practice of most tech companies — to increase user attention and interaction, and extend a user's network — is not always compatible with the best interests of children or the broader interests of society. Unless algorithms are brought firmly into the purview of the Bill, the mechanisms by which harms are created and spread will remain in place.

The Bill must include requirements for companies to address both the intended *and unintended* consequences of their algorithms and automated-decision making processes and the risks they create for children.[59] Harm can be caused not only by content recommendations on social media, but by the ranking, recommendation, search and targeting algorithms used on e-commerce sites, streaming services and app stores. For example, Amazon has been found to recommend knives to young people buying school rucksacks as part of their 'frequently bought together' feature.[60] While Amazon, as an e-commerce site, will be in scope of the new regulations, it is at best unclear from the proposals that as drafted will prevent similar harms as a result of its search and recommendation algorithms.

Services must consider the impact of their automated decision-making processes when conducting risk assessments and be required to report to Ofcom the input data that powers these systems, the criteria and 'goals' against which they are optimised, and the outcomes they produce. Ofcom must also have the required independence and research capability to ask the right questions of companies about the social outcomes of the algorithms, and the expertise to interrogate the answers.

5Rights will shortly be publishing its 'Pathways' report that looks at the role of system design in children's online experiences. It reveals the way in which children are offered inappropriate content and contact even when they have identified that they are a child. Pathways offers irrefutable evidence that should spur the government to take a closer look at the role of algorithms in automating and promoting harmful outcomes for children.

---

[58] "Examples of a safety by design approach include: default safety settings, clearly presented information, positive behavioural nudges and user reporting tools that are simple to use." Online Harms White Paper: Full government response to the consultation, Department for Digital, Culture, Media & Sport and Home Office, December 2020.

[59] The draft EU regulation on AI states: "certain artificial intelligence-empowered practices have significant potential to manipulate natural persons, including through the automated adaptation of misleading user interfaces, and to exploit a person's vulnerabilities and special circumstances."

[60] Amazon's 'frequently bought together' feature suggests 14-year-old buys knife with his school rucksack, The Telegraph, September 2019.

Advertising

The Online Safety Bill is an opportunity to consolidate regulation and address the repeated failure under the existing regime to prevent targeted advertising to children online. While 'user-generated' influencer adverts will be in scope, 'commercial' advertising remains under the remit of the Advertising Standards Authority (ASA) subject to voluntary rules contained within the CAP and BCAP Codes.[61] Given the central role advertising plays in the design and operating practices of most tech companies, this is a significant omission.

The exclusion of most advertising from scope will fail to address the significant harms created by the inappropriate commercial pressures to which children are routinely exposed online, from the advertising of age restricted products to the sale of personal data for real-time bidding.[62]

Research shows that tech companies have profited from the spread of misinformation through paid-for advertisements, with companies making up to $1 billion a year in advertising and other revenues from the anti-vaccine industry.[63]

Privacy International conducted research into the sharing of personal data by mental health websites. The research found that of the 136 mental health sites analysed, 97.78% contained a 'third-party' element and 76.04% contained third-party trackers for marketing purposes. Google, Facebook and Amazon trackers were present on a number of the web pages analysed.[64]

A recent investigation discovered an advertisement for a gambling service featuring a clown in an app popular among children (DOP: Draw One Part). The app's developer 'Say Games' is based in Belarus. When reported to the ASA, they responded saying UK advertising rules did not apply to marketing communications originating from foreign media.

In most circumstances, this would make the complaint outside of the ASA's remit. However, due to the exceptional circumstances of there being no possible

[61] Rule 5.4.2 of the CAP Code states "Marketing communications addressed to or targeted at children must not make a direct exhortation to children to buy an advertised product or persuade their parents or other adults to buy an advertised product for them." Put simply, marketers should not actively encourage children to buy an advertised product. Additionally, the Code states, "advertisers should ensure that there is nothing within an ad that is addressed to, targeted at or features a child that could result in a child's physical, mental or moral harm."

[62] Update report into adtech and real-time bidding, Information Commissioner's Office, June 2019.

[63] The Centre for Countering Digital Hate has shown that tech companies make up to $1 billion a year in advertising and other revenues from the anti-vaxx industry. The Anti-Vaxx Industry: How Big Tech powers and profits from vaccine misinformation, Center for Countering Digital Hate, 2020.

[64] Your mental health for sale, Privacy International, September 2019.

referral route to the Belarusian advertising regulator, the ASA was able to open a case.

The Online Safety Bill is an opportunity to bring advertising regulation onto a statutory footing, introduce proper enforcement powers and ensure children are no longer exposed to online advertising that has a detrimental effect on their physical, mental and moral wellbeing. Instead, the proposals will leave most advertising to remain out of scope of statutory regulation and create further fragmentation across the regulatory system.

Financial and consumer harms

Children are more susceptible to online scams due to the vulnerabilities associated with age and developmental capacity. Since the coronavirus outbreak in February 2020, the average number of Instagram frauds reported each month has increased by more than 50%.[65] More and more young people are also falling victim to scammers who steal photos to set up fake profiles advertising explicit images in exchange for money.[66]

There are also considerable financial risks to children in online games, notably in the form of random reward features such as loot boxes, micro payments and other insidious pay to play schemes.[67] British children collectively spent £270 million on loot boxes and other in-app purchases in 2019.[68] 76% of young video game players believe that online video games actively try to make you spend as much money as possible.[69] Loot boxes and other random reward features may be addressed in the upcoming review of the Gambling Act. If they are not, they must be brought into scope of the Online Safety Bill, so that the considerable financial harms that have devastated children and families are not left to fall between the cracks of regulation.

The widespread mislabelling of age restrictions in games downloaded through app stores must also be addressed. Enforcing consistency in age labelling is a necessary first step to ensuring children are offered age-appropriate experiences and should be relatively simple to achieve. This is another example of how restricting scope to services that host user-generated content or facilitate interaction between users creates a loophole, in which app stores that routinely mislabel content will remain out of scope of the regulation.

With financial and consumer harms out of scope, children remain at risk from online scams, gambling style features and inappropriate commercial pressures that can lead to the accrual of debt, financial losses and service/contract lock-ins.

[65] 'I was scammed out of £17,000 on Instagram', BBC News, January 2021.

[66] 'Fake accounts used my pictures to sell sex', BBC Scotland, March 2021.

[67] Loot boxes in video games – call for evidence, Department for Culture, Media & Sport, September 2020.

[68] Young People Losing Millions to Addictive Gaming – REPORT, Safer Online Gambling Group, August 2019.

[69] The Rip-Off Games: How the new business model of online gaming exploits children, Parent Zone, August 2019.

The Online Safety Bill must account for all known harms to children, both individual and collective, immediate and cumulative. Most importantly, it must look beyond malicious content and bad actors to the design features and automated decision-making processes that create risk. In the following chapter we discuss the proposals for how regulated companies will address harm and where the Bill can better tackle harm upstream.

Chapter 3: Duties on regulated companies

The new duty of care means that companies in scope will have to meet certain 'duties' to their users to keep them safe. Central to meeting the duty of care will be risk assessment and putting in place systems and processes to improve user safety.

Companies will be expected to take action to prevent user-generated content or activity on their services causing significant physical or psychological harm to individuals by:

- Completing a risk assessment to identify risks associated with their services and taking "reasonable steps" to reduce the risks of harms they have identified occurring.
- Having effective and accessible reporting and redress mechanisms
- Addressing anonymous online abuse that is illegal through effective systems and processes, and taking action to prevent the use of their services for criminal activity
- Complying with information requests from Ofcom.

In addition, category 1 services will be required to:

- Undertake regular risk assessments to identify legal but harmful material, covering both the priority categories set out in secondary legislation and any other types of harm present or at risk of arising
- Set clear and accessible terms and conditions which explicitly state how they will handle the priority categories of legal but harmful material established in legislation, and any others identified by them through their risk assessment
- Enforce terms and conditions consistently and transparently
- Publish transparency reports with information about the steps they are taking to tackle online harms.

These duties must be clearly defined and put on a statutory basis. To ensure a floor of protection that is consistent across the sector and encourage compliance, the regulator must issue templates for risk assessments and set minimum standards for terms and conditions and transparency reporting.

Child Risk Assessments

The requirement for services likely to be accessed by children to carry out regular child safety risk assessments is encouraging. The full response does not set out the considerations that will be included in child risk assessments, but it states that Ofcom's Codes of Practice will include guidance for companies on the risk assessment process. Ofcom will also need to establish the "reasonable steps" that companies will need to take following their child risk assessments, to ensure risks are properly mitigated.

No environment is entirely risk free, but a safety by design approach will ensure platforms and services are equipped to identify and mitigate risk. The 4 Cs framework classifies online risks as content, contact, conduct and contract (sometimes referred to as commercial) risks.[70] The framework has been widely adopted around the world, most recently by the Organisation for Economic Co-operation and Development (OECD),[71] the European Commission, and the United Nations.[72]

The 4 Cs


Content. A child or young person is exposed to harmful material (e.g., age-inappropriate content, pornography, extreme and real-life violence, discriminatory or hateful content, disinformation, content that endorses risky or unhealthy behaviours such as anorexia, self-harm, suicide).


Contact. A child or young person participates in activity with a malign actor, often, but not always, an adult (e.g., child sexual exploitation, grooming, harassment, stalking, blackmail, unwanted sexual advances, location sharing).


Conduct. A child or young person is involved in an exchange, often, but not always, peer-to-peer, as either a perpetrator or victim, sometimes both (e.g., bullying, sexting, revenge porn, trolling, threats and intimidation, peer pressure, loss of control of digital legacy/footprint).


Contract (also referred to as commercial risks). A child or young person is exposed to inappropriate commercial contractual relationships or pressures (e.g., compulsive use, gambling, targeted advertising, hidden costs, unfair terms and conditions, loss of control of personal data).

Figure 1. 'The Risks' adapted from Towards an Internet Safety Strategy by 5Rights Foundation

[70] The 4 Cs: Classifying Online Risk to Children, CO:RE Short Report Series: Key topics, Sonia Livingstone and Mariya Stoilova, 2021.

[71] Children in the digital environment: Revised typology of risks, *OECD Digital Economy Papers*, No. 302, OECD Publishing, OECD (2021), Paris.

[72] General comment No. 25 (2021) on children's rights in relation to the digital environment.

Done properly, the outcome of an assessment will show companies which parts of a product or service may need to be disabled, redesigned or that require risk warnings in order to keep children safe.

A child risk assessment takes an innovator or company through an eight-step process.

1. **Know your customer** - who is it that you are impacting (in this case a child or children)
2. **Scope** - interrogate the impact of your service, including the impact on any children who should not be, but are, using it
3. **Gather evidence** – create a risk management system that identifies risks, in line with children's rights and safety-by-design principles
4. **Consult** - in and outside your organisation. Solutions may come from surprising places including children themselves
5. **Assess/analyse/appraise** - what you discover may be surprising or obvious and different risks are likely to require different mitigation strategies
6. **Recommend** - this is your plan of what to do
7. **Publish and report** - transparency gives confidence to users and regulators. It also provides learning for others and sets a bar for your organisation
8. **Monitor and review** - digital products and services are rarely static. Small changes can have big impacts and constant vigilance and iteration are necessary.

These eight steps can be used to create a product, to assess an existing product or to look at the intersection between products that may together create risk. They must reveal known harms, unintended consequences and emerging risks, and take into account not only content but contact, conduct and contract risks, as per the 4 Cs risk framework.

Risk assessment is a norm across most sectors. While the process may be entirely familiar, many of the questions relating specifically to children will not be. 5Rights will be publishing a Child Impact Assessment template in summer 2021.

Safety by design

Safety by design should be the single biggest driver of a safer digital world for children. A safety by design framework has the potential to transform childhood experiences online, however in the government response it is unclear if it will carry statutory weight and form part of the Bill itself or part of Ofcom's Codes of Practice.

As the primary mechanism by which companies are to fulfil their duty of care and proactively address risk, the principles of safety by design must be set out on the face of the Bill and the implementation requirements must be legally enforceable.

A safety by design approach that is child-centred would include high privacy settings by default, age-appropriate published terms, positive nudge techniques, proportionate and robust age assurance mechanisms and the removal or omission of risky features such as private or anonymous messaging, among other measures. It also includes due diligence and risk mitigation strategies such as regular assessment of automated systems and algorithms, and effective staff training. Australia's e-safety Commissioner has published a set of principles for safety by design which would serve as a useful basis for the safety by design framework.[73] 5Rights is also working with the Institute of Electrical and Electronics Engineers to produce a standard for an Age Appropriate Digital Services Framework for publication in summer 2021, the provisions in which can be applied to a child-centred approach to safe design.

Age assurance

Companies will be expected to take "reasonable steps" to prevent children from accessing age-inappropriate content and to protect them from other harms. They will need to conduct regular child safety risk assessment to identify legal but harmful material on their services, assess the risks this material poses to children of different ages, and take proportionate action to respond to identified risks.

There is considerable confusion around when, where and how services need to assure the age of their users. The government must urgently clarify its expectations and how services can meet them. This is essential to avoid companies using excessive profiling under the guise of age assurance or introducing crude and restrictive age verification solutions that rely on parental controls, or which freeze children out of spaces they have a right to access. Equally, companies cannot be allowed to continue to use weak, 'tick-box' age assurance methods that fail to ascertain the true age of users.

The government has acknowledged that services should not use age assurance technologies to block children from content or services, but where appropriate, to protect children from a service (or part of a service) or enhance a child user's experience by tailoring features to the age of the user.

- 5Rights' paper *But how do they know it is a child? (2021)* looks at the different approaches to age assurance and the ways in it can be introduced to better protect and support children. It includes a set of standards for age assurance; that it is privacy-preserving, rights-respecting, proportionate to risk and purpose, easy for a child to use, accessible and inclusive, and enhances a child's experience, rather than merely restricts it, offering a high level of security, transparency, accountability, and clear routes to challenge and redress.

[73] Safety by Design, eSafety Commissioner, May 2019.

The government has indicated that the aims of the Digital Economy Act 2017[74] for commercial providers of pornography "to have robust age verification controls in place to prevent children and young people under 18 from accessing pornographic material",[75] will be fulfilled by the Online Safety Bill.
This is a promise long overdue, and the effective and trusted use of age assurance solutions must be mandated to prevent children being exposed to pornography at a young age.

It is imperative that the government sets out standards in regulation to bring clarity to when age assurance is needed, and credibility and consistency to the different age assurance tools and solutions available. Age assurance without trust will benefit neither industry nor children. Trust can only be achieved with a mixed economy of age assurance solutions that reflect the multiple scenarios in which age assurance is required, backed up by enforced regulatory standards.

Moderation, reporting and transparency

Effective content moderation, reporting and removal are central to the duty of care. But these measures focus on addressing harm *after it has been caused* and do little to tackle the systems and processes that create risk. The focus must be shifted upstream to address the technical practices and operating systems that facilitate the spread of harmful content through risky recommendation systems or pernicious design features.

Moderation

The response does not go into detail about the standards of moderation that will be required for companies to fulfil the duty of care, but it does say that companies will need to consider the systems and processes they have in place to address harmful content, including their moderation procedures.

Ofcom's Codes of Practice must be mandatory and require companies to invest in effective (including human) moderators, relative to the scale and nature of the business, and in more sophisticated AI moderation systems to mitigate the spread of harmful content before it reaches new audiences. The Codes of Practice must demand a higher bar of standards for moderation, putting greater emphasis on support and mandate that all moderation systems are linked with user reporting and redress.

[74] Digital Economy Act 2017.

[75] Online Harms Statement, The Rt Hon Baroness Nicky Morgan, 16 October 2019.

R;pple Suicide Prevention
R;pple is an online monitoring tool currently in development that redirects users searching for harmful keyword or phrases to different forms of mental health support.[76] The R;pple tool acts as an interception guiding users away from self-harm or suicide content towards mental health resources. This will provide a much stronger intervention than is currently offered by search engines and demonstrates the kind of moderation tools to protect users from harmful content.

Figure 2: Screenshot from Ripple Suicide Prevention, obtained 8 April 2021.

User reporting and redress

Companies will be required to put in place effective and accessible user reporting and redress mechanisms for content and activity which is harmful to children. Mechanisms for reporting should be readily available, but reporting puts the responsibility for addressing harm on those experiencing it. In many cases, it is not possible for a child to understand a service's community standards (in which they define what kind of content and behaviour is not allowed), to recognise why something is harmful or inappropriate, and to know how to report it.

While swift, effective reporting is an important provision for children, relying on user reporting requires a child to understand the harm and that they have a right to be treated differently. This is simply not the reality for many children, who may feel shame, who may not understand what is happening, who might be scared their device will be taken away, or who may not trust that the system will take care of them. Investment in

[76] How the R;pple tool works, R;pple Suicide Prevention, 2021.

moderation should be a price of doing business, but to prioritise investment in reporting mechanisms, at the expense of looking at systems and practices that create risk or result in harm, is antithetical to the government's stated aim of reducing risk for users online.

Transparency

> Category 1 services will be required to publish transparency reports and provide information to the regulator about the steps they have in place to assess risk of harm.

Category 1 services make up a tiny percentage of digital services, and while transparency does not in itself offer accountability, in conjunction with the duty of care and safety by design approaches, it carves a path to setting standards and ensuring that they are met. Transparency requirements should be adopted across the sector, irrespective of the nature and scale of the service. These requirements must be met with the necessary resources and capability of the regulator to respond supportively and proportionately, and enforce sanctions when needed.

Central to transparency reporting must be the requirement to publish information about the algorithms used to prioritise, recommend and moderate content, rank search results and target and profile users. These algorithms are very often the main drivers of harm and the way they are deployed is largely unknown to both users and regulators. Services must make clear the purposes for which algorithms are deployed, the data they are using to feed the algorithms, and the outcomes they produce. Importantly, these outcomes must include both the intended and unintended consequences. With transparency must come the required regulatory resources and expertise to understand the effect of automated decision-making systems, to identify appropriate actions and steps to take to mitigate harm, and sufficient enforcement powers to enact those measures. Companies must not be permitted to use commercial sensitivity to avoid transparency obligations, and where there are commercial sensitivities, the regulator must have the power to maintain private oversight.

Ofcom will also be responsible for producing an annual report to summarise key findings and insights from the reports that companies have produced. The response suggests this report will help "users and parents understand the differences between online services and make informed decisions about which ones they use."[77] Young people rarely have any *real* choice about the services they use, from social pressure or simply because certain services are essential for their education, access to healthcare

[77] Transparency, Online Harms White Paper: Full government response to the consultation, Department for Digital, Culture, Media & Sport and Home Office, December 2020.

or way of life.[78] It is not advisable or fair to expect children or their parents to choose the services they use based on transparency reporting.

The value of transparency reporting lies not in the availability of information itself, but in the way it enables the regulator to hold companies to account after that information has been understood and appropriate enforcement measures have been taken. As currently framed, the government has not grasped the opportunity of transparency reporting to drive cultural change for the sector to take responsibility for the safety of their services.

Terms and conditions

Category 1 services will be required to establish what is acceptable on their services in their terms and conditions and explicitly state how they will address legal but harmful material. They will be expected to consult with civil society and expert groups when developing their terms and conditions, to ensure they meet user needs and build on existing best practice on how to effectively tackle different types of harmful content and activity.

Services likely to be accessed by children will also need to make specific provisions for their users when presenting terms and conditions and mechanisms for redress.

Ofcom's Codes of Practice will define what appropriate levels of risk-based and proportionate protection for children look like and set out the measures companies need to take, as set out in their terms and conditions.

If services are to address legal but harmful content and activity in their terms and conditions, and if their fulfilment of the duty of care will be judged on the content of those terms, it is imperative that Ofcom sets out minimum standards for published terms and the way they are presented. Ofcom should also mandate that terms are presented in forms that are accessible to young people, at a time when they are most likely to read them.

The Bill must set out the exact range of issues that published terms are supposed to cover, and what the steps are for compliance and enforcement. It should also be clear that what is generically referred to as 'terms and conditions' are often presented in multiple documents (22 documents in some cases)[79] many or most of which would have to be read to understand the full agreement between the child and the company.

[78] 60% of respondents to an open public consultation on a New EU Competition Tool said that consumers don't have sufficient choices and alternatives regarding online platforms. Consultation outcome, Single Market – new complementary tool to strengthen competition enforcement, European Commission, 2020.

[79] Twitch, a global livestreaming service, has 22 different published documents to make up their terms and conditions. A handful of separate terms include: Community Guidelines, Privacy Notice, Privacy Choices, Ad Choices, Channel Points Acceptable Use Policy, Cookie Notice, Bits Acceptable Use Policy, Accessibility Statement and Transparency Report. See more from: Legal, Twitch, date accessed 1 April 2021.

Codes of Practice must set minimum standards for both the content and presentation of published terms, or we risk recreating a regime in which companies are allowed to mark their own homework.

Chapter 4: Duties and powers of the regulator

Ofcom will be responsible for setting out how companies in scope are to fulfil the duty of care in statutory Codes of Practice and will have the power to assess whether the steps they have taken to do this are sufficient. It will have a duty to consider the vulnerability of children and other vulnerable groups and will need to prioritise the protection of children in its approach to enforcement. Importantly, it will accept super-complaints from representatives taking action to address systemic issues affecting children.

Ofcom will have powers to:

- Require companies in scope to publish annual transparency reports
- Gather additional information from companies to inform its regulatory activity
- Interview tech company employees to understand how the company is complying with the duty of care
- Issue directions for improvement and notices of non-compliance
- Levy fines of up to £18m or 10% of global turnover
- Take measures to disrupt a company's business activities in the UK, including blocking access entirely
- Pursue enforcement action against a parent company that owns or controls the non-compliant company.

It is not yet sufficiently clear what the Codes of Practice will cover, how much of the Codes Ofcom will be empowered to impose nor the full extent of their powers to enforce. As previously stated, the Codes of Practice should be seen as intrinsic to the duty of care and should include a safety by design framework that is enforceable and applicable to any digital product or service that impacts children.

Enforcement

The efficacy of the regime will be determined in large part by how it is enforced. Ofcom must be sufficiently empowered and resourced to enforce the new regime. In particular, it will need in-house expertise and sufficient oversight powers to understand and audit the automated decision-making systems used by services. It must also have the power to interview and engage with those who design systems and are responsible for their impact. Without the requisite expertise, Ofcom will not be able to take a systemic approach to reducing risk.

Codes of Practice

The Codes of Practice will set out how companies should:

- assess the risk of harmful content and activity occurring
- put in place appropriate governance systems for managing risk
- moderate content for different types of harmful content
- implement tools to support users to manage harm
- put processes in place to allow users to report harmful content or activity and to appeal the takedown of their content
- understand the impact of online safety measures on freedom of expression and introduce appropriate mitigating measures.

There is too great an emphasis on addressing harm after it has been caused and not enough on risk assessment and mitigation. Beyond content moderation, reporting mechanisms and appropriate governance, the Codes of Practice must set a high bar for 'designing out' all categories of risk. For companies to make effective changes to the design and operation of their services, these Codes of Practice must be practical and implementable. Codes must include standards for age assurance, child impact assessments, risk management and mitigation, safety by design, presentation of terms and community rules, transparency and user support and redress.

Consultation with relevant stakeholders, particularly with children and other vulnerable users, is critical to ensure the Codes address all risks that users face, including new and emerging risks. All Codes must be on a statutory footing, such are their importance in reducing the risk of harm.

Company director liability

The proposed regulatory framework reserves the right to issue criminal sanctions against individual company directors, but only when they have failed to comply with information requests from the regulator.

As set out in the response, the government would need to be persuaded to introduce director liability on the basis of significant failure across the market. It is highly improbable that the egregious and continued failure by one company would result in sanctions being brought in for all companies in scope. What is more, the government has delayed the introduction of the power until at least two years after the regulation comes into force, following a review of the regulatory framework. This is not the swift, sharp, regulatory action needed when companies fail to engage.

The Online Safety Act should follow the precedent set by the Gambling Act 2005[80] and the Companies Act 2006[81] to hold individual responsible directors to account where they have failed to comply with the regulatory regime. Without individual director liability, it is hard to see how the largest tech companies, whose enormous wealth and cash reserves can easily absorb even the heaviest fines, will be sufficiently incentivised to comply with the duty of care.

User complaints and action

The proposed legislative framework will not empower Ofcom to consider individual complaints or arbitrate on individual cases, nor is there any indication that the government intends to establish an independent resolution mechanism. This means that there is no pathway for an individual or group to take action on the basis of a breach of the duty of care. While Ofcom will accept super-complaints, this does not tally with the conventional understanding of a duty of care in negligence law — it is more closely related to a conventional model of statutory regulation. It also deviates from the established process in data protection law, where the ICO investigates individual cases of alleged infringement of data protection laws.

The detail of the super-complaint system has not been laid out, but it is clear that it must be specifically designed to work for children. Without help, children cannot understand the risks that services pose, how they can complain when things go wrong, or even what constitutes a violation.

It is crucial that the legislation allows charities and others to represent children in super-complaints without having to identify individual users who have suffered harm. This would allow systemic issues to be resolved, without having to have children who understand and can articulate the harms they have suffered as a result of the breach of the duty of care and who may then have a permanent digital footprint linking them with that breach.

In the absence of such a super-complaint system Ofcom's remit must be extended to include the power to take individual complaints from minors, unwelcome and unwieldly as that would be.

Duties to businesses and promoting innovation

Ofcom will need to assess the impact of Codes of Practice on small businesses to ensure regulatory requirements are "proportionate" and do not place an "undue burden" on businesses. We welcome proportionate regulation and support innovation that recognises and anticipates child users. However, the government's emphasis on 'unburdening business' in a Bill designed to reduce online harms could undermine its

[80] Part 5 and Schedule 7 of the Gambling Act 2005 concern operating licences issued by the Gambling Commission, including powers to revoke licenses and impose financial penalties.

[81] Companies Act 2006.

commitment to "make the UK the safest place in the world to go online."[82] Innovation is crucial for a thriving economy and should be supported by the government, but the Bill is being brought forward because the tech sector has failed to put safety before profit. The tech sector is made up largely of private businesses, and like all other private businesses, they should be subject to the laws, expectations and duties that ensure they act in a manner that protects their users, particularly the most vulnerable.

An Online Safety Bill is no place to pay tribute to the "move fast and break things" gospel of disruptive tech. Nor is it helpful to frame innovation as fundamentally at odds with the design of responsible and safe services. Indeed, designing safe and responsible digital services is good for business, good for growth and good for market longevity. Smaller companies need more support and 'off the shelf' tools from the regulator rather than permission to harm.

Digital and media literacy

The government has committed to publishing an Online Media Literacy Strategy to promote greater media literacy among children and young people and to support parents to understand and prevent the risk of harmful activity online.

The full response states "internet users want to feel empowered to manage their own online safety." A child cannot and should not be expected to manage their own online safety in an environment that is not designed for their safe use. For example, companies routinely introduce children to unknown adults and allow direct messaging that is known to enable grooming but expect children and parents to be responsible for preventing unsolicited contact.

Children want digital environments to be made safer, so that they can play, learn and explore without fear of harm.[83] Empowering users to be responsible actors in the digital world is crucial but neither children, nor their parents, should be made to feel responsible for mitigating the harms that must be addressed by the Online Safety Bill at the level of system design.

This tendency to responsibilise children for their own safety is often present in digital literacy programmes run by private companies or interest groups. Programs such as Google's 'Be Internet Awesome'[84] and Facebook's 'My Digital World'[85] offer resources and workplans to schools at little or no cost, both in the UK and around the world, but teach children to accept certain service design elements as 'unavoidable' risks when in fact they could and should be tackled at a design level by those very same companies.

[82] Conservative Manifesto, 2019.

[83] 5Rights: By Young People, 5Rights Foundation.

[84] Be Internet Awesome with Google, Google.

[85] Facebook launches virtual digital literacy programme with CcHUB, JAN, others, BENJAMIN DADA, June 2020.

Ofcom's media literacy evaluation framework must set standards for providers of digital literacy programs to children, particularly those who are, or are funded by, tech companies.

Digital and media literacy should support children to understand the purposes and outcomes of the technologies they use, both through clear presentation of information at design level ('literacy by design') as well as through the teaching of digital and in particular, data literacy, which includes understanding the purposes, common practices and likely outcomes of digital engagement. Digital literacy programmes must give young people the ability to understand the common uses and abuses of technology. Importantly, they must understand that risks are created not only by adult content or 'bad actors', but also by the design and operation of platforms and services.

Ultimately, it must not be that we educate children to navigate a digital world that is inherently unsafe, and which systematically puts them in harm's way. Online safety and digital literacy are supplementary to the fundamental shift needed in the design and operation of online services to improve safety and should support children to be sophisticated users of a system that is already fit for purpose. The online safety Bill offers a unique opportunity for this narrative shift.

Upholding children's rights

The UN Convention on the Rights of the Child is the single most important expression of children's rights and the needs of childhood,[86] and is the basis for much domestic legislation in relation to children around the world. The UN Committee on the Rights of the Child adopted general comment (No. 25) 2021 on children's rights in relation to the digital environment on 2nd March 2021.

The Committee's adoption of the general comment makes explicit that children's rights apply in the digital world. It sets out how legislators and regulators should implement and account for children's rights in the digital world and their obligation to ensure businesses respect the rights of children. As a signatory to the Convention, the UK government should, through the Online Safety Bill:

- Make businesses responsible for ensuring they are responsive to the presence of children, including all digital services that impact on children, not just those designed for them or which they choose to use
- Provide default ways of enabling and protecting children including privacy, safety, and security by design; in effect, child rights by design (this includes training needs for all relevant professionals)
- Understand the digital world as interconnected and pervasive, considering a broad range of technologies, business practice and contexts

[86] Convention on the Rights of the Child, OHCHR.

- Take a holistic approach that emphasises the full range of children's rights including their civil rights and freedoms, and that gives necessary attention to protection without becoming restrictive or overly-protective
- Tackle the commercialisation of childhood, with provisions on child labour, data exploitation, advertising, profiling etc
- Emphasise processes (general measures) designed to ensure effectiveness: due diligence, child rights impact assessments, independent monitoring, oversight and accountability, and child-friendly remedy and redress
- Seek to ensure that discrimination is neither tolerated nor automated and explicitly mentions the many groups who are currently discriminated against, mandating their inclusion
- Consider emerging technologies from rights perspective and seeks to be future-proof (at least the next 5-10 years) by using technology-neutral language
- Demand tailored data protection for children, drawing on and advancing international best practice and recognising that privacy is both a right and a means to children's other rights in a digital world
- Emphasise the rule of law, regulation and standards to drive a better experience for children. Any actions must be evaluated in advance for possible costs to children's other rights and must be lawful, proportionate and necessary
- Recognise that all children are not the same and insist on creating and maintaining a disaggregated evidence base to target interventions and evaluate improvements.

Every four years, state parties are required to report to the Committee on the Rights of the Child how they have implemented the Convention and its protocols. The UK is due to make its latest report during 2021. The Government must now seize the opportunity to use the Online Safety Bill to fully implement the Convention rights as set out in general comment (No. 25) 2021 on children's rights in relation to the digital environment.

Timing and the wider regulatory landscape

While welcoming Ofcom's role, there is an undeniable danger of both regulatory overlap and gaps between the various regulatory authorities with online harms in their remit. These include the Advertising Standards Authority, the Competition and Markets Authority, the Gambling Commission and the Information Commissioner's Office. Cohesion across the regulatory landscape will be critical if the government is to fulfil its manifesto commitment "to make the UK the safest place in the world to be online."[87] The establishment of the Digital Regulation Cooperation Forum (DRCF) in July 2020 between ICO, CMA and Ofcom recognised the importance of cooperation and alignment. The efforts of the DRCF to develop joined-up regulatory approaches and build shared skills and capabilities should be supported by a duty on the face of the Bill for Ofcom to

[87] Conservative Manifesto, 2019.

co-operate with other agencies in the regulation of online harms. This kind of duty exists in other legislation, for example, the Parliamentary Standards Act 2009.[88]

The Bill will have a number of dependencies, including proposed reforms to the communications offences, the review of the Gambling Act 2005 and a new statutory code of conduct brought in by the CMA's Digital Markets Unit.[89] These dependencies are likely to have an impact on the priority categories of harm contained within secondary legislation, and on the timing of the Bill passing into law. As stated previously, any known harms that are not included in the Bill should be accounted for in other legislation.

Timing

It is unclear when the draft Bill will finally be laid before Parliament or when it will be passed into law. Many key elements of the regulatory regime are to be contained in secondary legislation. This is likely to involve further consultation, drafting and decision-making before we have clarity on some of the most important aspects of the new regulation. Following the passage of the Act through Parliament, responsibility will be placed in the hands of Ofcom for setting out how companies are to fulfil the duty of care.

Many of the provisions for children under the new regime will be contained in secondary legislation, Codes of Practice or accompanying guidance. If we look at the timeline for the Age Appropriate Design Code[90] as an indication, it will be some time before the changes brought in under the Online Safety Bill take effect. Amendment 123 of the Data Protection Act 2018[91] which required the ICO to create the Code was laid in 2017, the section was commenced under a Statutory Instrument in July 2018, the draft Code was laid before Parliament in June 2020 and will enter into force fully in September 2021.

The problems children face now need to be addressed urgently. It is likely that a child of 11 getting their first (or second) smartphone will be 15 or 16 by the time the new regulation has come into force and its effects are felt. The government needs to use existing powers to offer immediate protections for children. It should also make a commitment to develop Codes of Practice concurrently with the Bill so they can be implemented without further delay as the Bill becomes law.

[88] Which under section 10A requires the Independent Parliamentary Standards Authority to work with other bodies including the Director of Public Prosecutions and Commissioner of the Metropolitan Police. Parliamentary Standards Act, 2009.

[89] New competition regime for tech giants to give consumers more choice and control over their data, and ensure businesses are fairly treated, Press Release, Department for Business, Energy and Industrial Strategy, November 2020.

[90] Age appropriate design: a code of practice for online services, Information Commissioner's Office, 2020.

[91] Data Protection Act 2018.

Protecting regulation

The government should make a commitment to protecting online safety legislation in any future trade negotiations. It would follow the example the government has set by putting children's protections on the same footing as workers' rights and the environment in the recent amendment to the Trade Bill.[92]

With the ongoing debate in the US about Section 230,[93] it is more important than ever that the protection of children online is put above the commercial interests of big tech firms. Section 230 is already controversial and has been criticised for giving tech firms the latitude to ignore the law and the needs of users. In Canada, the free trade agreement between the United States, Canada and Mexico saw the inclusion of Section 230-style protections for tech firms. Canada is the base for Pornhub, the largest pornography site in the world. When Pornhub was found to be monetising child rape and child sexual abuse material,[94] the Canadian Government representative in the Senate, Senator Marc Gold, had to admit that "there are provisions in the North American Free Trade Agreement that make it difficult to deal with a company like Pornhub."[95]

Encouragingly, both Republicans and Democrats want change, and the US Supreme Court has criticised the way Section 230 lets online services off the hook for promoting illegal content, and for refusing to police their own platforms. While US Congress is likely to consider reform, it is not a given that the new administration will act swiftly or in the UK's interests. The voice of the tech lobby is powerful, and it is vital that that the principle of non-regression is applied to the protections for children contained in the Online Safety Bill.

[92] Online protections for children and vulnerable users were included as an area requiring statutory protection in any international trade agreement under Amendment 6B. Division 5, Trade Bill, February 2021.

[93] Section 230, US Communications Decency Act 1996.

[94] The Children of Pornhub, New York Times, December 2020.

[95] 2nd Session, 43rd Parliament, Volume 152, Issue 20, Senate of Canada, December 2020.

Conclusion

An Online Safety Bill worthy of its name must address the reality of the impact of digital technology on children's lives. As the specific requirements under the duty of care are drafted, the government must ensure that companies take responsibility for the risks they create. They must look beyond harmful content to tackle upstream the systems and design features that engineer risk into products and services.

The scope of the Bill must be extended to provide proactive duty of care for all places, products and services that impact on children regardless of the size or nature of their business. SMEs and start-ups should be supported with tools and products provided by the regulator and the regulator must take a proportionate risk-based approach so these companies can fulfil their responsibilities to children. Where specific harms are better dealt with in other legislation, the government must set out how these harms will be addressed, in what timeframe and where, without creating a fragmented and complex regulatory landscape that creates additional compliance challenges for businesses.

The efficacy of the measures set out in legislation will come down to how they are enforced. Ofcom needs to exercise its power as the regulator if it is to be taken seriously by tech companies. Any regulator has a duty to take a proportionate approach to enforcement under the Regulators' Code, so the emphasis on keeping business 'on-side' is unnecessary and risks undermining the very purpose of the Bill.

The success of regulation, and therefore the Bill itself, will be measured by the ability of young people to engage with the digital world safely and in line with their rights and development capacity. As the details of the Bill are worked over in the coming months, the government has the opportunity make good on its promise to provide the highest level of protection for children online. Requiring online services to provide safe, secure environments by design is not aspirational. It is what children, parents and civil society expect, and it is the very minimum we should expect of an Online Safety Bill.

Building the digital world that young people deserve

5RIGHTS FOUNDATION

Website: 5rightsfoundation.com
Twitter: @5RightsFound
Email: info@5rightsfoundation.com