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* Since 1913, ADL's mission has been to "stop the defamation of the Jewish people and to secure justice and fair treatment to all." Dedicated to combating antisemitism, prejudice, and bigotry of all kinds, as well as defending democratic ideals and promoting civil rights, ADL is a leading voice in fighting hate in all forms, including online. ADL has gained particular experience in this space since we launched our Center for Technology and Society (CTS) in 2017. CTS leads the global fight against online hate and harassment. In a world riddled with antisemitism, bigotry, extremism, and disinformation, CTS acts as a fierce advocate for making digital spaces safe, respectful, and equitable for all people.
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**	*Is this response confidential? – N* ADL's response will provide evidence as it pertains to user-to-user services rather than search. Our response will focus on evidence relating to the presence or quantity of illegal content including but not

a criminal offence and will be reported to the police.	limited to unlawful harassment on (1) mainstream social media and (2) online multiplayer games. **I. Illegal content, especially harassment, is common on mainstream social media platforms.** Despite tech companies' public commitments to improving safety on their platform, online harassment is still common. ADL's 2022 Online Hate and Harassment survey revealed that 2 in 5 Americans (40%) experienced some type of online harassment in the course of their lives, with 1 in 10 (12%) having experienced severe types of harassment—defined as physical threats, sustained harassment, stalking, sexual harassment, doxing, and/or swatting, severe harassment of some kind—in the past 12 months. Data from the same survey also shows that marginalized or minoritized identity groups—including Jews, women, people of color, and LGBTQ+ people—experience hate-based online harassment (i.e., targeted attacks or abuse of marginalized people because of their race, ethnicity, religion, gender, sexuality, physical appearance, gender, identity, or disability) at disproportionately high levels. According to the study, 65% of people from marginalized groups who experienced online harassment reported being targeted for an aspect of their identity, compared to 38% of people from non-marginalized

groups. Moreover, the incidence of severe harassment was generally higher, as well as online stalking (12% vs. 6%) and sexual harassment (12% vs. 5%), regardless of the reason.

In addition, the study reveals identity-specific differences in the incidence of harassment, its growth trend, and the type of abuse endured. In particular:

- LGBTQ+ people are more likely than any other marginalized group to experience harassment. 66% of LGBTQ+ users surveyed experienced harassment compared to 38% of non-LGBTQ+, with 1 in 2 (53%) attributing the targeting to their sexual orientation.
- Asian Americans also reported the most significant increase in online harassment in the last two years (from 11% in 2020 to 39% in 2022), tracking closely with the rise in anti-Asian incidents offline. Furthermore, 62% attributed the harassment to their physical appearance and 53% to their race or ethnicity, compared to 34% and 23% of non-Asian Americans.
- Women were more than twice as likely to report ever experiencing sexual harassment online as men were (14% vs. 5%), with 2 in 5 attributing the harassment to their gender (vs. 1 in 7 of men). The intersectionality matrix also seems to be at play, with 81% of non-white women attributing being harassed to aspects of their

	identity (vs. 61% of white women). ● Although Jewish respondents experienced online harassment at similar rates as non-Jews, they were more likely to attribute harassment to their religion (37% vs. 14%). An overwhelming majority of respondents who experienced harassment said that the abuse happened on Facebook (68%), with Instagram, Twitter, and YouTube following far behind (26%, 23%, and 20%, respectively). Similar trends were also observed in the last 12 months. Notably, Facebook's primacy still holds when comparing the proportion of platform users to the proportion of those who reported harassment on the platform. To a large extent, Facebook's role in enabling and amplifying online harassment can be explained by a business model that optimizes for user engagement and the company's overreliance on algorithmic AI/ML systems to moderate content. First, AI and ML-based tools deployed during the moderation stage are not fit in situations where there is a need to assess context and make subjective decisions, allowing much harmful content to go undetected. Second, as hateful, harassing content often has high engagement rates, once it evades detection from content moderation systems, it is spread and amplified by platforms' ranking and recommendation algorithms faster than other types of content.

II. Harassment experienced by gamers in online multiplayer games is also an increasingly worrisome phenomenon.

Online multiplayer games are a subset of the overarching video game market. Although gamers in Britain still prefer single-player gaming, 21% of British gamers favor online multiplayer games. Furthermore, this share progressively increases with the number of hours spent playing video games each week, reaching an astounding 37% for those who play for more than 21 hours.

For the third consecutive year, an ADL survey on American gamers[6] found that harassment experienced by adult gamers is both alarmingly high and on the rise. Not only five out of six adults (83%) ages 18-45 experienced harassment in online multiplayer games, but 71% experienced severe abuse, including physical threats, stalking, and sustained harassment.

Furthermore, the same survey shows that the largest increases in identity-based harassment occurred among respondents who identified as women (49% in 2021, compared to 41% in 2020), Black or African American (42% in 2021, compared to 31% in 2020), and Asian American (38% in 2021, compared to 26% in 2020). Although LGBTQ+ players did not experience a significant rise in the amount of harassment experiences (38% in 2021 versus 37% in 2020),

	the share of LGBTQ+ respondents experiencing harassment on online multiplayer games is still of concern. Although the above research is U.S. and not U.K. based, it is important to acknowledge that it can be indicative of larger global trends and should be considered during this phase of online safety regulation.
Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content presented by your service?	*No response.*
Question 4: What are your governance, accountability and decision-making structures for user and platform safety?	*No response.*
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* **Providers of online services can, at a minimum, adhere to and build off of efforts contained within California Assembly Bill 587.** Earlier this month, California lawmakers passed a bill out of the legislature, Assembly Bill 587, requiring covered social-media companies to publicly disclose their corporate policies regarding online hate, racism, disinformation, extremism, harassment, and foreign political interference, as well as key metrics and data around the enforcement of those policies. The bill is currently awaiting signature on

	the Governor's desk. Notably, the bill, which ADL sponsored, provides a blueprint for clear and accessible "terms of service" (ToS) and transparency reports. Specifically, this transparency legislation requires covered social media companies to post their ToS in a manner reasonably designed to inform all users of the existence and contents of the terms of service. It also requires the ToS to be available in all languages in which the social media company offers product features, including but not limited to menus and prompts. These ToS would also require (1) a way to contact the social media company to ask questions about the ToS, (2) a description of how users can flag content, groups, or other users that they believe violate the ToS, and "the social media company's commitments on response and resolution time," and (3) "a list of potential actions the social media company may take against any item of content, or a user, or group of users, including, but not limited to, removal, demonetization, deprioritization, or banning." These requirements begin to shift the treatment of users from products to consumers. Lastly, the bill requires covered social media companies to submit to the Attorney General two semiannual ToS reports containing information related to content moderation policies and data related to the application of those policies in practice. These

	reports, which will be made available to the public in a central location, would include: (1) the current ToS of the social media company and a complete and detailed description of any changes since the last report, (2) a complete and detailed description of content moderation practices used by the social media company, including any rules or guidelines regarding automated content moderation systems, and (3) information–deidentified and disaggregated–on content that was flagged by the social media company as content belonging to the predefined categories mentioned above (e.g., hate speech, extremism, harassment), including the total number of flagged items of content; the total number of actioned items of content; the number of times actioned items of content were viewed by users, shared, and the number of users that viewed that content before it was actioned; the number of times users appealed social media company actions and reversals of those actions on appeal.
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?	*No response.*
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* **Online services, especially user-to-user services can adopt anti-hate by design patterns to enhance transparency, accessibility and ease**

	of use of reporting and complaint mechanisms. ADL developed a Social Pattern Library, or a collection of design patterns and principles for mitigating the presence and spread of online hate and harassment in social platforms. Of the 32 patterns, eight deal with various aspects of enhancing reporting requirements or complaint systems. These patterns include: batch comment reporting option, batch content report option, comment report, hate severity report option, livestream streamer report, livestream viewer report, targeted characteristics report option, and voice chat report. Two of the eight patterns enhance the transparency of reporting and complaint systems. They include hate severity report option and targeted characteristics report option. Six of the eight patterns enhance accessibility and ease of use of reporting and complaint systems. They include batch comment reporting option, batch content report option, comment report, livestream streamer report, livestream viewer report, and voice chat report.
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?	*Nor response.*

Question 9: If your service has a *complaints* mechanism in place, how are these processes designed and maintained?	*No response.*
Question 10: What action does your service take in response to *reports* or *complaints*?	*No response.*
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* To improve content moderation to better protect users, platforms must enforce existing rules equitably and at scale. Inconsistent, as well as unfair enforcement of ToS undermine any effectiveness they might offer. Platforms should also hire additional content moderators and train them well. Additionally, ADL research underscores the need for platforms to invest resources to better train automated detection systems to better enforce policies. Specifically, automated detection systems should be trained with known samples of violent extremist speech and data from extremist sites. Furthermore, platforms should include specific linguistic markers in detection algorithms and deemphasize profanity in toxicity detection to bolster white supremacy detection approaches. In general, platforms and products should be designed with the experiences of communities targeted

	by hate in mind—employing an anti-hate-by-design model—like the Online Hate Index, ADL's antisemitism machine-learning classifier, which models this approach. Of course, increased transparency to ensure these practices are being implemented is crucial to increase the trust and safety of users.
Question 12: What automated moderation systems do you have in place around illegal content?	*No response.*
Question 13: How do you use human moderators to identify and assess illegal content?	*No response.*
Question 14: How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services?	*Is this response confidential? – N* This response will focus on the safeguards that should be in place to generally protect users' privacy. Another pattern from the aforementioned "Social Pattern Library" addresses user privacy and encourages account privacy settings to mitigate the presence and spread of online hate and harassment in social platforms. Privacy protections help to eliminate network or campaign harassment before it begins by allowing users to hide their accounts and the information and content posted on them, making it difficult for harassers to identify them as a specific demographic for targeting. Account privacy settings can include

	options such as: public, followers only, or followers of followers. The setting should be applied to all posts by default, unless specific privacy settings are applied to individual posts.
Question 15: In what instances is illegal content removed from your service?	*No response.*
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	*No response.*
Question 17: What other sanctions or disincentives do you employ against users who post illegal content?	*No response.*
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* **Platforms must take an anti-hate by design approach to prevent harm.** Such an approach would help to address aspects of design that could prevent harm, in some cases allowing users to control what they encounter and even restrict functionality. Of the 32 patterns in ADL's "Social Pattern Library" 23 additional patterns not yet discussed that work to prevent harm should be enhanced and deployed more widely by industry. These patterns include: block user, comment filter setting, feed filter setting, flagged link post interstitial,

	flagged link reshare interstitial, hateful comment interstitial, hateful post interstitial, hateful post reshare interstitial, hide comment option, livestream broadcast delay, livestream comment rules, livestream commenting rules violation, livestream ended notification, livestream rejoin prompt, livestream streaming rules violation, personal bubble setting, posting content and comments, read article prompt, user muting option, voice chat muted user indicator, voice chat muted/removed notification, voice chat rules, voice chat rules violation.
Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact of harm from illegal content?	*No response.*
Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	*No response.*
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?	*No response.*
Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?	*No response.*

Question 23: Can you identify factors which might indicate that a service is likely to attract child users?	*No response.*
Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?	*No response.*
Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?	*No response.*
Question 26: What information do you have about the age of your users?	*No response.*
Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	*Is this response confidential? – N* **Transparency is key to addressing harmful and illegal content, including unlawful hateful and harassing content, on digital platforms.** The ability to independently quantify hateful and harassing content at scale and compare results between and among different platforms is necessary to assess whether platform policy or product changes, or other interventions reduce their spread. Without an adequate level of transparency, lawmakers and watchdogs lack the understanding and evidence to regulate social media

	platforms and hold them accountable. For user-to-user mainstream social media services, transparency reports should show data from user-generated, identity-based reporting. For example, if users report they were targeted because they were Jewish, such data can be aggregated to become a subjective measure of the scale and nature of antisemitic content on a platform, a useful metric to researchers. As previously mentioned, California's social media transparency bill (AB 587) provides a blueprint for not only clear and accessible Terms of Service but also transparency report guidelines. Part of the bill requires covered social media companies to submit to the Attorney General two semiannual reports containing information related to content moderation policies and data related to the application of those policies in practice. Specifically, the data includes information–deidentified and disaggregated–on content that was flagged by the social media company as content belonging to the predefined categories (online hate/racism, disinformation, extremism, harassment, and foreign political interference), including: the total number of flagged items of content; the total number of actioned items of content; the number of times actioned items of content were viewed by users, shared, and the number of users that viewed that content before it was actioned; the

	number of times users appealed social media company actions and reversals of those actions on appeal. There is also evidence to support the notion that transparency reports should include standardized reporting categories. For example, a recent ADL report that investigated how hate and harassment manifest in Minecraft recommended industry-wide standardization of content moderation reporting categories to better understand the frequency and nature of hate in online spaces. Such reporting would include defined categories and violating offenses with clear descriptions. Standardization like this would help facilitate future research, particularly in regard to documenting how moderation actions change user behavior over time. One such example that could be used as the foundation of this effort is ADL and Fair Play Alliance's Disruption and Harms in Online Gaming Framework.
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	*No response.*