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.

Not confidential

The Center for Countering Digital Hate (CCDH) is a research and advocacy organisation working to disrupt the dissemination of online hate and reform the digital architecture which underpins its growth. Our research has evidenced the repeated failures by social media companies to protect users and enforce their own terms of service. Our advocacy seeks to increase the economic, reputational and political cost for supporting and profiting from online hate. This has included successfully forcing platforms to remove dangerous content and campaigning for legislation that changes the fundamental business model underpinning this toxic ecosystem.

CCDH research has shown dangerous threats to children using social media and search services. Our research has exposed how a video sharing platform’s algorithm directed young users to body image and self harm content; how age restrictions in virtual reality platforms failed to prevent minor abuse; and how search results for terms related to body image (a matter highly relevant in adolescence) featured pathways into violent online communities. Overall, CCDH believes without significant reform, children will continue to encounter content which threatens real world harm on user-to-user platforms and search services.

The work of the Center is carried out by two organisations, which operate collaboratively in carrying out their shared mission. CCDH US is a US nonprofit 501(c)(3) corporation headquartered in Washington DC, and CCDH UK is a UK nonprofit company headquartered in London.

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

Not confidential

Not applicable to CCDH.

Question 3: What information do services have about the age of users on different platforms (including children)?
Not confidential
Not applicable to CCDH.

Question 4: How can services ensure that children cannot access a service, or a part of it?
Not confidential
Not applicable to CCDH.

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?
Not confidential.
Not applicable to CCDH.

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?

Not confidential

CCDH research has evidenced the presence of content that is harmful to children on user-to-user and search services. The full catalogue of our research can be found in our research archive; this answer cites evidence concerning child users and content harmful to children.

CCDH's report "Deadly By Design" examined how the social media platform TikTok promoted eating disorder and self-harm content into childrens' feeds. Accounts were created for 13 year old users in the USA, UK, Australia and Canada, recorded and analysed to identify the types of content directed to users by the TikTok algorithm. The report showed the frequency with which body image, mental health, self harm and eating disorder content are encountered by TikTok's youngest users. Videos posted to hashtags associated with eating disorder content that we identified had over 13.2 billion views, often evading moderation by using coded hashtags. TikTok recommended eating disorder and self-harm content to new teen accounts within minutes: suicide-related content within 2.6 minutes and eating disorder content within 8 minutes. Body image and mental health content was the most frequently encountered, on average every 39 seconds. This research evidenced the presence of content that is harmful to children being directed towards child users despite TikTok being in full knowledge of users' age.

CCDH's report "The Incelosphere" analysed a user-to-user digital forum for men who identify as 'incels', meaning involuntary celibates who feel unable to form sexual relationships. We identified users aged between 15 and 17 amongst the forum's most active members. The site features misogynistic, racist, antisemitic and homophobic language, with 16% of all posts featuring misogynistic slurs. Comments about or mentioning rape were posted every 29 minutes on average. Over a quarter of forum users posted pedophilia keywords, with analysis determining 53% of posters were supportive. This user-to-user service exposes child users (the teen users identified) to harm while also promoting illegal harms to children offline (pedophilia).

As part of the same report, CCDH investigated the role of search services in surfacing this forum, discovering that Google searches for body image or unemployment keywords featured on the first page of Google search results. This demonstrates the ease with which users, including child users, who become concerned about their appearance could encounter dangerous user-to-user forums like the incelosphere.

CCDH has catalogued the harms to children using new virtual reality user-to-user services. Our report "Horizon Worlds Exposed" evidenced bullying, sexual harassment and harmful content encountered by minors in Meta's flagship VR product. At the time of the research, the virtual reality product was limited to users aged 18+, but child users were present in the majority (66%) of the most popular social spaces in Horizon Worlds. Meta has since announced that it will open the VR social network to teens aged 13 to 17.

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?

Researchers identified 19 incidents of abuse directed at child users by adults, including sexually explicit insults and racial, misogynistic and homophobic harassment. Minors were also identified within virtual strip clubs serving alcoholic drinks for in-app money. In earlier CCDH research on Meta's VR Chat, researchers found child users being exposed to abusive behaviour, such as graphic sexual content, bullying and racism, every seven minutes on average.

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

Not confidential

CCDH recognises the large corpus of medical and social science research on the damaging impacts on children from exposure to the types of harm we evidence in our research.

Suicide and self-harm content in particular have been shown to impact child users, with exposure increasing the likelihood of users performing imitative acts (see findings of Arendt et al. in the journal *New Media & Society,* for example). Although this is not a primary focus of CCDH research, we have partnered with expert organisations better placed to evidence these impacts, including our recent work with the Molly Rose Foundation to produce a parents' guide for childrens' TikTok usage.

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

Not confidential

Not applicable to CCDH.

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?

Not confidential.

Services must consider the age demographics of their user base as a risk factor in the UK. For example, the short form video app TikTok we studied in "Deadly By Design" has a high proportion of young and teenage users. As such, the priority harms we identified (eating disorder, self-harm and suicide content) have potentially greater impact on children in the UK than the same content would have on another platform.

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

Not confidential.

CCDH advocates a global standard of social media regulation called the "STAR Framework" with advice for governance, accountability and decision-making structures towards platform and user safety.

STAR's "S" stands for the principle of safety by design. Safety has not been prioritised or implemented to a sufficient degree by any part of these companies' governance structures. Safety by design changes this, requiring that service providers take proactive steps to ensure products are safe, particularly for child users, before public launch. For too long, child users have been part of a sociological experiment conducted by those who govern social media platforms in pursuit of novel ways to generate profits.

"T" stands for transparency, which CCDH sees as fundamental to accountability. The information asymmetry, in which the platforms withhold and obscure information about their products, makes it impossible to hold these services accountable for child user and platform safety. Making algorithms, rules of enforcement and business economics more transparent is critical to instituting proper accountability structures for user safety.

"A" stands for answerability to democratic and independent bodies. Regulation is most effective where there are accountability systems in place for harm caused, particularly where there is a risk of inaction because of profit motives and commercial factors. Frequently, accountability systems include an enforcement and independent pathway for challenging decisions or omissions. These external requirements, such as contained in the Online Safety Bill, will spur the creation of governance and decision-making structures within platforms to accommodate their new responsibilities.

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

“R” stands for responsibility for companies and their senior executives. This is a critical component of new decision-making structures in company governance, as personal liability makes real the consequences of actions/inactions leading to user harm.

In conclusion, CCDH believes that governance, accountability and decision-making structures based on our STAR framework are necessary for child user and platform safety.

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)?

Not confidential.

CCDH has not produced research on this topic.

Question 12: How do terms of service or public policy statements treat ‘primary priority’ and ‘priority’ harmful content?[1]

Not confidential.

CCDH has analysed user-to-user and search services’ public policy statements and terms of service. Often, these terms and statements identify the types of content we here consider ‘primary priority’ and ‘priority’ harmful content as violations of their community standards. Examples include policies on eating disorder content at TikTok, hateful conduct policy at Twitter, and threats of violence at Instagram.

However, CCDH research shows platforms consistently fail or refuse to act on harmful content, even when it breaches their own terms and policies. For example, despite

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

Question 12: How do terms of service or public policy statements treat ‘primary priority’ and ‘priority’ harmful content?[1]

TikTok’s public policy on eating disorder content, CCDH found a massive corpus of content in researching our report “Deadly By Design” – much of which remains accessible to UK users without attached user warnings.

In our research series “Failure to Protect” CCDH evidenced Instagram, Facebook and Twitters’ failure to act on abusive content. In “Hidden Hate”, CCDH showed how Instagram failed to act on 9 out of 10 reports of threats of violence sent to users in the study. The platforms’ failure to enforce their own rules affects all users, but is particularly worrisome when it comes to children.

In summary, CCDH finds that services currently address and say they treat seriously the types of content here deemed ‘primary priority’ and ‘priority’ in terms of service, but fail to enforce these in practice.

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

Not confidential.

CCDH has undertaken research on the accessibility of reporting and complaints mechanisms for children using Meta’s VR app store. In our report “Facebook’s Metaverse” researchers identified 100 potential violations of Meta’s policies for VR in 11 hours and 30 minutes of recording user behaviour in the app. Abusive behaviour exposed by the recordings included graphic sexual content, bullying and abuse, grooming and threats of violence. Just 51 incidents could be reported to Meta using a web form created by the platform for this purpose, as the platform refuses to examine policy violations if it cannot match them to a pre-defined category or username in its database. None of the 51 reports of policy violation were acknowledged by Meta in any way. This research is an example of the shocking lack of accessibility to a complaints and reporting mechanism.

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?

Not confidential.

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?

Reporting mechanisms must be accessible and user friendly, especially for child users. CCDH research suggests that this is particularly lacking in novel, non-conventional social media like virtual reality. As detailed in question 13, the complaints mechanisms available to users of Meta's VR app store were not easily accessed and required multiple user steps to reach. Once the complaints mechanism had been found, the functionality was limited given the platform did not take complaints for matters outside pre-classified criteria.

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)?

Not confidential.

A foundational step towards best practice in this area is acknowledgement and receipt of complaint. In our Metaverse study, CCDH did not receive a single acknowledgement or case file, let alone an enforcement action, in the 51 complaints lodged about content and user activity raising likely breaches of Meta's rules.

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.

Not confidential.

CCDH has not produced research on this topic.

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?

Not confidential.

CCDH has not conducted research on this topic.

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

Not confidential.

Following publication of CCDH's research on TikTok, in which we evidenced the platform's algorithm directing 13 year old users to self-harm and eating disorder content, the company responded by adding a limited number of warnings and resources to the content we identified. However, there was a massive difference between jurisdictions: in the US, 71% of the videos using coded hashtags CCDH identify now carrying warnings for users. In the UK however, that number falls to just 7%.

CCDH supports the inclusion of content warnings, signposting and assistance resources as services platforms can use to mitigate risks to child users. However, we are concerned about the variance in jurisdictional application, particularly as our research revealed fewer protections for child users in the UK compared to their American counterparts.

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)?

Not confidential.

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)?

CCDH has not conducted extensive research on this topic. But it may be useful to note that our research found TikTok's algorithm and recommender systems were more adept at identifying, and thus mitigating the harms to children posed by, self-harm and suicide content than for eating disorder hashtags and eating disorder content. This indicates that platforms are capable of instituting design choices to mitigate the risks of priority harmful content, but need to broaden their work to other areas of the harms landscape facing child users.

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

Not confidential.

Services must invest in content moderation teams and content moderators to protect child users. A recent CCDH study involved assessing content moderation on social media. As part of the business restructuring following Elon Musk's takeover of the platform in October 2022, content moderators and moderation teams were fired en masse. CCDH examined how Musk's takeover and restructuring influenced user speech on the platform. We found an atypical increase in hate speech: finding homophobic, misogynistic and antisemitic slurs increased substantially (from a 39% increase in use of a homophobic slur to a 67% increase in use of an antisemitic slur) compared with the average for the year-to-date prior. Although this study did not examine hate speech specifically directed at children, these tweets are not age restricted and visible to young users. This study and other research leads CCDH to the view that improvements must be made to content moderation, particularly in substantial resourcing and dedicated human moderators, to deliver greater protections for child users.

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

Not confidential.

Question 21: What automated, or partially automated, moderation systems are currently available (or in development) for content that is harmful to children?
Not applicable to CCDH.

Question 22: How are human moderators used to identify and assess content that is harmful to children?
Not confidential.
Not applicable to CCDH.

Question 23: What training and support is or should be provided to moderators?
Not confidential.
Not applicable to CCDH.

Question 24: How do human moderators and automated systems work together, and what is their relative scale? How should services guard against automation bias?
Not confidential.

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

Not applicable to CCDH.

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?

Not confidential.

CCDH is not an online service provider but has offered evidence in response to question 12 showing that removals are refused or inconsistently enforced even in circumstances where content violates terms and conditions.

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

Not confidential.

CCDH has not produced research on this topic.

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?
Not confidential.
In our research "Horizon Worlds Exposed" CCDH found child users on a platform that, at the time, claimed to be exclusively for adult users. It is clear that some children will attempt to/successfully circumvent age restrictions. But this is not sufficient to explain or excuse the lack of age requirement verification and enforcement CCDH encountered on the platform.

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?
Not confidential.