Your response

Please refer to the sub-questions or prompts in the annex to our call for evidence.

Question	Your response
Question 1: Please provide a description introducing your organisation, service or interest in Online Safety.	*Is this response confidential? – N* TrustElevate is an age verification and parental consent service. We verify children's ages by checking against authoritative data sources and verify the relationship between that child and their parent/guardian from whom we then acquire (or not) consent for data processing, in line with GDPR Article 8. Our age verification process also allows relying parties to determine whether a child belongs to a particular age band (e.g. 5-9, rather than just >/<18) and deliver age-appropriate services in compliance with the Age Appropriate Design Code. TrustElevate was built to enhance safeguarding and protect children and their data online while empowering parents with the level of oversight they can expect in offline contexts. This is a short video showcasing TrustElevate: https://www.youtube.com/watch?v=DuPz-4OM-Yk Our business model relies upon the payment of relying parties like online platforms, including those providing social media and gaming services, and banks on a per managed child per annum basis. Our checks underpinning the age verification process are repeated at agreed intervals [each year] to ensure accuracy over time and to avoid exceeding data retention limits. The service is free to use for parents, children and schools. We have also developed a Child Rights Impact Assessment (CRIA) which we are developing with a number of partners, including auditors. A CRIA is a series of interrelated decision trees requiring engineers, data scientists, and commercial teams to consider risks, harms and safeguards associated with product features made accessible to children in specific age bands. A parallel instrument would be a Data Protection Impact Assessment (DPIA). We are expecting to roll-out in Q4 2022.
Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and	*Is this response confidential? – N* The four types of illegal content under EU law are a) child sexual abuse material (CSAM); b) racist and xenophobic hate speech; c) terrorist content; and d) content that violates Intellectual Property Rights. Individual member states may have identified additional categories that they have made illegal in their country. Of course, illegality does not mean that content of this nature has been stopped in its tracks:

search services? **IMPORTANT: Under this question, we are not seeking links to or copies/screenshots of content that is illegal to hold, such as child sexual abuse. Deliberately viewing such images may be a criminal offence and will be reported to the police.**	• On Instagram alone, 1.5 million items of content were actioned on Child Endangerment: Sexual Exploitation • On Facebook alone, 16.5 million items of content were actioned on Child Endangerment: Sexual Exploitation • On Facebook alone, 2.5 million items of content were actioned on Dangerous Organizations and Individuals: Organized Hate • On Facebook alone, 16.1 million items of content were actioned on Dangerous Organizations and Individuals: Terrorism • On Instagram alone, 481.3K items of content were actioned on Dangerous Organizations and Individuals: Organized Hate • On Instagram alone, 1.5 million items of content were actioned on Dangerous Organizations and Individuals: Terrorism The Internet Watch Foundation's 2021 report revealed that they assessed 361,062 reports in 2021, taking action against a record-breaking 252,000 URLs containing images or videos of children being raped and/or suffering sexual abuse. Variables relevant to the prevalence of illegal content on a user to user or search service include the type of content supported or made available by the service. At the highest level of this, the difference between Professionally Produced Content (PPC) and User Generated Content (UGC), due to the levels of oversight and accountability associated with each and their publishing parties, makes for an important variable. In the context of user-to-user services, this may not seem a particularly salient issue but one thing that might influence the quality of user generated content might be the dynamics of a service - if the platform serves UGC alongside PPC, on equal standing, it might be that the environment is a more professionalised one, one in which a curated image and brand is important to users. If advertising and branded content is something the service wants and seeks out, it may in turn feel greater incentive to 'clean up' its environment to make it more appealing to advertisers. The standing of ads on the service's interface may shape this dynamic, too: pop-ups, banners and sidebars may create a different user experience from one where branded content is part of a user's feed. This is not to suggest that the presence of advertising is necessarily a positive thing, only that it and the concerns of brands and advertisers are deeply connected to platforms' approaches to engaging users and delivering the content of interested parties. The next level down from this is the type of content that the service delivers: can users share videos, images, audio or text via the service? Or all of them, or perhaps some combination? Is there an element of pre-selection, as is the case with emojis or GIFs? Each type of content presents different types of risk, although text-based and audio content may be perceived to represent lesser risk of actual harm to those users reading/hearing the content. However, their power to communicate plans, actions, methods, etc. may be just as powerful in propagating the planning or recreation of crime and/or recidivism. There are also different types of text: hyperlinks pose a risk in terms of potentially connecting a user to another digital service

wherein illegal content is stored, albeit not on the original service where the text-based communication took place.

Images and videos can present graphic and upsetting material to users, an event that can itself be traumatising. We have also borne witness, globally, to the radicalisation of individuals and communities by the profit incentive to drive user engagement by showing users increasingly extreme/graphic media. This relates to a process, distinct from, though still relevant to, the discussion of format here.

Comparisons of harms based on format are, of course, difficult or even impossible to do. Different services support different content formats on their platforms, users may use one formats more than others within an individual service, and different metrics like scale, use, reporting use and accuracy muddy the picture. The Alan Turing Institute has determined that the proportion of hosted content which is actioned for being hateful or harassing amounts to:

1. 0.001% of content posted on Facebook.
2. 0.001% of videos posted on YouTube.
3. 0.0001% of content posted on Reddit.
4. 0.2–0.3% of users on Twitter

Note that these are not broken down by type of content - but that will be a factor. As will the shareability of that content.

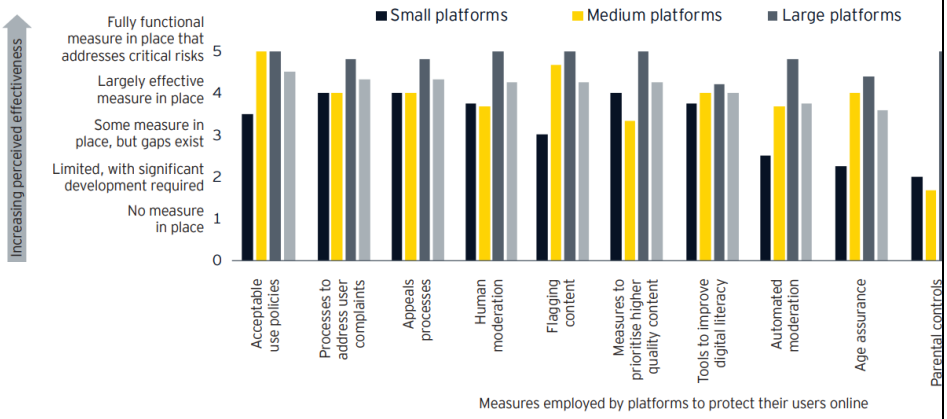
Shareability was notably raised as an important issue in the legislative scrutiny of the Online Safety Bill in the House of Commons. Frances Haugen testified as to its significant role in boosting the appearance and prominence of hate speech in users' feeds and in emboldening other users to 'pile on' in online conflicts, leading, in many cases, to online bullying, harassment and even hate speech. This is also tied in with the visibility or connectability of users with other users. Some services may not enable public profiles, or non-public profiles may be the default (public meaning accessible/visible to everyone, not just other users/accounts the profile has connected with). As noted in the Government response to the Joint Committee report on the draft Online Safety Bill, there are "risks created by virality and the frictionless sharing of content at scale", which can be "mitigated by measures to create friction, slow down sharing whilst viral content is moderated, require active moderation in groups over a certain size, limit the number of times content can be shared on a "one click" basis, especially on encrypted platforms, [and] have in place special arrangements during periods of heightened risk (such as elections, major sporting events or terrorist attacks)". While these measures are not in place, it can only be expected that illegal content is being shared to the detriment of users.

The recommendation of moderation specifically targeted at viral content is a significant one and highlights the role of moderation in combatting the proliferation of illegal content. The use or otherwise of pre-moderation by a service is a major variable in the presence of illegal content on a service.

Pre-moderation entails that a user's content is subject to review by a moderator or moderation system before it is visible to other users. Some services do this in limited cases; for example, a service may pre-moderate a user's first five posts, since it has been found that users most commonly act in violation of the service's Community Guidelines or Terms of Service (or the law) early on. Or they may pre-moderate the posts of a user who has been reported multiple times, or suspended several times, such that it may be appropriate to terminate their account. In each case, the role of pre-moderation will be an important variable in terms of the presence of illegal content on the service. As are the types of reactive moderation a user to user service employs to combat the presence of illegal activity.

Reactive moderation most commonly involves automated detection of potentially violative content, human review of content reported by users or flagged by automated systems. Triage is an important feature that not all services employ. The below graph, from an EY report, reveals the perceived effectiveness of user protection measures, including human moderation and content flagging, broken down by the scale of the service.

Platforms' perceived effectiveness of measures to protect users online by platform size



Note here the inclusion of age verification and age assurance - while it isn't possible to say that enforcing an age gate will reduce the quantity of illegal content on a user to user service, it will minimise the exposure of children and young people to illegal content on that service.

We can see, then, that although this graph is the outcome of a self-report survey, scale is an important variable. Smaller platforms have more limited resources to combat the illegal content they do host. Some may actively be designed and built to support the sharing of illegal content and are sufficiently small to fly under the radar.

The functionality of services, as well as their scale, is also a major factor: the duration of a post ('stories' last 24 hours, Snapchat 'snaps' can last as little as 1 second), for example, will affect the quantity of illegal content on the service. Encryption, too, plays a role here - although exactly what role is a contested issue. Some people regard it as a stalwart in the fight for user privacy, others a trojan horse for illegal activity. TrustElevate's position is that the appropriateness of encryption relates to context and that there are always different balances to be struck between safety and privacy depending on the situation: there is no one size fits all approach. For example, where children are communicating and in particular where children are able to communicate with people of other age groups, the application

	of encryption would be a misstep and would represent a lapse in the exercising of a service's duty of care to its vulnerable users. In other contexts, where children are not present, it may be more appropriate to implement encryption for privacy purposes.
Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content presented by your service?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 4: What are your governance, accountability and decision-making structures for user and platform safety?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 5: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?	*Is this response confidential? – N* While this might have been a worthwhile question at one point in time, when considered against the backdrop of companies such as Meta force feeding users cookies and disbanding their Responsible Innovation team, Google firing their Ethical AI team and Patreon cutting their cybersecurity team, it is clear that regulators need to tackle the core issue of profits being more important than a duty of care towards users. What these companies have demonstrated is the following: while they recognise the harm that their technology can facilitate, they no longer want to hold up a mirror to themselves or invest in, recognise, or support efforts to develop responsible and ethical technology that keeps users secure. Given such a massive dereliction of duty, it is not clear what value there is in examining the positioning of ToS and public policy statements. Except, of course, that companies engage in these discussions and then 'demonstrate their commitment to safety' by making a few changes that ultimately amount to window dressing, but which policy teams can point to illustrate willingness to protect users. These minor tweaks could be made - but as outlined in the opening paragraph this is not a core issue. Terms of service and public policy statements should be made available on or via the homepage of a platform or company's site or app. Where they are grouped and stored on a separate page from the homepage, this separate page should be no more than one click

away, with users being able to open the individual policy documents with no more than one additional click from there.

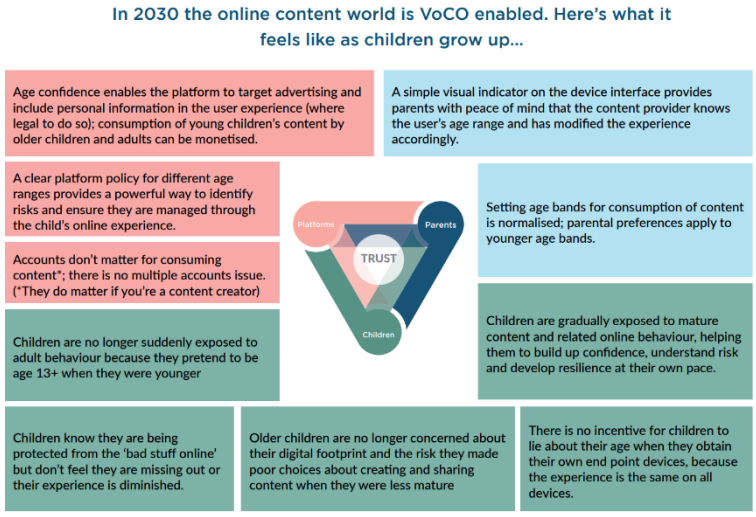
Accessibility should include age-appropriateness and the salient points of such documents should be communicated to the youngest users of the platform by way of images or video where those users might not reasonably be able to read. To ensure a standard approach to age-appropriateness, service providers should look to the age-bands (0-5, 6-9, etc.) of the Age Appropriate Design Code and its recommendations on how best to convey information to users in those age bands.

There is great scope for standardisation in terms of the language used in the instructions provided and their underpinning concepts. The provisions of the platform's terms of service should be based on objective, verifiable criteria. Additionally, the tallying of these instructions and guidance around prohibited content, contact, conduct and commercial activity with accountability and transparency metrics could present a benefit to users. In looking to provide users (and their parents) with meaningful and representative information about their platform's policy, principles and processes around child rights due diligence, platforms should conduct a Child Rights Impact Assessment (CRIA), as called for in the General Comment 25 on the United Nations Convention on the Rights of the Child. CRIAs are designed to calibrate functionalities with risk and mitigation strategies in direct relation to the age-bands of users. It enables companies to determine how best to mitigate harm while furnishing them with a score that is intended to be useful to users and their parents in assessing whether they want to use the platform - a parallel might be found in colour-coded nutritional labelling. The use of a CRIA score in conjunction with terms of service and public policy documents would supplement users' understanding and summarise their contents in a readily comprehensible manner.

Platforms often reply on the following rationale that it is important for platforms not to over-disclose their detection and moderation processes so as to avoid equipping bad actors with the tools to evade their systems. While this may be true in relation to users at the most granular level, it is vital that this rationale does not apply to disclosure to the regulator or in communicating to users what they can expect of the service provider in providing a safe environment. Regulators routinely work with banks and gambling operators to assess the efficacy of the measures put in place to combat, for example, fraud and money laundering. This regulatory oversight includes regular reviews of the types of fraud detection and security tools these companies deploy. Furthermore, information about the range of anti-fraud services is readily available online.

Social media and gaming platforms should be subject to transparency requirements enforced by the regulator so that there is a shared knowledge and understanding of the threat detection measures a company operates. Users should be able to report to the regulator when their complaints have not been handled to their satisfaction in line with the requirements stipulated in the Telecommunications Act. In addition, presenting clear instructions and outcomes fosters a sense of mutual accountability between

	users and platform and can be important in encouraging positive user behaviour. In the current absence of UK regulatory oversight, online service providers could align themselves with current best practices from other industries around action and notice procedures. Users should be notified of any detected and actioned violations of the terms of service or community guidelines and be plainly provided with a pathway for appeals. All of this information should be provided by the platform in a timely manner and using clear, age-appropriate language. Appeals should be easily conducted and oversight must be levied over platforms' responses to them and responsibilities in remedying any issues raised by users through the reporting and appeals processes. Platforms should share publicly information about how reporting and appeals processes have flagged issues with existing practices internally and led to any changes. Users and non-users can and do influence platforms' terms of service and public policy. This should be cemented in platforms' governance frameworks and communicated to users to maintain the sense of mutual accountability and transparency. There is a precedent for establishing what should be included in Terms of Service and other public documents. In February 2009, the European Commission, in partnership with 'social networking sites' (SNSs), drafted the Safer Social Networking Principles for the EU. However, they were created in the era of self-regulation, such that platforms were not compelled to comply or align themselves with such principles, despite the fact that they had a hand in writing them. We need to move beyond just principles and implement binding rules to which platforms must adhere with consistent oversight by the regulator. With rules in place for users and platforms, it should be the case that those rules are equally communicated: the guidelines and minimum standards set for user behaviour should be represented to users, as should the guidelines and minimum standards that platforms must meet at risk of penalty. This information about platforms should also include information about appropriate pathways for the reporting of platforms should they be seen to be not fulfilling their duties or meeting requirements. In sharing this information transparently, we can expect there to be greater incentives toward consumer protection and consequences for not protecting end-users. The

	same report laid out more benefits in the diagrams below: 
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?	*Is this response confidential? – Y / N (delete as appropriate)*
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?	*Is this response confidential? – N* Reporting mechanisms should include the minimum number of clicks and steps for a user to quickly submit a report or complaint with ease while equipping the receiving party/platform with sufficient information to assess the report and determine the appropriate response. This means that the reporting mechanism should be accessible within one click of where an offence or violation of the T&Cs or Community Guidelines could occur. For example, in a dating application, there should be a reporting pathway available to users on other users' profiles and within any chat functionality or, in an online game, users should be able to report another user during any interaction regardless of the interaction being chat, voice or even video based. The reporting mechanism should be visible at these points and made plain to users by using either an easily recognisable symbol, such as a flag, or words like 'Report', 'Make a complaint', etc. This

should be the case across registered and non-registered users: where interaction may occur or where a person can view another person's post, these principles should stand.

It should be possible, where appropriate, for a user or non-user to identify the item of content or incident in the report - this might mean that individual posts have a report button connected with them or they should be able to identify the item in the reporting pathway by way of selection (if there is a limited number of posts that they could be choosing from) or free text input box. The person should be able to select from a list of reasons why they are reporting the content. In creating reporting mechanisms and pathways, it is important to take into consideration the functionalities of the platform and the means by which a user would be violating Community Guidelines or Terms and Conditions. In considering those means, the platform should be able to more accurately determine the ways in which a user may have violated. 'Other' should be included as a possible reason, in the event that the reasons provided have not captured the offence, followed by a free text input box. Where a user or non-user is known to be a child, either by way of registration or a verification of age further upstream, the reporting process should be differentiated by providing prompts in age-appropriate language, as described in the AADC.

The flow, typically, should be Report > What are you reporting? > For what reason? > Further information. This may, of course, vary from platform to platform. If it is the case that a user has selected a reason that indicates a serious violation/harm has occurred or is is expect to, or an automated detection system has picked up on keywords in the free text input box that would indicate something to the same effect, the platform must provide signposting for additional support during the reporting process or immediately after. Platforms should work with mental health organisations so that those organisations delivering services online and their abuse management systems can direct users to support. See the RAMP Guide: Risk awareness and management guide delivering mental health services online guide. Children need support in a timely manner and these organisations working together could increase the capacity to ensure people get that support designed to mitigate harms

Where non-users are concerned, contact information should be required from the user so that additional information may be obtained if necessary and/or notice of the outcome may be provided to the non-user. This contact information should likely include name and email address and/or telephone number. Where non-users are able to see and report content, it could be the case that a child is reporting an item of content. It should be possible for a child to indicate that they are a child as part of the reporting process and indicate that they will, in lieu of their own, provide their parent's contact information.

The UK Children's Commissioner report in 2019 highlighted the impact a lack of transparency or non-responsiveness following a report can have on children. A primary school child was reported as having said that "[i]t makes you feel like they don't care if they don't respond". Another told them that "I think they should make it more simple so that you can call up for reports

	instead of having to type", underscoring the significance of platforms providing alternative methods for reporting, which should be a requirement of them moving forward.
Question 8: If your service has *reporting or flagging* mechanisms in place for illegal content, or users who post illegal content, how are these processes designed and maintained?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 9: If your service has a *complaints* mechanism in place, how are these processes designed and maintained?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 10: What action does your service take in response to *reports* or *complaints*?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 11: Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?	*Is this response confidential? – N* Ultimately, it is the case that the pressure would be taken off reactive moderation procedures if preventative measures were taken to ensure the safety of users in the first place. One such preventative measure is ensuring that the platform is age gated and delivering age appropriate services to its users. To do so, a platform must ensure that content, where possible, is age rated using a balance of automated systems, human review, and self-classification. Age classifications of media content vary from country to country. Moreover, many age classification labels cannot be processed automatically by computers. The EU-co-funded pilot project MIRACLE intends to digitalise all age classification labels so that all labels speak the same language. In the US, Common Sense Media is working on a similar initiative. The knowledge and expertise exists to address a number of these

	knotty problems, but what is required is the willingness of companies or sufficiently strong regulatory oversight. While it is the case that the implementation of age rating for user generated content may present a challenge, it is not the case that we cannot hold higher standards for the content to which our children are exposed than for general audiences, wherein adults may make decisions according to their own risk appetite. That is not to suggest that harmful content should proliferate on platforms for adults, only that we must set a higher bar for our children. The question of age rating content also raises the issue of who is responsible for doing so. Dealing with the quantities we see online, it would not be reasonable to expect a single classification body to age rate all content online. Depending on the age bands that have access to the content and the governance of the framework of the platform, there may be a significant role here for moderator review, trusted reviewers (community moderators in the style of trusted flaggers, e.g.) as well as initial self-classification. There is a role for self-classification of content and it is already incorporated into the upload process on YouTube, whereby the platform asks the user whether the content being uploaded is child-friendly. There is the issue, then, of being sure that users understand what constitutes 'child friendly' both broadly and within the context of the individual platform, perhaps, which calls to mind Q7 as well as the sense of personal responsibility that user is taking on in self-classifying their content. Does anonymity have a role to play here? Discussions around the Online Safety Bill have raised the question of whether anonymity can continue online or whether users should have to verify their identity and connect their account with that offline identity in order to post or, at the very least, connect with someone who has connected their offline and online identities in order to prevent abuse, harassment and pile-ons on social media. The condition of identifying yourself in relation to an item of content may discourage bad actors from flouting the self-classification system or posting potentially harmful content more generally. Marking oneself as the source of a negative item of content or tying oneself to a piece of content intended to harm children could disincentivise bad actors and flag them as ones to be mindful of for platforms. It is worth noting that monetisation of user generated content entails that an uploading user must connect their account with their bank details (associated with their real life/offline identity) in order to capitalise on their content. It may be valuable for those platforms with this functionality/capability to learn from other sectors around combating bad actors and bots, etc. through connected identities/channels. (Below response was kept confidential)

Question 12: What automated moderation systems do you have in place around illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 13: How do you use human moderators to identify and assess illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 14: How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services?	*Is this response confidential? – N* The current age-gating processes in place are weak and not fit for purpose. One of the more commonly used methods of age checking online is the use of the Email Plus method, introduced by the US Children's Online Privacy Protection Act, whereby a user who might be a child (based on a self-assertion of age or the service's target audience) is invited to share their parent's email address. An email is then sent to the parent who agrees to allow the child to access the service. The email plus mechanism has not met either user's or service providers' needs because children are known to often provide their own email address or one to which they have access and a large proportion of those emails sent are never actually opened because they are redirected to spam folders. Another commonly deployed method is that of credit card checks. Per the Pas 1296 Age checking code of practice, "[s]ome merchants regard holding a credit card as a proxy for indicating a cardholder's age (i.e. only over-18s have credit cards). However, entry of credit card details only confirms that the customer has access to, or knowledge of, credit card details. It is not possible to infer the person to whom that credit card was issued is the person entering those details." And, more recently, platforms have implemented systems whereby a cookie is dropped when a child enters a Date of Birth below 13 such that if they subsequently try to re-enter a different, earlier date of birth to appear older and gain access they will be blocked from doing so. However, children have figured out a workaround by clearing their cookie caches (that is, if they failed to lie about their date of birth in the first instance). Plus, age estimation by way of biometric or behavioural analysis processes inordinate amounts of children's sensitive data without checking with their parent(s). The use of estimation also raises the issue of generating and

introducing to the ecosystem synthetic data points. The insertion of synthetic data points (age estimates) into the data ecosystem is not helpful from a child rights perspective: fallible age estimation associating individuals with incorrect age bands could exclude them, may be upsetting and will be disruptive to their ability to exercise their data rights. Using hard identifiers like credit cards or other kinds of official documentation for age verification (not just checks for parental responsibility as described above) also routinely exclude people without access to such identifiers and any use of this system must provide alternatives.

What is becoming increasingly clear is that there is an ecosystem-wide need - not for 'checks' or estimations but for **verification** of age. When differentiating service provision in line with AADC age bands, a level of accuracy is required that the above methods cannot provide. And when acquiring parental consent on behalf of the youngest children in line with the GDPR Article 8, it is critical that that consent is verifiable and not simply provided by the child via another email account they've created or to which they have access.

The debate around whether age verification can and/or should be implemented in the digital ecosystem has been around for a long time. As the technical feasibility of the proposition has become plain, the appropriateness of it has been contested by some camps. Some have posited that the widespread adoption of age verification would be restrictive to children's online experiences, excluding them from digital life and limiting their agency.

However, this position is problematic: its starting position is negative and runs contrary to a growing body of child-centric research which has centred children's desire for an internet designed for them, an internet in which they can talk to their friends and people their own age without having to worry about 'weirdos', for example. It has historically been framed as a restriction but rather represents an opportunity for children and for companies to better deliver their services to users. Just as in sports or even school classes, where children are grouped according to the age band to which they belong, it should be the case that children are provided with services and opportunities in line with their age group.

Youth Activism is a growing feature that will shape how companies respond – here are some examples:

- The #DesignItForUs https://www.designitforus.org campaign led by @logoffmovement https://www.logoffmovement.org
- and @technicallypoli (https://www.youtube.com/playlist?list=PLkaxTNca-z8jo3N9CkL2G7Iv0BFHXfR0V) demonstrates youth support for the California Kids Code

The incorporation of youth voices in the dialogue around the enforcement of standards and guidelines is critical in advancing the protection of online communities, especially those including children and young people. The

	Social Switch's video details the outputs of a programme of research involving children to determine what harm means to children and young people. Taking a child-centric approach, as in the Social Switch's programme of research, is critical to the redesign of the internet to be a place not only for adults but also designed with children and young people in mind. The video details children and young people's experiences around reporting and removal and find these mechanisms wanting: children and young people report a lack of transparency, consistency and responsiveness that leads to feelings of confusion, frustration and upset. The children and young people in the video want clearer rules around what is acceptable and what is not online, greater transparency on boundaries, options to determine what it is that they see online inclusion in the process of improving online spaces. Platforms should align their products and processes with these expectations and must incorporate children and young people's voices in doing so. Concerns around privacy, too, can be mitigated in relation to the standards that underpin identity attribute checking and the numerous provisions of the Data Protection Act that require companies to respect child rights in relation to not being profiled, restrictions around the processing of sensitive personal data, etc. Further safeguards with respect to users' data can be found in the implementation of zero data, zero knowledge models and the tokenisation of user data such that their personal information is not being insecurely transferred from one party to another.
Question 15: In what instances is illegal content removed from your service?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 17: What other sanctions or disincentives do you employ against users who post illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*

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?	*Is this response confidential? – N* Central to the issue of mitigating harm is the question of what is harmful to who. Harm is not readily quantifiable in terms of how much harm was done to an individual user or the widespread impact on groups who may be affected by the spread of a type of content. However, we do understand that children and young people self-report having more harmful experiences than adults and seeing more harmful content, including but not limited to harmful misinformation. We also know that children have poorer coping strategies when dealing with harm by virtue of their age and the stage of cognitive development they're at, which may impede them for being able to interpret or verbalise what they have experienced or to have a concept of the harm in the short term or, indeed, the long-term effects of what they have experienced. That is why they are afforded special protections in law. The fact is that the functionalities and design features deployed to effectively prevent harm must be deployed along the lines of age bands. Harm impacts children and young people differently from adults and may be facilitated in its delivery by different means. While there is overlap in the platforms children and adults use, there are also swathes of platforms that are far more popular with one group over the other, and that is also true of age bands within those two groups. Across these various platforms and contexts, any number of design features may be deployed to effectively prevent harm. But the fact is that they will vary widely according to the current functionalities, community standards and norms and, ultimately, the audience of those platforms. Each platform must implement Safety by Design principles and incorporate trust and safety training in their design and product teams' processes to ensure the best outcomes in mitigating harm in their digital environments. They must also, in doing so, ensure that their current functionalities and standards align with their audience. Effective age gating must be put in place to ensure that harm mitigation strategies are appropriately targeted. Age verification will ensure that platforms know the ages of their users and deliver their services accordingly. Laws and regulations state that, where a service is being delivered to a child, the highest safety and privacy standards must be applied - for example, profiles set to private by default, etc. How can these highest standards, known to be effective in mitigating harm, be applied in environments where users' ages are not checked? On a rollercoaster ride, safety checks must be conducted and standards must be met to ensure that the ride can continue to operate safely. This means that *anyone could* use the ride. It ensures that things will run smoothly and the ride itself is functioning properly. However, there are further standards that people must meet to ensure that *they* will be safe on the ride. Children must be of a certain height to ride - this isn't inappropriately inhibiting their access to an experience. This isn't restricting their

	agency or minimising their need to develop risk management skills, it's taking the responsibility to ensure that they are not taking unnecessary risks on an experience designed for adults. These same principles apply to digital experiences on platforms. It may be that one platform is the theme park and one functionality is one ride (e.g., children over 13 may be on TikTok, but have to be 16+ to livestream) or that the broader digital ecosystem is the theme park and one platform is one ride (e.g. to shop on an adult website, you must be over 18) but the point remains the same. It is also worth noting here that children are known to be early adopters of new technologies; imagine a world where the first safety checks of a new rollercoaster ride were performed mostly with children, putting them on the first line of risk and harm. Age verification is the critical tool in platforms' toolkits to ensure that children are protected and their rights and upheld by being delivered with services designed with them in mind. Safety by Design cannot be universal: child Safety by Design is necessarily different from that which would be appropriate for adults. As noted in 5Rights' response to the Online Harms White Paper, "Nearly one billion children are growing up in an environment that **systematically fails to recognise their age**, and in doing so, fails to uphold the protections, privileges, legal frameworks and rights that together constitute the concept of childhood." Protecting children encompasses different matrices of rights and protections than those involved in protecting adults and we must all do our best to ensure they are effectively upheld using sometimes overlapping but also sometimes distinct means. In ensuring we are doing our best in doing so means implementing the best practices available to us. Age *verification* as opposed to age assurance is the best means of checking users' ages available. That said, not all age verification methods are created equal. Using trusted third parties is an important step in ensuring that platforms are not checking their own homework, incentivised as they are to allow more users per month, for example. Data minimisation principles should be upheld, ensuring that only the minimum amount of data to conduct the verification of age and to acquire verifiable parental consent (per GDPR Article 8) is processed. TrustElevate's zero trust, zero data model upholds children's data protection rights while effectively checking their ages against existing authoritative data sources, meaning no new data needs to be gathered or synthetically generated.
Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact	*Is this response confidential? – Y / N (delete as appropriate)*

of harm from illegal content?	
Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
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?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?	*Is this response confidential? – N* Age assurance technologies range from those offering an informed estimation of age to financial industry Know Your Customer-level verification of age. The difference between the two cannot be overstated and nor can the potential impact on the users' whose ages are being checked or on the data ecosystem. Age estimation offers a solution to the problem of determining a user or individual's age based on visible or perceptible attributes without authentication via comparison with previously verified attributes. This solution assumes that people age (and present that ageing) in a predictable manner across age bands and regional and socioeconomic differences, to name only a couple of major variable factors, and that those users or individuals cannot influence their presentation to circumvent age estimation measures to produce an estimate that suits their requirements. As noted by Huerta et al., biometrics or image-based age estimation presents such a challenge because of the uncontrollable nature of ageing (and the role of environment

	in determining the ways in which people age, individuality of traits and features, masking of faces by hair, facial hair, glasses and makeup) and the challenges posed by the need to gather sufficient and appropriate training data. Collecting and correctly utilising training data has historically been a challenge to industry and regulators alike. Indeed, one of the recommendations from the UK-government funded Verification of Children Online (VoCO) project was "exploring accessibility to testing data, to improve accuracy in age assurance methods". (Below response was kept confidential)
Question 23: Can you identify factors which might indicate that a service is likely to attract child users?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 26: What information do you have about the age of your users?	*Is this response confidential? – Y / N (delete as appropriate)*

Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	*Is this response confidential? - N* A critical measure in developing a digital ecosystem in which children and young people are protected from harm and in which their rights are upheld is the development and implementation of a Child Rights Impact Assessment (CRIA) A CRIA is a series of interrelated decision trees (currently laid out in .xls) requiring engineers, data scientists, and commercial teams to consider risks, harms and safeguards associated with product features made accessible to children in specific age bands. The ICO's Age Appropriate Design Code, IEEE's Standard for an Age Appropriate Digital Services Framework and General Comment 25, adopted by the United Nations Convention on the Rights of the Child, had highlighted the need for a CRIA. The development and deployment of a CRIA is critical to ensuring digital platforms are both transparent and accountable concerning their prioritisation of the protection and wellbeing of children and young people. The CRIA enables product developers/service designers to see the risks associated with different features in relation to different age-bands. In exposing interrelationships between child rights, product features sets and associated harms impacting children, it will be possible to mitigate risks to children's safety, including sexual predators, at the design stage of products and features. (Below response was kept confidential)