Your response

Please refer to the sub-questions or prompts in the annex of 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 (delete as appropriate)* The Antisemitism Policy Trust is a charity, funded by philanthropic donations, that works to educate and empower parliamentarians and policy makers to address antisemitism. For more than ten years, the Trust has provided the secretariat to the All-Party Parliamentary Group (APPG) Against Antisemitism. The Trust has advised the government, opposition parties, policy makers, civil servants, and regulators, including Ofcom, on policies relating to antisemitism, hate crime and online abuse. We have produced briefings and provided written and oral evidence about online safety, especially with regards to online antisemitism, both in relation to illegal and legal but harmful content. The Trust Chief Executive, Danny Stone, gave evidence to Parliamentary Committees scrutinizing the Bill, including the Joint Committee on the Draft Online Safety Bill, the Petitions Committee, and the Public Bill Committee. Our recommendations were adopted or advanced by some of these bodies, and by the DCMS Select Committee.
Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services?	*Is this response confidential? – N (delete as appropriate)* Social media platforms host large volumes of antisemitic content. This includes both legal and illegal material that can incite hate crime, violence and terrorism against Jewish targets in the UK and

internationally, but can also serve to frighten, intimidate or seek to exclude Jewish people and anti-racists from online spaces. The Community Security Trust's (CST) recording methods mean that it only details incidents where the victim or perpetrator is based in the UK, and the incident must be pro-actively sent to the organisation. If it were recording all online antisemitism, the number would be almost unmeasurable. For 2021, the majority of incidents recorded by CST occurred on mainstream platforms such as Twitter and Facebook. Most of the incidents (509 out of 552 in 2021), fall under the category of Abusive Behaviour. Forty incidents include harassment and threats. CST stated in its 2021 annual report that 'social media has been used as a tool for coordinated campaigns of antisemitic harassment, threats and abuse directed at Jewish public figures and other individuals.'

CST publishes incidents reported to it by the general public, most of whom only use mainstream platforms. However, we are deeply concerned about small and more extreme platforms such as Bitchute, 4Chan and 8Chan, that host both extreme content and illegal content that radicalises people and inspires violence against Jews.

A briefing by the Antisemitism Policy Trust, published in August 2020, provided examples of the connection between online and offline harms, citing examples of attacks against Jewish targets (for example, the Pittsburgh Synagogue attack) and against Muslim targets (for example, the Finsbury Park and Christchurch mosque attacks). In all of these, attackers participated in extreme online forums where they were either radicalised to the point of attacking Jews and Muslims, or inspired others to commit acts of hateful violence. The terrorist who killed eleven congregants and injured six others in a synagogue in Pittsburgh in 2018, promoted his hateful, antisemitic agenda on the social media platform Gab – where he also posted minutes before attacking the synagogue. Based on testimonies after the attack, he had consumed large amounts of racist and other material online which had incited him to violence. Another briefing by the Antisemitism Policy Trust, on anti-Jewish misogyny, referenced a study by the American NGO Media Matters, which found a

staggering 180% increase in posts containing both antisemitism and misogyny on the far-right anonymous message board 4chan between 2015 and 2017.

Illegal content also involves harassment of Jews, who are targeted only because they are Jewish. This includes threats of violence, rape and death against public figures and members of the public. For example, Dame Margaret Hodge MP and former MP Luciana Berger have been targeted by antisemites online, receiving illegal abuse that included death threats. One of the men who harassed Ms. Berger with antisemitic and misogynistic abuse that made her concerned for her safety, received a two-year jail sentence. Research into online antisemitism by the Antisemitism Policy Trust, in collaboration with the Community Security Trust (CST) and the Woolf Institute, found that antisemitism flourishes on Instagram, with antisemitic hashtags often associated with conspiracy theories. Such hashtags are sometimes attached to posts that have no direct relationships to the content of the post, meaning that they are displayed to a large pool of users, who in most cases are not actively looking for antisemitic content, but are exposed to it and to its harmful influence. A case of antisemitic supply rather than demand. Our research concluded that Instagram requires improved algorithmic filtering to address conspiracies and antisemitic hashtags, and that it needs to improve its community standards against hateful content. A recent report by the Community Security Trust (CST) about antisemitism and the Covid conspiracy movement, also found a link between antisemitism and conspiracies. Its report showed that conspiracy theorists, especially those on the far right and far left, regularly incorporate antisemitic conspiracy theories and tropes into the messages. This includes blaming Jews for creating the virus in order to control populations, for benefiting from the pandemic, and for creating dangerous vaccines. Examples of these types of false charges were found on Instagram, Facebook, Twitter and also on smaller platforms. Some of the content explicitly called for violence against Jews.

Another study by the Antisemitism Policy Trust, in collaboration with the Community Security Trust (CST) and the Woolf Institute, found that Twitter

hosts vast amounts of antisemitic content. Our researchers estimated that there are up to 1,350 explicitly antisemitic tweets in English posted and available to UK users every day. This amounts to nearly half a million explicitly antisemitic tweets a year – two tweets for every Jewish person in the UK- giving the content a very wide potential reach =. Some of this content would likely cross the threshold . Of course, each tweet must be assessed and there is a vast amount of material. To give a sense of some of the content, one tweet mentioned in the study calls for denial of 'equality, voting rights' for Jews, as well as a call to 'take their land and demolish their homes … and ghettoisation'. Another tweet read: 'Little Jews… I long for your suffering. Which is coming.'

As a recent report by the Community Security Trust (CST) found, far-right extremists have migrated from the large, mainstream social media platform into smaller platforms, such as Gab, Bitchute and Telegram, where antisemitic content flourishes. This includes illegal material that calls for violence and terrorism against Jews. On Telegram for example, CST found posts glorifying far right terrorists including Thomas Mair and David Copeland, and calls to kill Jews. The platform 4chan was also found to host threads containing explicit calls to kill Jews. Similar posts, containing violent, antisemitic comments, were found on Bitchute, including images with phrases including: 'all Jews must die', and images of people aiming weapons accompanied by threatening language. The information is easily accessible to anyone, and CST concluded that 'the quantity and spread of this incitement poses an urgent and ongoing terror threat to Jewish communities.' The Trust has made clear its position on the categorization of small platforms but whilst, at the time of writing, they fall outside the scope of Category 1 as invisaged in the Online Safety Bill, Ofcom will need to look at the illegal materials to ensure they are removed, but should also undertake the research Government has said is necessary before bringing in the relevant categorization methods, the team will find a large volume of harmful antisemitic material accompanying the illegal content.

In addition to CST's findings in relation to

Telegram, a report published by Hope not Hate, Amadeu Antonio Stiftung and Expo in October 2021, Antisemitism in the Digital Age, found considerable presence of antisemitic and other racist and extremist content on every platform explored in the research. Telegram hosted some of the most extreme content due to its lack of moderation. This included 120 groups sharing and glorifying the Christchurch shooter's manifesto, and a channel promoting the antisemitic New World Order conspiracy (which alleges, among other conspiracies, that Jews are plotting to rule the world). That group grew by 90,000 users from its launch in 2021. This research furthers that from a Hope not Hate report which found that Telegram hosts Nazi channels spreading antisemitic white supremacist propaganda. Some of these channels provided training and instructions for guerrilla operations, use of weaponary and real world attacks. Hope not Hate concluded that conspiracy theories, including many antisemitic ones, are used to radicalise people into far right doctrines, and 'can motivate disruption and violence.'

The Anti-defamation League (ADL) published an Online Holocaust Denial Report Card, probing Holocaust denial on large platforms, including Facebook/Instagram, YouTube, Twitter, TikTok and Reddit. It investigated a range of factors, including which companies had an explicit Holocaust denial policy, how effective their efforts were in addressing Holocaust denial, and the actions taken against Holocaust denial. The results were grim: Holocaust denial was prevalent and easily accessible on most platforms; Out of nine platforms, only two took action against Holocaust denial content reported to them, even though five platforms have an explicit Holocaust denial policy and all have a general hate policy; Five out of nine did not have effective product level efforts to address Holocaust denial.

Hope Not Hate, in its report Antisemitism in the Digital Age, also found that Holocaust denial (as well as celebrating the Holocaust, diminishing it or mocking it) is prevalent online, especially in far right spaces, such as Reddit and 4chan. Holocaust denial is just one form of antisemitism that exists in these spaces alongside calls for terror attacks against Jews. Holocaust denial was also found to be present in far-left groups and within some

online Muslim communities. According to the authors of the report, this brings ‘a normalisation of extreme antisemitism, especially with regard to the Holocaust.’

Although Holocaust denial is not illegal in the UK, such content can be linked to more extreme and illegal content that incites violence and threats against Jews. In addition, despite not being illegal, in 2018 Holocaust denial was prosecuted under the Communications Act 2003 in the case of blogger Alison Chablos. Chablos wrote songs that mocked the Holocaust and was convicted of three charges, including sending offensive messages through public communications. She was sentenced to twenty weeks’ imprisonment. This form of denial has also previously been prosecuted under public order legislation. Holocaust denial – which is commonly found on both large and small platforms – should therefore be taken seriously, as it may be illegal in the UK under some circumstances.

Search services have also been found, through their systems, to direct people to hate materialsand racist content that is legal, but can easily direct users to more extreme and illegal content when they follow search prompts. Google’s autocomplete algorithm has been found to suggest antisemitic, racist and sexist content to users; typing the word ‘are Jews’ in the Google search bar, prompted an autocomplete suggestion ’…. evil?’. This produced results that demonise or incite people to hate Jews.Our own research into Google found that its lauded ‘SafeSearch’ option produced as many antisemitic results as its regular search. For example, when searching for the term ‘Jew jokes’ with the SafeSearch option disabled, 48% of the results produced by Google were found to be antisemitic – a high proportion in of itself. However, the same search phrase with the SafeSearch option enabled, produced an even greater proportion of antisemitic results – 57%. This places at risk children and other vulnerable people, who wrongly assume that they are protected by Google’s SafeSearch option.

As Andrew Percy made clear in his speech during the Report Stage of the Online Safety Bill, a Google search for the seemingly innocent words ‘desk ornaments’ has produced top search results

that included swastikas, SS bolts and other Nazi memorabilia. Amazon's Alexa produced an antisemitic conspiracy theory in response to a search. It suggested – based on a single comment posted on Amazon's website – that the Jewish American-Hungarian philanthropist George Soros is responsible for all of the world's evils – a common trope based on antisemitic conspiracies. This is information that could reach millions of users around the world. Microsoft Bing was found to direct users to hateful searches with the autocomplete "Jews are bastards" and Google's image carousel highlighted pictures of portable barbecues to those searching Jewish baby stroller.

Antisemitic search results come up in other languages too. Last year the Trust found that asking Siri, in Spanish, "do the Jews control the media?" prompted a response of articles including details of "Jewish control international media" and an article arguing that "A world famous sociologist claims that the Jews control the media".

Despite the risk of exposure to harmful but legal content on large search platforms and on voice search assistants, as outlined above, which can easily lead users to increasingly extreme and even illegal content.

As will be clear from the material above, antisemitism (including Holocaust denial, as we have set out) sits on the boundary at the threshold of what is legally permissible. We have sought to outline the scale of antisemitism online, and indicate where we have seen, and where we believe the problems are greatest. Judging the extent to which illegal materials are online and being delivered to users through systems will be the job of Ofcom, Starting any investigation with legal harms, in relation to antisemitism, we are confident albeit regret that illegal harm will follow, and accompany the other racist bile you will find easily discoverable online.

Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content

Is this response confidential? – Y / N (delete as appropriate)

presented by your service?

Question 4: What are your governance, accountability and decision-making structures for user and platform safety?

Is this response confidential? – Y / N (delete as appropriate)

Question 5: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?

Is this response confidential? – N (delete as appropriate)

Terms of service should be presented in a clear, visible, and concise way, understandable by all likely users, and ergo in a selection of languages. These should include an explanation of what is not allowed on a platform, and the expected repercussions for those violating the terms and conditions, along with a clear and accessible information on how decisions are made, the timeframe in which they are made, and the rights of users to challenge those decisions. Platforms should also make clear their policy regarding cooperating with law enforcement on illegal content published by users, such as death threats.

As will be suggested in more detail in the answer to Q11, encouragement of friction in systems might well be encouraged. For example, use of artificial intelligence (AI) to identify phrases that suggest content is in violation of a platform's Terms, and which automatically directs users to those Terms (preferably to the specific condition the content may apply to) –even before the content is published might be a helpful intervention. This would allow users to quickly assess whether they may risk violating the Terms and change the content of the information they are about to make public. Some form of explanatory video or other material on how enforcement decisions are taken might also be advisable, especially as not everyone will take the time to read details terms. When users on Ebay are believed to have abused its terms, in respect of so-called 'shill bidding', for example, they are forced to do a 'test' to ensure they have understood the platform's terms before re-gaining access to their account. Something akin to this could be a strong incentive to users to understand

and abide by Terms.

Some of the largest platforms, including Twitter, YouTube and Facebook, have strengthened their Terms and Conditions to include prohibiting abusive content against people who have protected characteristics. Unfortunately, enforcement has been largely patchy and ineffective. As a result, online discourse continues to be abusive and harmful. Terms of service should therefore not only be accessible, but enforced consistently and effectively by platforms. Fast action is key to effective enforcement. To that end, some level of culpability and transparency in relation to poor enforcement would be welcome, as this will highlight to users a platforms commitment to getting it right, in respect of its own Terms.

The Antisemitism Policy Trust developed a draft Code of Practice on Hate Crime and Wider Harms which was then adopted, improved and published by the Carnegie Trust. In that document, specific reference was made to the importance of Community Guidelines, and the recommendation included that companies must explain their policies (and how these are developed, enforced and reviewed, plus the role of victims' groups and civil society in developing them) on harmful content, including what activity and material constitutes hateful content, including that which is a hate crime, or where not necessarily illegal, content that may directly or indirectly cause harm to others. A specific list of harms was included in the document.

Question 6: How do your terms of service or public policy statements treat illegal content? How are these terms of service maintained and how much resource is dedicated to this?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 7: What can providers of online services do to enhance the transparency, accessibility, ease of	*Is this response confidential? – N (delete as appropriate)*

use and users' awareness of their reporting and complaints mechanisms?

Reporting should be easily accessible. Facebook for example, has a very accessible reporting mechanism on its platform that can be easily and quickly accessed by users.

Companies should be transparent about how complaints are handled, and provide detailed explanation about the outcomes of complaints to the complainant. Vague information, such as that content simply did not violate Terms, should be replaced with a more detailed explanation.. When a complaint is rejected, the complainant should receive information on how the decision can be challenged. The same should apply if a complaint is upheld, and the person being sanctioned wants to challenge the decision.

In the aforementioned draft Code of Practice the Trust developed, we stated that:
- Companies must have reporting processes that are fit for purpose in respect of hate crime and wider harms, that are clear, visible and easy to use and age-appropriate in design. Thought should be given to reporting avenues for non-users.
- Companies must have in place clear, transparent and effective processes to review and respond to content reported as illegal and harmful.
- Companies must have in place effective and appropriate safeguards in full respect of fundamental rights, freedom of expression and relevant data protection regulation. This includes, specifically, taking reasonable steps to ensure users will not receive recommendations to criminal, hateful or inappropriate content.

Though moves have been made to simplify reporting processes, context can be key, and so space to apply context would be helpful. For example, the Trust had cause to report a picture taken from behind of former MP Luciana Berger, and posted to the 'Redwatch' group page before it was banned from Facebook. This was far-right stalking of a high profile Jewish individual. Without the explanatory context, the photo was deemed not to have breached the platform's rules but on contacting staff, with the explanation, it was removed. Context can help aid decision making and save time for everyone. Having the option to add it would be helpful, and Ofcom might make

such a suggestion in a code of practice.

Question 8: If your service has *reporting or flagging* mechanisms in place for illegal content, or users who post illegal content, how are these processes designed and maintained?

Is this response confidential? – Y / N (delete as appropriate)

Question 9: If your service has a *complaints* mechanism in place, how are these processes designed and maintained?

Is this response confidential? – Y / N (delete as appropriate)

Question 10: What action does your service take in response to *reports* or *complaints*?

Is this response confidential? – Y / N (delete as appropriate)

Question 11: Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?

Is this response confidential? – N (delete as appropriate)

While the Trust believes that a focus on systems will be more effective than individualised content moderation, human and artificial intelligence (AI) moderation will continue to play an important role in removing illegal harmful material.

Effective moderation can be achieved by providing clarity of what constitutes a priority harm, ensuring moderators are sufficiently trained to recognise these, and act on such content quickly before it can have a wide reach. We are, and have long been particularly concerned about the quality assurance of moderator training. We believe that platforms should have greater transparency about moderator training, so that it can be assessed and available for independent scrutiny. Running training for one of the largest global technology companies which runs a search engine amongst many other platforms, we were told that training we ran on antisemitism had directly impacted some of its efforts, and we know that input we have given to TikTok, Microsoft, Twitter and others has helped improved and extend some of the systems they have for moderation of

antisemitic content. There is little point in having training on antisemitism developed in-house by tech companies, or outsourced to generalised 'training experts', proper independent scrutiny would be welcome and will ensure antisemitism is more effectively moderated, and not over-moderated, especially when it comes to difficult decisions on – for example – the Middle East conflict vis-à-vis anti-Jewish tropes.

More investment should also be put into AI to improve its capabilities and aid human moderators in reviewing large volumes of content. Smaller platforms do not enjoy the same funding for developing AI as large social media platforms and search engines, and more thought needs to be given as to how to help those platform have better access to advanced AI, for example with Ofcom facilitating the sharing of learning which does not compromise a companies competitive edge.

The risk assessment processes Ofcom is due to initiate should also help in this area, for example, the Antisemitism Policy Trust report on Google SafeSearch, referenced in an earlier question, outlined that at the back end, Google's categorization of images included terms like 'spoof', 'racey' and so on. Whilst we appreciate that there is a balance to be achieved between complexity and simplicity when it comes to machine learning, any risk assessment would highlight that such simplicity will lend itself to racist content filtering through systems and so additional thought could and should be given to better tagging methods.

Specifically, when it comes to illegal antisemitic harassment and incitement, moderators should be made aware of the different antisemitic code words used, especially by those on the far-right. Phrases such as 'Holocaogh', which is a call by the far-right to spread coronavirus among Jews in order to kill them, or 'ZOG' – Zionist Occupying Government, which claims that governments are controlled by Jews, and (((echo))) - a symbol used to highlight names of Jewish people or organisations owned (or perceived to be owned) by Jews, that has been used to incite harassment and death threats against them. To achieve this, platforms, as we stated above, would ne wise to use third sector experts to help.

Question 12: What automated moderation systems do you have in place around illegal content?

Is this response confidential? – Y / N (delete as appropriate)

Question 13: How do you use human moderators to identify and assess illegal content?

Is this response confidential? – Y / N (delete as appropriate)

Question 14: How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services?

Is this response confidential? – N (delete as appropriate)

We have discussed with Google some of the issues pertaining to this very concern. For example, when searching for 'Happy Merchant' or 'Greedy Merchant' on Google images, a great deal of racist material is returned. Similarly concerns arise in relation to anti-black racist materials and homophobic content. At present, Google is trialing a warning to appear above relevant searches. We welcome any friction that can be introduced to systems, and educational or warning material that is introduced, particularly as it might impact children's use of a service. So far as sanctions are concerned, we have seen there been poorly applied.

A high profile example is that of Richard Kylea Cowie, aka Wiley, a notorious rap artist who went on an antisemitic rant to his 950,000 followers on social media platforms, including Twitter and Instagram,. The platforms were slow to act decisively. A report from the Community Security Trust (CST) about this incident showed that on Twitter alone, Wiley's antisemitic comments received roughly 355,813 impressions for each minute that Twitter failed to act, exposing this content to a huge audience. Similarly, his videos on Instagram were viewed a total of 1,441,028 times before Instagram finally banned him from the platform a day later. The suspension should

	have kicked in far earlier for an account that was repeatedly breaking platform Terms. We have major concerns about the ease of access to service for banned accounts, and the lack of apparent monitoring for similar usernames linked to banned accounts. It would be helpful if platforms had a specific reporting option to send details of accounts suspected to be spawns
Question 15: In what instances is illegal content removed from your service?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 17: What other sanctions or disincentives do you employ against users who post illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 18: Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry?	*Is this response confidential? – N (delete as appropriate)* We have heard from companies like Public.io and Bertie Vigden at the Turing Institute (and on his own behalf) about work they are doing towards improving efficiency and developing tools in this area, they would be best placed to advise on this question and should they not respond to Ofcom's call for evidence, we would strongly suggest that contact is made with them.
Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact of harm from illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*

Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 21: How do you mitigate any risks posed by the design of algorithms that support the function of your service (e.g. search engines, or social and content recommender systems), with reference to illegal content specifically?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 23: Can you identify factors which might indicate that a service is likely to attract child users?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 26: What information do you have about the age of your users?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 27: For purposes of transparency, what	*Is this response confidential? – N (delete as*

type of information is useful/not useful? Why?	*appropriate)* It is important that Ofcom liaise with regulated entities in relation to potential 'gaming' of their rules by bad actors, but perhaps more important is the sharing of knowledge about some gaming when it occurs. Too often we have found platforms act in a silo in this regard.
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*

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