

Your response

Please refer to the sub-questions or prompts in the <u>Annex</u> of our Call for Evidence.

Question	Your response
Question 1: Please provide a description introducing your organisation, service or interest in Online Safety.	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. While 5Rights works exclusively on behalf of and with children and young people under 18, our solutions and strategies are relevant to many other communities.
	Our focus is on implementable change and our work is cited and used widely around the world. We work with governments, inter- governmental institutions, professional associations, academics, businesses, and children, so that digital products and services can impact positively on the lived experiences of young people.
Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services?	Confidential? – N Prevalence of illegal content Evidence suggests the presence and quantity of certain types of illegal content, such as Child Sexual Abuse Material (CSAM), is rising sharply. The Internet Watch Foundation (IWF), for in- stance, has this year reported a 186% increase in sexual imagery involving 7–10-year-olds, and has recorded an "exponential increase" in self- generated content created using webcams or smartphone cameras. During the same period, an 137% increase in self-generated imagery specifically showing boys aged between 7-13 years old was also documented. The amount of CSAM detected by the charity reached record

levels during 2021 - rising to 15 times that detected 10 years ago.¹

Problem of defining "illegal content" The Online Safety Bill categorises content into illegal and legal content, but in practice these parameters are often difficult to distinguish. While certain online content is undisputedly illegal, such as CSAM, the concept of "illegal content" is ambiguous. It can be difficult to determine the intention behind and the intended audience for content posted online, challenges that are more prevalent in the online than the offline world. Existing criminal law is out of step with the digital world, and law enforcement agencies struggle to apply the criminal law to online settings.²

Codes of Practice should clarify types of content and activity online that "amounts to" an offence. The government's amendment NC14 attempts to place some of the responsibility on service providers to distinguish between illegal and legal content.³ This amendment will put the burden on service providers to interpret what amounts to illegal content on the basis of whether the thresholds for committing an offence have been met to their knowledge. This is likely to lead to some providers taking an overzealous approach to content removal and others to leave up content if they do not have enough relevant information to conclusively determine whether the content has met the high threshold of illegality.⁴ Ofcom must clarify what content meets the threshold of illegality as well as how content which falls short of this threshold should be addressed.

Drivers of illegal content

Platform design that amplifies and facilitates virality enables the spread of illegal content. Features such as livestreaming and direct messaging, meanwhile, allow for easy and direct com-

¹ 'Understanding how platforms with video-sharing capabilities protect users from harmful content online', Ernst & Young LLP, 08/2021 link

² 'Modernising Communications Offences A final report', Law Commission, 20/07/2021, <u>link</u>

³ 'Fact sheet on changes to the illegal content duties within the Online Safety Bill', DCMS, 23/08/2022 link

⁴ 'Fact sheet on changes to the illegal content duties within the Online Safety Bill', DCMS, 23/08/2022 link

munication between adults and children, creating significant safeguarding risks. In order for children to be protected from illegal content online, Ofcom's guidance must focus on the design features and operating practices that account for the rapid increase in the presence and quantity of illegal content online.

Enforcement of terms of service

Most mainstream service providers have community guidelines and terms of use which stipulate that illegal content, from infringement of copyright to CSAM and hate speech, are not allowed on their services. Yet these terms of service can be changed instantaneously to permit the presence of illegal content. For example, Meta temporarily allowed the use of hate speech on its platform, in the context of Russia's invasion of Ukraine.⁵ Further, existing terms and conditions are also not consistently applied across all users. Tweets initially posted by public figures were left online whilst the same tweets posted by a dummy account were removed by Twitter for violating its community guidelines "on violence and incitement."6

Reporting and redress mechanisms

Children are not always able to determine whether content meets the threshold of illegality, and only a third of children know how to use online reporting or flagging functions (32%) and just 14% had ever used them.⁷ Even when children do report content, they often have no way of checking the status of their report, if it is ever addressed. A 14-year-old member of the 5Rights youth advisory group shared their experience of reporting content to an online service: "I would report it and block it all, but I'd never know if that person was taken off. There was a point where I just stopped reporting it and making complaints and stuff, because what's the point?"

Inadequate age assurance

A lack of effective age assurance on services enables adults to have access to child-only

⁵ 'Meta's decision to allow hate speech against Russians is troubling and can impact other conflict areas, say experts', Thomson Reuters, 17/03/2022, link

⁶ Big Brother Watch Twitter, 11/03/2022 <u>link</u>

⁷ 'Children and parents: media use and attitudes report 2022', Ofcom, 30/03/2022 link



⁸ 'Pathways: How Digital Design puts children at risk',5Rights Foundation 07/2021 <u>link</u> P.21

⁹ 'Pathways: How Digital Design puts children at risk',5Rights Foundation 07/2021 <u>link</u> P.63

 $^{^{10}}$ 'Facebook apps used to groom 24 children every week', The Telegraph, 10/2021 $\underline{\mathsf{link}}$

 $^{^{11}}$ 'Facebook apps used to groom 24 children every week', The Telegraph, 10/2021 $\underline{\sf link}$

¹² 'Record high number of recorded grooming crimes lead to calls for stronger online safety legislation', NSPCC, 08/2021 link



¹³ 'What Happened When Dylann Roof Asked Google For Information About Race?', NPR, 10/2017 <u>link</u>

¹⁴ 'Amazon's 'frequently bought together' feature suggests 14-year-old buys knife with his school rucksack', The Telegraph, 09/2019 link

¹⁵ 'HIDDEN HATE: What Google searches tell us about antisemitism today', Antisemitism Policy Trust, 01/2019 <u>link p.8</u>

¹⁶ 'The 2022 Dirty Dozen List', National Centre on Sexual Exploitation, 03/2022 Link

must also reflect the gendered elements of illegal content. Over three-quarters (79%) of young women have experienced online harm in 2021 in the form of sexist comments (35%), cyberflashing.¹⁷ This trend of gender-based online abuse and harassment is corroborated by research from Revealing Reality which showed that more than a third of the girls surveyed (5,000 14-18 year olds) said they had first been asked to send a nude image when they were 13 or younger.¹⁸

Pressure on law enforcement resources

Whilst the amount of illegal content online increases, law enforcement's capacity and resources have not increased to meet this challenge. The police foundation report found law enforcement is "simply overwhelmed" by the amount of CSEA cases.¹⁹ If companies design features and services with safety in mind this will help tackle harm upstream consequently relieving pressure on law enforcement.

Emerging risks

As user-to-user services and search services evolve, so do the forms of illegal content found on them. Evidence suggests that illegal content and activity is prevalent in augmented reality environments, notably in the Metaverse. A BBC researcher who set up an account registered as a 13-year-old witnessed grooming, sexual material, racist insults and a rape threat.²⁰ Search services through connected devices also present a new range of risks of illegal content. Amazon's smart speaker, Alexa, suggested the dangerous viral trend that first spread on Tik-Tok known as "the penny challenge" to a 10year-old girl using the device.²¹

¹⁷ 'ONLINE VIOLENCE AGAINST GIRLS AND WOMEN TO BE INCLUDED IN ONLINE SAFETY BILL' GirlGuiding, 02/2022 <u>link</u>

¹⁸ 'Not Just Flirting', Revealing Reality, 07/2022 <u>link</u>

¹⁹ 'TURNING THE TIDE AGAINST ONLINE CHILD SEXUAL ABUSE,' The Police Foundation, 07/2022 <u>link</u>

 $^{^{20}}$ 'Metaverse app allows kids into virtual strip clubs', BBC 02/2022 <u>link</u>

²¹ 'Alexa tells 10-year-old girl to touch live plug with penny', BBC, 28/12/21, link

Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content presented by your service?	Confidential? – Y / N
Question 4: What are your governance, accountability and decision-making structures for user and platform safety?	Confidential? – Y / N
Question 5: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?	Confidential? – Y / N SRights research has shown that many services popular among children and young people set out their terms of service in legalistic documents, sometimes over 11,500 words in length with a 'readability' score requiring a university education. The Age Appropriate Design Code already re- quires services likely to be accessed by children to provide their published terms, policies and community standards in concise, prominent formats with language suited to the ages of child users. ²² To fulfil this requirement, provid- ers should ensure terms are comprehendible, of an appropriate length, clearly presented, easy to find, introduced at the right moments, and understandable to all young people, no matter who they are, how old they are, or where they come from. In particular: - Services should avoid the use of jargon and define words or phrases that are likely to be unfamiliar. They should either be pitched in language that the youngest likely user can readily comprehend, or offered in different versions tailored to different age groups.

²² <u>Policies and community standards</u>, Age Appropriate Design Code, Information Commissioner's Office



23

Web Content Accessibility Guidelines (WCAG) 2.1, Web Accessibility Initiative, World Wide Web (W3C)

²⁴ <u>Safety Hub</u>, Yubo

²⁵ <u>Twitter Data Dash</u>, Twitter

	which providers can deploy to make their terms suitable for children of all ages. ²⁶
Question 6: How do your terms of service or public policy statements treat illegal content? How are these terms of service maintained and how much resource is dedicated to this?	Confidential? – Y / N
Question 7: What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms?	Confidential? – Y / N Providers should recognise that reporting and complaints mechanisms are not in themselves sufficient risk mitigation measures, but rather tools to account for situations when something has gone wrong. That said, inadequate reporting and complaints mechanisms are central to the widespread lack of trust children have in online services. Members of 5Rights' youth advisory group frequently share their experiences of trying to report other users for violations of terms and community guidelines, and their frustration at the lack of transparency around the complaints process. Often, we hear that young people have stopped reporting harmful content or activity because the process is too onerous and frequently comes to nothing.
	To provide adequate reporting and complaints mechanisms, services should have prominent and accessible tools, tailored to different age ranges, that are highlighted during the sign-up process with instructions for how to use them. When they are moved to report, complain or seek redress, children and their parents should also have access to expert advice to help them understand their rights and support their decision-making. Users should be given expected response times that are proportionate to the seriousness of the report being made, including by responding immediately to children who appear to be in distress. To prevent the onus being placed on children to chase up the status of their

²⁶ <u>Tick to Agree: Age Appropriate Presentation of Published Terms</u>, SRights Foundation

	complaint, providers should inform them of action taken, including by granting access to the status of the reports, communicating actions clearly and offering an opportunity to provide feedback.
Question 8: If your service has <u>reporting or</u> <u>flagging</u> mechanisms in place for illegal content, or users who post illegal content, how are these processes designed and maintained?	Confidential? – Y / N
Question 9: If your service has a <u>complaints</u> mechanism in place, how are these processes designed and maintained?	Confidential? – Y / N
Question 10: What action does your service take in response to <u>reports</u> or <u>complaints</u> ?	Confidential? – Y / N
Question 11: Could improvements be made to content moderation to deliver greater	Confidential? – Y / N
protection for users, without unduly restricting user activity? If so, what?	As with reporting and complaint mechanisms, moderation is one tool among many to keep users safe, but should never take the place of upstream risk mitigation measures that reduce the likelihood of harm. By taking a safety by design approach, providers will be less dependent on after-the-fact moderation and less likely to be drawn into disputes about unfair content removal.
	Providers should deploy human moderators who have received training in how to identify risks to child safety, including knowledge of risks to different groups of children and the full range of content and activity that is illegal or might be harmful to a child. This also includes knowledge of the stages of child development and awareness of how children's capacities, vulnerabilities and behaviour change as they grow. Training must also cover the service's policies and what constitutes a violation of published terms, and how they can be implemented effectively and fairly. Moderators should be aware of when and how to intervene and to whom issues should be escalated when appropriate.

	Where possible, providers should make use of automated technology, ideally at the point of upload, to detect and remove illegal and harmful content and activity before users are exposed to harm. Investment in all moderation systems should be proportionate to the risks associated with the service, and should be subject to regular evaluation to ensure good practice is upheld and investment directed at the parts of the service most in need of it. If responsible user behaviour is to be encouraged, clear penalties must be applied fairly and consistently, and users should also be given the opportunity to appeal decisions or escalate unresolved appeals to expert third parties or regulators. Livestreaming app Yubo uses real-time technology to intervene at the most opportune
	moments, explaining to users why warning messages have appeared, a livestream taken down or other action taken.
Question 12: What automated moderation systems do you have in place around illegal content?	Confidential? – Y / N
Question 13: How do you use human moderators to identify and assess illegal content?	Confidential? – Y / N
Question 14: How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services?	Confidential? – Y / N
Question 15: In what instances is illegal content removed from your service?	Confidential? – Y / N
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	Confidential? – Y / N
Question 17: What other sanctions or disincentives do you employ against users who post illegal content?	Confidential? – Y / N

Question 18: Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry?

Confidential? - Y / N

Rather than 'tacking on' safety features after a product or service has been developed, providers should take a safety by design approach and consider the best interests of children, including their rights to privacy, to freedom of expression and thought, to participation and to protection against all forms of discrimination and exploitation, at each stage of the design and development process.

Fundamental to harm prevention is risk identification and mitigation. Providers should assess the risks presented by each feature of their product or service, and of their service in the round, to identify known harms, potential risks and unintended consequences, considering both the likelihood of harm occurring and the severity of harm when it does occur. Through this process of risk assessment, providers will be able to identify features and functionalities that may need to be redesigned or disabled to keep children safe. This also offers providers the opportunity to consider positive changes to deliver enhanced, age-appropriate experiences for younger users.

5Rights' project Risky by Design examines design features, common to the services children use, that create risk for young people. It illustrates how these design features, driven primarily by commercial interests, can lead to harm, and includes practical suggestions for how these risks might be eliminated, mitigated or effectively managed. It categorises the risks into four groups known as the 4 Cs of online risk to children: content, contact, conduct and contract/commercial risks.

Content risks

Features that enable users to block, limit or filter certain keywords or types of content can prevent their exposure to harmful content. Users should be given the option of switching off comments and recommended content. For instance, Instagram has introduced two new features, *Favourites* and *Following*, to offer users



"New to you" - Discover more content on the YouTube homepage, Google

28

Twitter expands its crackdown on trolling and hate, BBC News, 16th February 2022

such as TikTok are deploying digital well-being prompts to remind younger users of tools to manage their experience, including setting screen-time limits.²⁹

Contact risks

Default settings should be configured to the highest levels of safety and privacy, and profile information, livestreams, videos and private messaging should not be accessible to anyone who is not already in contact with the child. Features that risk users, especially children, being exposed to people they do not know or do not wish to be contacted by should be switched off by default, and users should have the option of limiting who can see their information or initiate contact with them. For example, users should be empowered to decide to be visible to only those users who have had their identities verified.

Conduct risks

5Rights often hears from young people about the pressure they feel to be constantly visible, and to boost their popularity metrics (such as likes or followers). Features and functionalities can be redesigned to counter these effects. Switching off popularity metrics and ensuring that a diverse range of information and images are promoted to users can alleviate pressure to conform to certain image or cultural norms, or to show more of their lives and their bodies online.

Contract/commercial risks

A child's experience can be enhanced and commercial risks mitigated when games are playable without the use of features that incentivise or pressure players to spend money, and when paid-for randomised rewards such as loot boxes are not offered.

²⁹ Investing in our community's digital well-being, TikTok, 9th June 2022

Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact of harm from illegal content?	Confidential? – Y / N
Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	Confidential? – Y / N
Question 21: How do you mitigate any risks posed by the design of algorithms that support the function of your service (e.g. search engines, or social and content recommender systems), with reference to illegal content specifically?	Confidential? – Y / N
Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?	In the report 'But how do they know it is a child', 5Rights identifies ten distinct approaches to age assurance. These range from asking users to self-declare their age, to using hard identifiers to validate their age, to third party providers who can provide age tokens that validate ages or age ranges, to the use of biometrics and capacity testing to estimate age. As it stands, all approaches are undermined by the lack of common definitions, agreed standards and regulatory oversight, even though each can be done in a way the preserves the privacy of the user. The sector needs rules of the road not only to prove that their products are fit for purpose, but to prevent data-hungry services from discrediting the technology by demanding more data than necessary or failing to accurately establish age or age range. Even the most contentious forms age assurance, involving the user suffering any loss of privacy. This can only happen if the user consent to that specific use of their data, if that data is not shared with any third parties and if the result of the age assurance is not sold or shared to third parties. Age Tokens A promising area in the development of age assurance approaches is tokenised age checks. Age tokens offer a digital representation that

confirms a user is a specific age, above or below a specific age or within an age range. An age check provider verifies information, either provided by the user of from an official source, and generates a digital token. These may then be used in multiple settings as proof of age, without providing any additional data. The level of assurance a token provides depends on the initial method used by the provider who generates the age token. The systems often use hard ID 's such as passport or National Insurance cards to validate ages. The use of age tokens could be extended to allow institutions like GP surgeries and schools to issue tokens, as trusted sources that can easily verify people's ages.

Hard Identifiers

Even where the risk is high and therefore the age assurance system must be particularly rigorous, for example by requiring hard identifiers, this could be done in a way that ensures the user's privacy. Age assurance done using hard identifiers offer a high level of certainty that the user's age is being correctly verified, identity documents may show a user's date of birth 'within-record', such as a passport or birth certificate while others, such as a credit card, can act as a proxy, because you must be 18 or over in the UK to have a credit card. Hard identifiers are currently most commonly used by services that are restricted to users over 18.

Biometric Age Estimation

The potential for incorrect self-assurance, coupled with the problems with regard to the use of hard ID verification, has led to the rise of a new sector, where age verification is based on biometric data. Biometric age estimation tools use AI to assess the age of users. A facial image or voice clip that is collected once, then discarded, can establish an age range without verifying the unique identity of the child. The level of assurance a system provides will vary depending on the size and diversity of the data set on which the AI is trained and the standard of the technology. This will determine the error range that is provided. For instance, one system may give an estimated age of 10 with an error of 1 year either side, whereas another may give the same estimate but with a

	3–5-year error range either side. Both will be appropriate to use depending on the level of risk the service poses and the requirements of the check and the age of the user. What we have is an absence of governance ra- ther than an absence of technology, in which both third party providers and the services us- ing age assurance are not subject to agreed standards. Without agreed standards of effi- cacy and privacy it is up to each provider or ser- vice to set their own bar, which can result in poor practice.
	The impact and cost will depend on the implementation of the age assurance system within the particular service. For some the impact of using this technology may result in the service being able to keep adults out of their children/teen only spaces while for others it could result in the service being able to provide age appropriate messaging/functionality by default.
	The cost on the other hand will be dependent on the specific service used, how much age assurance checks will be able to be reused between services and the frequency with which the checks need to be carried out. Confidential? –N
Question 23: Can you identify factors which might indicate that a service is likely to attract child users?	There is a growing recognition of the fact that children need to be protected where they are and not where they are 'supposed to be'. This means that services that are not specifically aimed or targeted at children but are nonetheless likely to be used by under-18s, as well as those that are aimed or targeted at them fall under scope.
	The Age Appropriate Design Code uses the language of 'likely to be accessed' by children. The ICO states that they consider a service 'likely to be accessed' when there is evidence that 'the possibility of this happening needs to be more probable than not.' They further clarify that this 'recognises the intention of Parliament to cover services that children use in reality,' but not all services they could possibly access.

	The AADC clarifies the meaning of likely to be accessed by stating that it: 'If your service is not aimed at children but is not inappropriate for them to use either, then your focus should be on assessing how appealing your service will be to them. If the nature , content or presentation of your service makes you think that children will want to use it, then you should conform to the standards in this code . <i>If you have an existing service and children</i> <i>form a substantive and identifiable user</i> <i>group, the 'likely to be accessed by' definition</i> <i>will apply.</i> '
	The ICO outline two practical tests to establish whether a particular service is 'likely' to be accessed. <i>Firstly, services should consider the</i> <i>nature and content of the service and whether</i> <i>it 'appeals to children.' Secondly, considering</i> <i>the ease with which children can access the</i> <i>service and any measures 'put in place to</i> <i>prevent children gaining access.'</i>
Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?	Confidential? – Y / N
Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?	Confidential? – Y / N
Question 26: What information do you have about the age of your users?	Confidential? – Y / N
Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	Confidential? – Y / N
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	Confidential? – Y / N

Please complete this form in full and return to OS-CFE@ofcom.org.uk