making communications work
for everyone

Your response

Question 1: To assist us in categorising responses, please provide a description of your organisation, service or interest in protection of children online.

Is this a confidential response? (select as appropriate)

No

Nexus is Northern Ireland's leading specialist sexual trauma counselling and early intervention education charity. Our early intervention and prevention services present a host of training and information sessions to schools, clubs, youth programmes, etc. to young people to talk about relationship and sexuality education as well as how to stay safe online, particularly concerning engaging in content of a sexual nature.

Question 2: Can you identify factors which might indicate that a service is likely to attract child users?

Is this a confidential response? (select as appropriate)

No

According to the Office for National Statistics England and Wales[1], in the year ending March 2020, the most common online activities for children aged 10-15yrs were: watching videos online; messaging; playing online games; and social networking sites.

For social media and online video/music content, Ofcom's Media Use and Attitudes 2022[2] research found that YouTube, WhatsApp, TikTok, Snapchat, and Instagram were the most popular platforms for children and young people aged 3-17 yrs. From this data, we can conclude that children and young people are most frequently using online spaces that have the ability to view and share video and other media content. These platforms allow for users to create profiles to follow their favourite creators and to house their favourite content in one place, as well as share and message with other profiles. From our own work with young people in schools, the majority are using smartphones to communicate and consume online content, and the majority are using apps such as Snapchat, Instagram, and TikTok to do so.

[1] https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/childrensonlinebehaviourinenglandandwales/yearendingmarch2020

[2] https://www.ofcom.org.uk/__data/assets/pdf_file/0024/234609/childrens-media-use-and-attitudes-report-2022.pdf

Question 3: What information do services have about the age of users on different platforms (including children)?

Is this a confidential response? (select as appropriate)

No

For many online services, a profile is required to view the platform. To sign up for platform, a user must input basic data, such as a username, password, email, name, and a birth date. Of course, it is entirely possible to input a false birthdate, which raises concerns for the safety of young children on social media platforms.

For other platforms that cater to adults- including shopping channels, online stores, and adult content- a credit card or bank account information is required in order to make purchases.

Question 4: How can services ensure that children cannot access a service, or a part of it?

Is this a confidential response? (select as appropriate)

No

Some platforms such as Facebook, are implementing new strategies to tackle underage users, such as manual and AI screening to check profiles for indicators of age, such as birthday posts or milestone events[3].

And as stated above in Question 3, some platforms and sites have a built-in paywall that is only accessible with an account that has a credit card or online money service attached, which requires a minimum age to secure, such as PayPal[4] and Apple Pay[5]. For other services, they may require Government Identity Documents, Mobile Phone Account Records, and/or Social Proofing[6].

[3] https://about.fb.com/news/2021/07/age-verification/

[4] https://www.paypal.com/uk/webapps/mpp/ua/useragreement-full

[5] https://support.apple.com/en-gb/HT204506#:~:text=If%20you're%20under%20the,card%20in%20the%20Wallet%20app.

[6]https://avpassociation.com/avmethods/#:~:text=Users%20will%20typically%20submit%20an,the%20form%20of%20ID%20used.

Question 5: What age assurance and age verification or related technologies are currently available to platforms to protect children from harmful content, and what is the impact and cost of using them?

Is this a confidential response? (select as appropriate)

No

According to the British Standards Institution's Provision and Use of Online Age Check services (PAS 1296:2018, found on AVP Association's website)[7], there are a few different methods of age verification technologies, including manual human checks and AI recognition software:

- Government Identity Documents: passports, driving license, national identity card.
- Mobile Phone Account Records: mobile phone companies apply parental controls to new phones and sim cards which can only be removed by providing your age.
- Credit Card
- Biometric Age Estimation: using a trained AI software that analyses thousands of anonymous images of people with a known age, which can find identifying features to estimate age.
- Social Proofing: Using AI, it is possible to estimate the age of a user based on their online behaviour, such as their interests, their bio data such as school, sports team, etc., their friends lists, etc.

The use of Government Documents is not without controversy; requiring people to upload their personal ID information would exclude those who do not have these documents, or perhaps do not have access to the documents due to a myriad of reasons, such as not being able to afford a physical copy or are not able to access a copy of their documents due to them being withheld by a parent or partner.

AI technology is continuing to evolve; however, it could be limited by data protection laws, including the use of personal data, as well as the limit of available information pertaining to individual platforms to make an accurate determination of age.

Question 6: Can you provide any evidence relating to the presence of content that is harmful to children on user-to-user and search services?

Is this a confidential response? (select as appropriate)

No

[7] https://avpassociation.com/avmethods/#:~:text=Users%20will%20typically%20submit%20an,the%20form%20of%20ID%20used.

Question 6: Can you provide any evidence relating to the presence of content that is harmful to children on user-to-user and search services?

The Internet Watchdog Foundation released data for January to July 2022 that found 20,000 webpages of child sexual abuse imagery included 7-10 year olds, a 360% increase in primary priority content for this age group[8].

Question 7: Can you provide any evidence relating to the impact on children from accessing content that is harmful to them?

Is this a confidential response? (select as appropriate)

No

From our work, we report that children and young people are impacted by harmful content in the following ways:

- Unwanted attention, both online and offline by their peers and anonymous profiles online
- Bullying and harassment
- Blackmail, or the threat of reporting to a guardian, carer, etc.
- Mental health deterioration, including extreme stress, anxiety, and depression
- Risk of self-harm
- Shame and embarrassment
- Ostracization from peers

Question 8: How do services currently assess the risk of harm to children in the UK from content that is harmful to them?

Is this a confidential response? (select as appropriate)

No

The NSPCC[9] currently advises a full risk assessment of any new platform before sharing it with young people, including:

- Consider using a site that is specifically designed for use with children, such as an education platform.
- Are there adverts on the platform that can redirect young people to another site?
- Are there sharing tools available, and who has access to them? i.e., could a child share posts with other profiles, and vice versa?
- Who will have access to the platform?

[8] https://www.iwf.org.uk/news-media/news/20-000-reports-of-coerced-self-generated-sexual-abuse-imagery-seen-in-first-half-of-2022-show-7-to-10-year-olds/

[9] https://learning.nspcc.org.uk/news/2023/january/risk-assessing-online-platforms

Question 8: How do services currently assess the risk of harm to children in the UK from content that is harmful to them?

- What communication features are available, and are they available for the general public? i.e., the ability to send messages to profiles that are not 'friends'
- Does the platform allow young people to contact each other, staff, or the general public, in a group and/or a one-to-one basis?
- Will the platform be accessible outside of working hours?

The Information Commissioner's Office created a Children's Code Risk Assessment Toolkit[10] to help digital services conduct their own risk assessment in compliance with UKGDPR and UNCRC, including:

- Accountability for data protection and children's privacy, including low privacy default settings being applied to all accounts regardless of age.
- Identify and provide necessary support, protection, or development to young users.
- Establish whether the service is likely to be used or accessed by a child when it has not been designed for children specifically.
- Verification processes for determining the age of users.
- Implement appropriate privacy measures that are communicated to users in a way that they can understand.
- Upholding standards of use for users of the services
- Affects on children's mental and/or physical wellbeing by data driven incentivisation of continued or additional use of the service.
- Physical tracking of the location of a child resulting in the data being misused and compromising the physical safety of the child
- Profiling for targeted marketing and additional content delivery resulting in advertised behaviour that can be inappropriate and damaging to children.
- Awareness of children and/or guardians of the delivery of the service and its connectivity to other services

Question 9: What are the exacerbating risk factors services do or should consider which may have an impact on the risk of harm to children in the UK?

Is this a confidential response? (select as appropriate)

No

[10] https://ico.org.uk/media/for-organisations/documents/4020178/childrens-code-self-assessment-risk-tool.xlsx

Question 9: What are the exacerbating risk factors services do or should consider which may have an impact on the risk of harm to children in the UK?

Services which allow the following could be exacerbating the risk factors of harmful content for children:

- Sharing and posting images, videos, and audio content
- Anonymous profile or Avatar profile capabilities
- Anonymous chat service without a proper screening process (i.e., the ability to message someone without an ‘accept’ option, a warning that this person is someone you don’t follow, etc.)

Question 10: What are the governance, accountability and decision-making structures for child user and platform safety?

Is this a confidential response? (select as appropriate)

No

Staff, carers, people in a position of trust over children, etc. should all receive cybersafety and online child exploitation training, with a particular focus on grooming, language, laws, and consent.

Question 11: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements for children (including children of different ages)?

Is this a confidential response? (select as appropriate)

No

Having a terms of reference and user policy that is accessible and streamlined is key to user engagement and understanding. This includes simplicity in language and very clear reasoning for the policy, i.e., being clear on why underage users are not allowed to use the platform and the dangers of accessing the service underage.

User agreements tend to be very long documents that are easily skipped, even if it is required to scroll to the bottom of the policy before being able to continue with the account/website. In order to enhance clarity on terms of use, service providers need to simplify their content and make it user-friendly, just as they do for other tutorial sections of their platforms, such as how to create a profile, how to create a post, etc.

Question 12: How do terms of service or public policy statements treat 'primary priority' and 'priority' harmful content?[11]

Is this a confidential response? (select as appropriate)

No

YouTube doesn't allow content that endangers the emotional and physical wellbeing of minors, including primary priority and priority harmful content such as[12]:

- Sexualisation of minors
- Harmful or dangerous acts involving minors, including encouraging minors to do dangerous activities that may lead to injury
- Infliction of emotional distress on minors, including exposing minors to mature themes, simulating parental abuse, coercing minors, or violence
- Misleading family content, where the content is targeted to minors but contains sexual themes, drugs, violence, or death
- Cyberbullying and harassment involving minors.
- YouTube may add an age restriction to content that includes harmful or dangerous acts that minors could imitate, adult themes in family content, and vulgar language.
- This type of harmful content also applies to videos, video descriptions, comments, Stories, Community posts, live streams, playlists, and any other YouTube product or feature, as well as external links in user content.

YouTube has a separate suicide and self-harm policy[13], which doesn't allow content that:

- Gives instructions on how to die by suicide or engage in self-harm.
- Show graphic images of self-harm.
- Share content related to suicide or self-harm that is targeted at minors.
- Blurred imagery in combination with details or visuals that show the method of suicide.
- Suicide notes without sufficient context
- YouTube may add an age restriction, warning, or Crisis Resource Panel on videos and content that: is meant to be educational, documentary, scientific, or artistic; content that is of public interest; dramatizations or scripted content including animations, music videos, or clips from films, etc.

YouTube also has a nudity and sexual content policy. "Explicit content meant to be sexually gratifying" is not allowed on YouTube, and posting pornography may result in

[11] See A1.2 to A1.3 of the call for evidence for more information on the indicative list of harms to children.

[12] https://support.google.com/youtube/answer/2801999?hl=en&ref_topic=9282679#zippy=%2Cage-restricted-content%2Ccontent-featuring-minors

[13] https://support.google.com/youtube/answer/2802245?hl=en&ref_topic=9282679

Question 12: How do terms of service or public policy statements treat 'primary priority' and 'priority' harmful content?[11]

content removal or channel termination. Further policies are detailed here, and more are available on their website[14]:

- Videos containing fetish content will be removed or age restricted.
- Sexually explicit content featuring minors and content that sexually exploits minors is not allowed on YouTube. Content containing child sexual abuse imagery is reported to the National Centre for Missing and Exploited Children, a US based agency that works with global law enforcement agencies.
- Content that depicts clothed or unclothed genitals, breasts, or buttocks that are meant for sexual gratification is not allowed.
- Pornography, the depiction of sexual acts, or fetishes that are meant for sexual gratification is not allowed.
- Masturbation, using sex toys, fondling, or groping of genitals, breasts, or buttocks is not allowed.
- Non-consensual sex acts or unwanted sexualisation is not allowed.

Instagram's Community Guidelines[15], managed by Meta- the company that also manages Facebook- include primary priority and priority harmful content in their policies:

- Nudity is not allowed on Instagram, including photos, videos, and some digitally-created content that show sexual intercourse, genitals, and close-ups of fully nude buttocks.
- There are times when Instagram may remove images that show nude or partially nude children, even when this content is shared with good intentions by the guardians of the child.
- Content that contains credible threats, hate speech, serious threats of harm to public and personal safety, encouragement of self-injury and violence is not allowed.

Instagram/Meta also has a specific Child Sexual Exploitation, Abuse, and Nudity Policy[16]:

- Instagram will report any content or activity that sexually exploits or endangers children to the National Centre for Missing and Exploited Children.

[14] https://support.google.com/youtube/answer/2802002?hl=en&ref_topic=9282679#zippy=%2Cother-types-of-content-that-violate-this-policy%2Cage-restricted-content

[15] https://help.instagram.com/477434105621119/?helpref=hc_fnav

[16] https://transparency.fb.com/en-gb/policies/community-standards/child-sexual-exploitation-abuse-nudity/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fchild_nudity_sexual_exploitation

Question 12: How do terms of service or public policy statements treat 'primary priority' and 'priority' harmful content?[11]

- Child Sexual Exploitation is defined by Meta as "Content or activity that threatens, depicts, praises, support, provides instructions for, makes statements of intent, admits participation in or shares links of the sexual exploitation of children (real or non-real minors, toddlers, or babies)" [ibid].

- This includes content that features but is not limited to Sexual intercourse; Children with sexual elements such as sex toys, stripping, open mouth kissing; Content of children in a sexual fetish context; Content that support, promotes, or encourages participation in paedophilia.

- This policy also includes content that solicits child sexual abuse material, nude imagery of children, engaging in implicitly sexual conversations in private messages with children, obtaining or requesting sexual material from children in private messages, arranging or planning real-world sexual encounters with children, etc.

- Instagram/Meta includes a warning screen so that people are aware that certain content may be disturbing and limits the ability to view the content to adults aged 18 and over. This content includes videos or photos that depict police officers or military personnel committing non-sexual child abuse and "Imagery of non-sexual child abuse, when law enforcement, child protection agencies or trusted safety partners request that we leave the content on the platform for the express purpose of bringing a child back to safety" [ibid].

Please note that the information above is not an exhaustive list. You can find more information by visiting the websites cited below in the Footnotes.

Question 13: What can providers of online services do to enhance children's accessibility and awareness of reporting and complaints mechanisms?

Is this a confidential response? (select as appropriate)

No

There are several points that we believe online service providers need to consider when looking to enhance children's awareness of reporting:

- Firstly, a considerable majority of children will not report because they do not have faith in the reporting system
- Children who have engaged in sending sexually explicit images and/or videos oftentimes feels ashamed or embarrassed.
- Many children are unaware of what constitutes as a reportable offence and why decisions are made after a report.

Question 13: What can providers of online services do to enhance children's accessibility and awareness of reporting and complaints mechanisms?

- Lastly, there is also the fear of being in trouble with their guardians, school, authorities, etc.

With these points in mind, we would recommend that providers consult directly with children via a youth advisory forum in order to create child-friendly, de-stigmatising explainers and straightforward reporting and support mechanisms, with particular attention towards children with learning challenges, disabilities, language barriers, etc.

Question 14: Can you provide any evidence or information about the best practices for accurate reporting and/or complaints mechanisms in place for legal content that is harmful to children, or users who post this content, and how these processes are designed and maintained?

Is this a confidential response? (select as appropriate)

No

There are reporting functions on most social media platforms, which can be viewed from their pages. These reporting functions include reporting any kind of abusive content, meaning that there is no separate space for content that is harmful to children:

- Facebook[17]: The best way to report abusive content or spam on Facebook is by using the Report link near the content itself. You need to have an account or a ask a friend to make a report.
- Instagram[18]: The best way to report abusive content or spam on Instagram is by using the Report link near the content itself. You can also report a post or profile on Instagram. If you do not have an account, you can report using a form on their Help Centre webpage (linked in footnotes)
- Twitter[19]: Anyone can report abusive behaviour directly from a Tweet, profile, or Direct Message, as well as from a List, Moment, or Twitter Space. All reports are kept anonymous.
- TikTok[20]: You can anonymously report a video that you believe violates TikTok Community Guidelines. You need to have an account or a ask a friend to make a report.
- YouTube[21]: Reporting content is anonymous, which is then reviewed against YouTube Community Guidelines and can result in the video being removed or age restricted if the content is deemed not appropriate for younger audiences. Content that can be reported includes videos, shorts, channels, playlists, thumbnails, comments, live chat messages, and ads.

[17] https://www.facebook.com/help/reportlinks/

[18] https://help.instagram.com/2922067214679225

[19] https://help.twitter.com/en/rules-and-policies/twitter-report-violation

[20] https://support.tiktok.com/en/safety-hc/report-a-problem/report-a-video

[21] https://support.google.com/youtube/answer/2802027?hl=en&co=GENIE.Platform%3DAndroid

Question 14: Can you provide any evidence or information about the best practices for accurate reporting and/or complaints mechanisms in place for legal content that is harmful to children, or users who post this content, and how these processes are designed and maintained?

There are numerous charities and organisations that provide anonymous reporting tools as well as guidance on what is illegal and how to report to the relevant services and/or authorities. These organisations have specific sections on abusive content, inappropriate content, and content that is harmful to children, making it easier to navigate to the relevant reporting scheme:

From Childnet[22]:

- On many services, you can report content such as images, videos, and text as well as other users, comments, and adverts for content including impersonation, hate speech, violent or extreme content, pornographic content, harassment, threats, abuse and bullying
- You can make a report and get further advice by visiting Report Harmful Content hosted by the Internet Watch Foundation (IWF)[23]
- You can also make a report on a specific service, some of which are included on this site as well as other places to report to, such as CEOP, The Advertising Standards Authority, and Childline's Report remove tool

From NSPCC[24]:

- Report to CEOP if you suspect an adult is communicating with a child inappropriately
- Report to IWF using their anonymous reporting portal if you come across an indecent image of a child online, or you know a young person who has had a sexual image or video of themselves shared online.

Stop It Now[25]:

- Stop it Now helpline, secure chat, and secure email services that can provide advice for anyone worried about their own sexual thoughts, feelings, and behaviour towards children; anyone worried about another adult's online or offline sexual behaviour towards children; anyone concerned about a young person's sexual behaviour. This service is anonymous, and any details shared that identifies a child who has been or is at risk of being abused, then that information will be passed to the appropriate agencies

Internet Watch Foundation[26]:

- You can anonymously and confidentially report child sexual abuse pictures or videos on the internet, non-photographic child sexual abuse images.

[22] https://www.childnet.com/help-and-advice/how-to-make-a-report/

[23] https://reportharmfulcontent.com/

[24] https://www.nspcc.org.uk/keeping-children-safe/online-safety/online-reporting/

[25]https://www.stopitnow.org.uk/helpline/?utm_source=gov&utm_medium=safetybydesign&utm_campaign=gov-illegalharm

[26] https://www.iwf.org.uk/about-us/how-we-assess-and-remove-content/

Question 14: Can you provide any evidence or information about the best practices for accurate reporting and/or complaints mechanisms in place for legal content that is harmful to children, or users who post this content, and how these processes are designed and maintained?

- IWF analysts use technical internet tracing techniques to locate criminal content around the world.
- When content is traced to an INHOPE country, IWF follow INHOPE procedures and send a report to the INHOPE reporting system which then forwards the report to the relevant INHOPE hotline. INHOPE is a global network of hotlines dedicated to combatting all forms of child sexual abuse material (CSAM) on the internet.
- If criminal content is hosted in the US, IWF sends Simultaneous Alerts direct to the hosting company at the same time as sending it to the US Hotline NCMEC
- If criminal content is hosted in a country without an INHOPE hotline, IWF report it to the UK National Crime Agency who then forward on to Interpol
- IWF also has a membership service that shares blocking lists, alert services, and databases for making services a safer place for children.

Question 15: What actions do or should services take in response to reports or complaints about online content harmful to children (including complaints from children)?

Is this a confidential response? (select as appropriate)

No

The first step should be to assess the risk- is the report in regards to harmful content present, or that the child is the subject of harmful content?

Reports should be taken seriously and with full cooperation. When a report comes through, there should be an option to signpost the child to support services, as well as further details about what happens next. This could also include a referral code for the report.

Question 16: What functionalities or features currently exist that are designed to prevent or mitigate the risk or impact of content that is harmful to children? A1.21 in the call for evidence provides some examples of functionalities.

Is this a confidential response? (select as appropriate)

No

Question 16: What functionalities or features currently exist that are designed to prevent or mitigate the risk or impact of content that is harmful to children? A1.21 in the call for evidence provides some examples of functionalities.

In 2021, YouTube announced new Safety and Digital Wellbeing[27] options for children and young people, targeting default settings that could be used to mitigate risk:

- Adjusting the default upload setting to the most private option available for users ages 13-17
- Turning 'take a break' and bedtime reminders on by default for all users ages 13-17
- Additional parental controls on YouTube Kids, including the ability for a parent to choose a "locked" default autoplay setting.
- Removing overly commercial content on YouTube kids, such as a video that only focuses on product packaging or directly encourages children to spend money. Paid product placements will continue to be unallowed on the platform.
- On YouTube, the company has updated the disclosures that appear on "made for kids" content or supervised accounts when a creator identifies that their video contains paid promotions. The disclosures appear in an "easy to read" text and link to a "kid friendly" animated video which provides additional information on paid product placement.

In 2017, Instagram released new age-appropriate features[28]:

- Restricting DMs between teens aged under 18yrs and adults they don't follow. For example, when an adult tries to message a teen who doesn't follow them, that adult receives a notification that DM'ing them isn't an option.
- Using safety notice prompts to encourage teens aged under 18yrs to be caution in conversations with adults they are already connected to, or "friends" with on the platform.
- Safety notices in DMs will notify young people when an adult who has been exhibiting potentially inappropriate behaviour is interacting with them in their DMs
- Making it more difficult for adults to find and follow teens by restricting adults who have been "exhibiting suspicious behaviour" from seeing teen accounts in the "Suggested Users" section, prevent them from discovering teen content in "Reels" or "Explore", and automatically hiding their comments on public posts by teens.
- Adding a new step that, when someone under 18 signs up to Instagram, gives users the option to choose between a public or private account.
- If the teen doesn't choose 'private' when setting up their account, Instagram will send notifications later on in the app to highlight the benefits of a private account and reminding them to check their settings.

In 2023, TikTok released new features for teens and families to come into effect[29]:

[27] https://blog.youtube/news-and-events/new-safety-and-digital-wellbeing-options-younger-people-youtube-and-youtube-kids/

[28] https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community

[29] https://newsroom.tiktok.com/en-us/new-features-for-teens-and-families-on-tiktok-us

Question 16: What functionalities or features currently exist that are designed to prevent or mitigate the risk or impact of content that is harmful to children? A1.21 in the call for evidence provides some examples of functionalities.

- every account belonging to a user below age 18 will automatically be set to a 60-minute daily screen time limit. If the 60-minute limit is reached, teens will be prompted to enter a passcode in order to continue watching. For children under 13 years old, the daily screen time limit will also be set to 60 minutes, and a parent or guardian will need to set or enter an existing passcode to enable 30 minutes of additional watch time.
- If a teen user below the age of 18 opts out of the default time limit setting, they will be prompted to set a custom time limit.
- TikTok already implements a default private account setting to all account users aged 13-15.
- DM'ing is only available to users aged 16+, and to host a "LIVE", users must be 18.
- In 2020, TikTok developed Family Planning, which allows a guardian to link their personal TikTok account to their teen/child's account and set controls, including screen time management, restricted mode (limiting the appearance of content that might not be appropriate for all audience), and restrictions on DM'ing, including turning of DM'ing entirely.
- In this updated article, Family Planning will introduce custom screen time limits as well as a screen time dashboard which summarises time on the app, the number of times TikTok was opened, and a breakdown of total time spend during the day and night.
- Already in effect is the restriction on who can message users of any age, meaning that only approved followers can message each other, and no images or videos can be sent.
- Guardians will be able to enable a set schedule to mute notifications for their children. Accounts aged 13-15 already do not receive push notifications from 9pm and accounts aged 16-17 have push notifications disabled from 10pm

Question 17: To what extent does or can a service adopt functionalities or features, designed to mitigate the risk or impact of content that is harmful to children on that service?

Is this a confidential response? (select as appropriate)

No

Please see the responses to Question 16 and Question 5.

Question 18: How can services support the safety and wellbeing of UK child users as regards to content that is harmful to them?

Is this a confidential response? (select as appropriate)

No

Something that we see as Early Intervention and Prevention educators is that there is a large gap in educating guardians, carers, and other persons responsible for children's safety and education on the ways to promote a safe and accessible online experience with their young people. This also includes education around what to do when a child comes to a guardian/carer with a safety concern. These carers need to be supported and kept up to date so they feel more empowered in talking with children about keeping safe, how to report any dangerous or inappropriate behaviour, and how to find child-friendly spaces online.

Question 19: With reference to content that is harmful to children, how can a service mitigate any risks to children posed by the design of algorithms that support the function of the service (e.g. search engines, or social and content recommender systems)?

Is this a confidential response? (select as appropriate)

No

It is harder for search engines to detect content that is harmful to children, but certain companies are doing what they can to tackle the spread of CSAM.

Google Search Policy[30]:

- Block search results that lead to child sexual abuse imagery or material that appears to sexually victimise, endanger, or otherwise exploit children
- Filter out explicit search results if the search query seems to be seeking CSAM
- For queries seeking adult explicit content, Search won't return imagery that includes children in order to break the association between children and sexual content.
- In some countries, users who enter queries into Search that "clearly relate to CSAM" are shown a warning that child sexual abuse imagery is illegal, alongside information on how to report this content to trusted organisations.

Microsoft Digital Safety Content Report[31]:

- Code of Conduct specifies that their platform prohibits "any activity that exploits, harms, or threatens to harm children across our products and services –

[30] https://protectingchildren.google/#fighting-abuse-on-our-own-platform-and-services

[31] https://www.microsoft.com/en-us/corporate-responsibility/digital-safety-content-report?activetab=pivot_1%3aprimaryr3

Question 19: With reference to content that is harmful to children, how can a service mitigate any risks to children posed by the design of algorithms that support the function of the service (e.g. search engines, or social and content recommender systems)?

including but not limited to distribution of child sexual exploitation and abuse imagery (CSEAI), and grooming of children for sexual purposes".

- Employs hash-matching technology (Microsoft developed PhotoDNA) to detect CSEAI as well as in-product reporting for products such as Skype, Xbox, and Bing.

Question 20: Could improvements be made to content moderation to deliver greater protection for children, without unduly restricting user activity? If so, what?

Is this a confidential response? (select as appropriate)

No

Reporting systems and complaints procedures are frequently overrun with reports and do not have enough staff to sift through all the complaints. One way to tackle this problem would be to have a separate reporting procedure dedicated to content harmful to children. Many users can identify what the problem is with a post or user that they are reporting, however having a dedicated reporting scheme and analyst team to specialise in child safety concerns could free up the general reporting scheme as well as more efficiently tackle abusive and harmful content directed at children in a timely manner that would allow for swift action.

Another possible solution is increasing what the World Economic Forum calls "the human-curated, multi-language, off-platform intelligence"[32] into AI learning sets. One problem with AI and other machine learning content moderation is that AI services cannot always detect context and nuance; the World Economic Forum reported that "For example, robust AI models exist to detect nudity, but few can discern whether that nudity is part of a renaissance painting or a pornographic image. Similarly, most models can't decipher whether the knife featured in a video is being used to promote a butcher's equipment or a violent attack" [ibid]. With more human moderators inputting more diverse data from around the world through the internet, including the dark web, AI training sets will be able to refine the content they flag without infringing on free speech and user activity.

[32] https://www.weforum.org/agenda/2022/08/online-abuse-artificial-intelligence-human-input/

Question 21: What automated, or partially automated, moderation systems are currently available (or in development) for content that is harmful to children?

Is this a confidential response? (select as appropriate)

No

Google[33]:

- Uses deep neural networks (a series of algorithms meant to recognise underlying relationships in a set a data that mimics the way the human brain operates) for image processing to assist human reviewers sorting through images by prioritising the most likely child sexual abuse material (CSAM) content for review.
- Hash-matching technology creates a 'hash' or unique digital fingerprint for an image or video so that it can be compared to the 'hash' of known CSAM.
- Content Safety API helps organisations classify and prioritise potential abuse content for review using machine learning classifiers.

YouTube[30]:

- CSAI Match API is used for video hash-matching to tag and remove known CSAM videos as well as identify re-uploads of previously identified CSAM.

From the Cyberbullying Research Centre[34]:

- Natural language processing involves using machines to take human language in text or audio format and decipher what is meant, including subtleties, nuances, turn of phrase, tone, and colloquialisms.

Thorn[35], a non-profit that builds technology to "defend children from sexual abuse":

- Uses cryptographic, perceptual hashing and machine learning algorithms identify known and unknown CSAM that is flagged for further review.

IWF and The Lucy Faithfull Foundation developed a chatbot that, each time someone has searched for a word or phrase that could be related to CSAM on Pornhub's UK website, appears and asks the user "whether they want to get help with the behaviour they are showing". According to a Wired article, during the first 30 days of the chatbot's trial, Pornhub users triggered the system 173,904 times[36].

[33] https://protectingchildren.google/#fighting-abuse-on-our-own-platform-and-services

[34] https://cyberbullying.org/machine-learning-can-help-us-combat-online-abuse-primer

[35]https://safer.io/?__hstc=208625165.149ef04d5367c0be8591115cd48c34c1.1673530459640.1673530459640.1673530459640.1&__hssc=208625165.1.1673530459641&__hsfp=4029266239

[36] https://www.wired.co.uk/article/pornhub-search-child-abuse-chatbot

Question 22: How are human moderators used to identify and assess content that is harmful to children?

Is this a confidential response? (select as appropriate)

No

There is a limit to machine learning, artificial intelligence, and other computer-generated software designed to detect CSAM. Human moderators are able to pick up on the nuances, keywords, and slang terms that CSAM are tagged by in a way that machine technology cannot. As the World Economic Forum reports, "…threat actors use increasingly sophisticated tactics to avoid evolving detection mechanisms. This has resulted in the development of new slang, like child predators referring to 'cheese pizza' and other terms involving the letters 'c' and 'p' instead of 'child pornography"[37]. AI relies on data sets that contain familiar language, known online abuses and profiles, which means that any AI program can be quick to detect known CSAM, but detecting nuances is something that requires language markers that human moderators can pick up on at a much more reliable consistency.

Question 23: What training and support is or should be provided to moderators?

Is this a confidential response? (select as appropriate)

No

Besides what we have already recommended in our answers in previous sections, we would also recommend the following:

- Training on the dynamics of child-to-child image sharing and self-generated nude/semi-nude images
- Training on the language in terms of victim blaming and safeguarding.
- Training on child sexual exploitation, coercive control, and grooming.

Question 24: How do human moderators and automated systems work together, and what is their relative scale? How should services guard against automation bias?

Is this a confidential response? (select as appropriate)

No

[37] https://www.weforum.org/agenda/2022/08/online-abuse-artificial-intelligence-human-input/

Question 24: How do human moderators and automated systems work together, and what is their relative scale? How should services guard against automation bias?

It is necessary for human moderators and automated systems to work together to identify and remove CSAM, harassment, and other online abuses against children. Language researchers are training AI systems to spot potentially abusive language and imagery, including captions on social media posts. This requires multiple language researchers from a variety of sources and organisations to collaborate on cross-party platforms and portals to share information. In a study done by Cornell University[38], researchers found that user behaviour on a controlled, yet realistic, social media platform when shown harassment content were more likely to question AI moderators and the accountability of the AI system to pick up on ambiguous abusive content- content that is hard to detect without human moderator input to evaluate tone, context, and cultural language. The researchers concluded that "Even if AI could effectively moderate content… there is a [need for] human moderators as rules in community are constantly changing, and cultural contexts differ" [ibid].

For further information on automation bias, please see our answers to Questions 20 and 22.

Question 25: In what instances is content that is harmful to children, that is in contravention of terms and conditions, removed from a service or the part of a service that children can access?

Is this a confidential response? (select as appropriate)

No

Please see the response Question 12 for more examples.

TikTok will remove content if it includes[39]:

- Activities that perpetuate the abuse, harm, endangerment, or exploitation of minors (defined as any person under the age of 18)
- Depictions of abuse in digitally created or manipulated media.
- Offers to trade, sell, or directions for users off-platform to obtain or distribute CSAM.
- Engaging with minors in a sexual way or otherwise sexualises a minor.
- Promotion and normalisation of paedophilia and grooming behaviours
- Re-enactments of assault or confessions
- Soliciting real-world contact between a minor and an adult or between minors with significant age differences
- Solicitation of nude imagery

[38] https://news.cornell.edu/stories/2022/10/trust-online-content-moderation-depends-moderator

[39] https://www.tiktok.com/community-guidelines?lang=en#31

Question 25: In what instances is content that is harmful to children, that is in contravention of terms and conditions, removed from a service or the part of a service that children can access?

- Minor sexual activities including penetrative and non-penetrative sex, oral sex, or intimate kissing.
- Sexual fetish involving a minor.
- Exposed genitals, buttocks, the pubic region, or female nipples of a minor.
- Sexualised comments, emojis, text, or other graphics used to veil or imply nudity or sexual activity of a minor.
- Suggests, depicts, imitates, or promotes the possession or consumption of alcoholic beverages, tobacco, or drugs by a minor.
- Instructions targeting minors on how to buy, sell, or trade alcohol, tobacco, or controlled substances.
- Promotion of activities that may jeopardise youth well-being such as physical challenges, dares, and stunts.
- Depiction of physical abuse, neglect, endangerment, or psychological disparagement of minors
- Depiction, instruction, and promotion of suicide or self-harm
- Depiction, instruction, and promotion of disordered eating and dangerous weight loss behaviours
- Insults about an individual or group
- Wishes death, serious disease, or other serious harm to the user.
- TikTok interactive features to degrade others.
- Depiction of wilful harm or intimidation, including cyberstalking, blackmail, doxing, or trolling

Please note that this is not an exhaustive list. More information can be found on the TikTok Community Guidelines Policy referenced in the footnotes.

Question 26: What other mitigations do services currently have to protect children from harmful content?

Is this a confidential response? (select as appropriate)

No

From INHOPE, a global network of hotlines dedicated to combatting all forms of child sexual abuse material (CSAM) on the internet[40]:

- PhotoDNA: creates a ‘hash’ (unique digital signature) of an image which is then compared against hashes of other photos to find copies of the same image. When matched with a database containing hashes of pre-existing CSAM, PhotoDNA can report the image.

[40] https://inhope.org/EN/articles/what-kinds-of-technology-used-in-the-fight-against-child-sexual-abuse-materialare

Question 26: What other mitigations do services currently have to protect children from harmful content?

- AviaTor: Augmented Visual Intelligence and Targeted Research helps law enforcement prioritise NCMEC reports so they can work on efficiently identifying perpetrators.
- Safer: programme that allows technology platforms to identify, remove, and report CSAM by matching uploaded content against the hashes of known CSAM. This platform is shared by member businesses, allowing for all kinds of online services to use the portal.
- Child Protection System: scans file sharing networks and chatrooms to find computers that are downloading and sharing CSAM online. This system operates in 96 countries and collects 30-50 million reports a day, allowing the Child Rescue Coalition to expose hidden networks of abusers online.

Please see the responses to Question 16 and Question 21 for more examples.

Question 27: Where children attempt to circumvent mitigations in place on a service, what further systems and processes can a service put in place to protect children?

Is this a confidential response? (select as appropriate)

No

As with some of the online platforms described in previous answers, online providers should issue reminders on screen about the importance of being truthful about age and any warnings for potentially harmful content. This could also extend to sending push notifications to banking accounts, internet, or phone service provider accounts, etc. if the service is oriented for adults (such as pornography).

Question 28: Other than those covered above in this document (the call for evidence), are you aware of other measures available for mitigating the risk, and impact of, harm from content that is harmful to children?

Is this a confidential response? (select as appropriate)

No

Question 28: Other than those covered above in this document (the call for evidence), are you aware of other measures available for mitigating the risk, and impact of, harm from content that is harmful to children?

N/A