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* The Center for Countering Digital Hate (CCDH) is a not-for-profit NGO that seeks to disrupt the architecture of online hate and misinformation. CCDH has been at the forefront of unmasking how online platforms and search engines drive radicalisation, online harm and misinformation. The Center's work combines both analysis and active disruption of these networks and the online architecture enabling its rapid worldwide growth. We champion levers for change to increase the economic, political, and social costs of all parts of the infrastructure - the actors, systems, and culture - that support and profit from hate and misinformation (for example, climate change denial, sexual and reproductive health, anti-vaxx, antisemitism, and identity-based hate). The Center fulfils this mission in three primary ways. First, by producing research that exposes the actors, systems and culture that facilitate the spread of hate and disinformation on social media platforms. Second, by advocating for legislation that will ensure that social media platforms meet our STAR framework for addressing digital hate and disinformation, making them Safe by Design, Transparent, Accountable and Responsible.Third, by educating the public, civil society organisations and regulators about the dynamics behind the spread of digital hate and disinformation, enabling them to better address these problems and more effectively press for change. CCDH is independent, is not affiliated to any political party and does not receive money from technology companies. We believe it is impossible to serve honestly and without fear as an industry watchdog against harms an industry produces if they

	also pay our salaries. We have offices in London and Washington D.C., and connections globally. CCDH UK is a non-profit limited by guarantee and CCDH US is a 501(c)(3) non-profit.
Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and 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.**	*Is this response confidential? N* Much of the Center's work concerns the ways in which user-to-user and search services facilitate the spread of harmful hate and disinformation which falls short of constituting illegal content or behaviour. However, a number of our research projects have exposed harms caused by illegal content hosted by these services too. For example: • We have previously exposed Instagram's failure to remove extremist content linked to ISIS after it was reported to them using the platform's tools.[1] This content was easily accessible to UK users and would be considered illegal under UK law. It included graphic videos of beheadings and mass executions, and in many cases Instagram's systems had failed to apply warnings about graphic content to the footage. • Our recent Hidden Hate report investigated the way in which hatred, abuse and harassment is directed at high-profile women over private direct messages on Instagram in the UK and US.[2] Content sent over direct message included image-based sexual abuse, death threats and other threats of violence including rape threats. Much of this abuse could constitute illegal content, but our research showed that Instagram failed to act on 9 in 10 abusive direct messages when it was reported to them using their own reporting systems.

[1] "Instagram chiefs refused to axe ISIS propaganda account glorifying 9/11 and featuring execution videos", The Sun, 6 October 2020, https://www.thesun.co.uk/news/12862901/instagram-refusing-axe-isis-account-glorifying-911/

[2] "Hidden Hate", Center for Countering Digital Hate, 6 April 2022, https://counterhate.com/research/hidden-hate/

	In some cases, it is unclear whether harmful posts amount to illegal content. Our investigations of racist abuse directed at England football players around the 2021 Euro championship identified many examples of racist abuse and harassment.[3] We know that some of the abuse directed at players led to arrests and at least one conviction, but it is often impossible to know whether abuse meets this legal standard without information on its full extent, frequency, and severity without access to the account of the abused individual.[4] Transparency requirements and information accessed under those provisions will be critical for determining this.
Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content presented by your service?	*N/A*
Question 4: What are your governance, accountability and decision-making structures for user and platform safety?	*N/A*
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* Many of our reports have audited platforms' enforcement of their standards, frequently exposing their "failure to act" on content they state is not permitted. That work has demonstrated a consistent failure to act on the following when reported to them: • 87.5% of Covid and vaccine misinformation[5] • 89% of content featuring anti-Muslim hate[6]

[3] "Instagram fails to take down more than 94% of racist abuse accounts targeting England players after Euros", iNews, 15 July 2021, https://inews.co.uk/news/technology/instagram-racist-abuse-posts-england-players-after-euros-1102896

[4] "Euro 2020: Five people arrested over racist abuse of England players", BBC News, 15 July 2021, https://www.bbc.co.uk/news/uk-57848761

[5] "Marketplace flagged over 800 social media posts with COVID-19 misinformation. Only a fraction were removed", CBC, 30 March 2021, https://www.cbc.ca/news/marketplace/marketplace-social-media-posts-1.5968539

[6] "Failure to Protect: Anti-Muslim Hate", Center for Countering Digital Hate, 28 April 2022, https://counterhate.com/research/anti-muslim-hate/

	• 84% of content featuring anti-Jewish hate[7] • 94% of users sending racist abuse to sportspeople[8] • 90% of misogynist abuse sent to high-profile women over DM[9] • 100% of abuse reported in Meta's VR platform[10] • Users who repeatedly send hateful abuse[11] This demonstrates a significant gap between what platforms state in their terms of service and public policy statements, and the action that they take on content presented to them in user reports. This most fundamental gap has to be addressed by platforms enforcing their standards in a timely manner. We believe that there are two other steps beyond this that platforms could take to enhance the clarity of their terms and policies in accordance with our STAR framework: Safety by Design, Transparency, Accountability and Responsibility. First, platforms must make the private rules and guidelines used by their moderators to make enforcement decisions publicly available. Reports have exposed that moderation centres operated by large social media platforms follow much more detailed guidance than is available in those platforms' publicly stated policies. This helps create a gap between what content platforms say they will act on, and which content that will actually act on in practice. This is evident in

[7] "Failure to Protect", Center for Countering Digital Hate, 30 July 2021, https://www.counterhate.com/failuretoprotect

[8] "Instagram fails to take down more than 94% of racist abuse accounts targeting England players after Euros", iNews, 15 July 2021, https://inews.co.uk/news/technology/instagram-racist-abuse-posts-england-players-after-euros-1102896

[9] "Hidden Hate", Center for Countering Digital Hate, 6 April 2022, https://www.counterhate.com/hiddenhate; "I get abuse and threats online - why can't it be stopped?", BBC, 18 October 2021, https://www.bbc.co.uk/news/uk-58924168

[10] "New research shows Metaverse is not safe for kids", Center for Countering Digital Hate, 30 December 2021, https://counterhate.com/blog/new-research-shows-metaverse-is-not-safe-for-kids/

[11] "Twitter fails to remove 100 abusive misogynists", The Times, 13 January 2022, https://www.thetimes.co.uk/article/twitter-fails-to-remove-100-abusive-misogynists-z7nwg6d9t

	our own research which exposes a gulf between platforms' stated standards and their enforcement of those standards. Second, platforms must be much more transparent about the action they take in response to content or accounts that violate their standards. We know from dialogue with large platforms that some operate a 'strikes system' that allows accounts to post a number of violating posts in a set time period before triggering enforcement action, such as an account suspension. This contributes to a lack of clarity around what those standards are, as users can see accounts repeatedly violating standards without any visible consequences. Platforms should state in the policies what penalties they impose for policy violations, including detail on any strike systems they operate. They should also attach public information to accounts about the number of upheld policy violations or strikes that they have accrued.
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?	*N/A*
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* As noted elsewhere in our response, many of our reports have audited platforms' enforcement of their standards, frequently exposing their "failure to act" on content that violates their standards when it is reported to them. That work has demonstrated a consistent failure to act on the following when reported to them: • 87.5% of Covid and vaccine misinformation • 89% of content featuring anti-Muslim hate • 84% of content featuring anti-Jewish hate • 94% of users sending racist abuse to sportspeople • 90% of misogynist abuse sent to high-profile women over DM • 100% of abuse reported in Meta's VR platform

	• Users who repeatedly send hateful abuse This demonstrates clear failures in platforms' reporting systems. This most fundamental gap has to be addressed by platforms enforcing their standards in a timely manner, which may necessitate more investment in moderation staff and systems to protect user safety. Our research highlights three other ways in which reporting systems should be strengthened. First, reporting systems should be available in virtually every part of social media apps and web interfaces. Our Hidden Hate report found that it was simply impossible to report some content, such as voice notes sent over direct message.[12] In other cases, we have found that it is simply difficult for users to find a reporting function or that the reporting system does not allow the selection of appropriate categories for the report. Platforms can improve this by adopting a "safety by design" approach in line with our own STAR framework, ensuring that new features are accompanied by appropriate and well-functioning reporting systems which are part of the foundation of user safety on all platforms. Risk assessments and complaint duties under the Online Safety Bill will also assist with the responsiveness and effectiveness of complaints procedures for users, including children. Second, platforms should make reporting systems more flexible, allowing users to provide wider context or extra evidence as necessary. At present there are significant differences between platforms in terms of how much evidence users can submit and what grounds they may report content. Reporting systems on Twitter encourage users to attach other violating tweets from an account they are reporting, giving moderators more information on which to base a decision. Many other platforms do not offer similar options, risking poorer moderation decisions based on limited evidence.

[12] "Hidden Hate", Center for Countering Digital Hate, 6 April 2022, https://www.counterhate.com/hiddenhate

	Third, platforms should give clear explanations for any penalties they impose on content or accounts and allow users an opportunity to appeal. At present, platforms' reporting mechanisms both under- and over-moderate content on their platforms. They under-moderate when content or accounts that clearly, grossly and repeatedly violate their standards are allowed to remain on the platforms. They over-moderate when their systems, which are increasingly automated, remove content or accounts that a well-trained human moderator should have determined did not constitute a policy violation. The solution to this is to push for greater transparency and accountability. Decisions should be made transparent by declaring clearly and publicly the reason why a piece of content or an account has been removed or labelled. They should be made accountable by having an effective appeal mechanism so that poor moderation decisions are reviewed in a timely manner.
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?	*N/A*
Question 9: If your service has a *complaints* mechanism in place, how are these processes designed and maintained?	*N/A*
Question 10: What action does your service take in response to *reports* or *complaints*?	*N/A*
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* As noted elsewhere in our submission, at present platforms' reporting mechanisms both under- and over-moderate content on their platforms. They under-moderate when content or accounts that clearly, grossly and repeatedly violate their standards are allowed to remain on the platforms. They

	over-moderate when their systems, which are increasingly automated, remove content or accounts that a well-trained human moderator should have determined did not constitute a policy violation. The solution to this is to push for greater transparency and accountability. First, decisions should be made transparent by declaring clearly and publicly the reason why a piece of content or an account has been removed or labelled. Second, decisions should be made accountable by having an effective appeal mechanism so that poor moderation decisions are reviewed in a timely manner. Third, platforms must ensure that well-trained human moderators are involved in significant moderation decisions, such as the removal of influential accounts. Too often, platforms have had to walk back content or account removals that were triggered by badly designed automated systems or poorly-trained and overworked human moderators. Overreach of this kind can be minimised by ensuring well-trained human moderators are involved in the most significant decisions on content and accounts.
Question 12: What automated moderation systems do you have in place around illegal content?	*N/A*
Question 13: How do you use human moderators to identify and assess illegal content?	*N/A*
Question 14: How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services?	*N/A*

Question 15: In what instances is illegal content removed from your service?	*N/A*
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	*N/A*
Question 17: What other sanctions or disincentives do you employ against users who post illegal content?	*N/A*
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* Evidence shows that there are a wide range of design features that platforms can implement to prevent or reduce harm. Our submission will focus on three of these: effective labelling, inoculation, and increasing friction for harmful behaviour. First, studies carried out by platforms themselves have shown that labelling posts can have positive although extremely limited effects on how users interpret them and how they are shared.[13] Likewise, fact-checks or other forms of debunking can be effective, but only if they are properly designed and if they actually reach users exposed to harmful disinformation. Second, research by academics such as Sander van der Linden at Cambridge University has shown that it is possible to strengthen public resistance to harmful disinformation by exposing people to weakened forms of this content, accompanied by explanations of the motives and methods that accompany disinformation.[14]

[13] "Facebook Knows That Adding Labels To Trump's False Claims Does Little To Stop Their Spread", BuzzFeed, 17 November 2020, https://www.buzzfeednews.com/article/craigsilverman/facebook-labels-trump-lies-do-not-stop-spread

[14] Roozenbeek, J., van der Linden, S. Fake news game confers psychological resistance against online misinformation. Palgrave Commun 5, 65 (2019). https://doi.org/10.1057/s41599-019-0279-9

	Third, studies have shown that adding 'friction' to certain user behaviours likely to cause harm – effectively making particular actions slightly more difficult to perform, for example by increasing the number of button presses needed to execute them – can reduce the prevalence of harmful posts.[15] This has already been implemented on some platforms which issue users with warnings if their systems recognise that they may be about to share articles without reading them or post what appears to be hateful content.[16] However, we believe that these design features must be accompanied by primary research into emerging harms on platforms, carried out by independent civil society organisations, regulators and platforms themselves as part of their efforts to make their services safe by design. It is only through conducting investigations into the spread of hate or disinformation on a platform that it is possible to identify gaps in platform standards and enforcement, or platform features such as algorithmic amplification that are contributing to the spread of harmful content. Such research needs to be made accessible to the public: internal documents leaked by the Facebook whistle blower Frances Haugen show that platforms are otherwise willing to sit on research exposing serious harm being caused by their systems.
Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact of harm from illegal content?	*N/A*
Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	*N/A*

[15] Velásquez, N., Leahy, R., Restrepo, N.J. et al. Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. Sci Rep 11, 11549 (2021). https://doi.org/10.1038/s41598-021-89467-y

[16] "Can Twitter warnings actually curb hate speech? A new study says yes.", Protocol, 22 November 2021, https://www.protocol.com/policy/hate-speech-warnings-twitter

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?	*N/A*
Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?	*N/A*
Question 23: Can you identify factors which might indicate that a service is likely to attract child users?	*N/A*
Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?	*N/A*
Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?	*N/A*
Question 26: What information do you have about the age of your users?	*N/A*
Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	*Is this response confidential? – N* Transparency must be a pillar of efforts to address digital hate and disinformation. The principle of transparency underpins our STAR framework for legislative efforts to address these problems. We must distinguish between transparency systems that are accessible to the public, and those that would be accessible only to academics, researchers and regulators.

	The Center strongly believes that public access to information must be a priority for legislators and regulators. This is for three key reasons. First, the only way to provide the basis for a healthy and accurate public conversation about online harms is to ensure that the public has access to basic information about the content and accounts that are most popular or prevalent on platforms. Just as citizens are able to see the front pages of today's papers or listen to the news bulletins, they must have some shared access to meaningful and current information on what is trending or popular on social media platforms. Second, public transparency is necessary in order to ensure that any other information shared by platforms is accurate. We know that Meta has previously supplied academics with data that was later proven to be inaccurate, and this inaccuracy could only be detected by comparison with more readily accessible data from Meta's CrowdTangle analytics tool. Similarly, public transparency information which gives a current view of what is trending on a platform ensures that there is a strong set of publicly scrutinised data that can be compared to any transparency data that can only be accessed more privately. Third, communities themselves need to be able to research, identify and report forms of harm that may otherwise go unnoticed and unaddressed. This is important in order to avoid bias and blindspots for regulating existing and emerging forms of online harm and disinformation. While we understand that privacy and safety must be considered in terms of which data can be made fully public, the Center believes there should be transparency of algorithms, rules enforcement and economics, and that as much of this information as possible should be easily accessible to the public. At a minimum, algorithmic transparency should include:

	• Search algorithms and data – such as autocompleting a keyword and metadata used; • Recommendation algorithms and data – which curate content that a user may be interested in; • Ad-tech algorithms and data – that target users based on demographics and behaviour to optimise advertising; and • Moderation algorithms and data – that target content, users and groups that breach the law or the platform's / search engines terms and conditions / community standards. This should include internal metrics, such as the violative view rate. To help assess the impact of algorithms and products, and to identify emerging forms and trends of harm on platforms, the data above should be supported by public transparency on the most popular content on that platform (with the impact of algorithms controlled and shown). For example, Facebook's top 10 content: • Most liked • Most viewed • Most recommended. Transparency should include publicly accessible data, complemented by more access via a public API, which can be converted into a broader range of formats. There should be clarity about what meta-data is entered into the API to yield particular results. A live public service has the benefits of being faster, giving broader access, providing a public record, and being harder to falsify or mislead. Within a legislative framework, regulators and courts should have the right to access additional data to ensure legal duties are being complied with. Individuals should also have a clear right to access and share their own data. On transparency of rules enforcement, platforms and search engines need to have clear, accessible and responsive complaints/reporting systems, where terms and

	conditions / policies ("rules") have been breached. Transparency on rules enforcement means providing public access and data on: • Rules: content of terms and conditions, reporting pathways, and content moderation policies / practises / tools; and • Enforcement: on how terms and conditions/community standards have been breached, which rules are applied (including prioritisation and criteria), how and when. This data should include both overall violation rates of rules and by particular topics (e.g. COVID vaccine misinformation). Currently, in most countries transparency reports on content moderation and design choices are provided by technology companies on a voluntary basis. The UK Government noted that these voluntary reports (where they exist): "... often provide limited detail across important areas including content policies, content moderation processes, the role of algorithms in moderation and design choices, and the impact of content decisions." In addition, a common issue that we have experience of through our work is that, while companies may release a transparency report that states the total number of individual pieces of content related to a specific policy that has been removed or otherwise moderated, there is no data provided on what proportion of that type of content it comprised. The extent of this disparity between what was stated and what was known was also evidenced by Facebook Whistleblower, Frances Haugen, who advised that internal estimates were that Facebook may action as little as 3-5% of hate and about 6/10 of 1% of violence and incitement content on Facebook. On transparency of economics, this should include greater transparency over adverts: specifically, understanding where, when, by whom, and using which data.

	One option for achieving this is to require advertisers to publicly declare, on their websites, the domains on which their adverts appear. This creates a driver for corporate accountability, i.e. that consumers' money is not being funnelled to content that fundamentally harms individuals, communities and society. This type of information is often provided to advertisers by brokers, some of which are updated in real time. This requirement would simply ensure that advertisers disclose the URLs of the pages on which their adverts appear—but not other information, such as performance data or targeting criterion. It wouldn't create a duty for advertising organisations to conduct costly studies—but by making these URLs publicly available, it will make it easier for researchers, journalists, authorities and the public to instantly access the relevant information. This creates an accountability ecosystem of enabling legislation, transparent corporate behaviour and civil society/ other companies doing the checking. There are organisations such as GDI and NewsGuard that can provide the "checklist" for advertisers. CCDH's Stop Funding Misinformation has a much shorter and much more focused "Blacklist".
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	*N/A*

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