OFCOM CONSULTATION:

PROTECTING PEOPLE FROM ILLEGAL HARMS ONLINE

Institute for Strategic Dialogue (ISD): Consultation Response - February 2023

Overview:

ISD welcomes the opportunity to respond to this important consultation on illegal harms, and commends Ofcom's openness and engagement during the development of the proposals and throughout the consultation process to outline the rationale for the approach taken and respond to questions from civil society stakeholders. We also recognise the challenges the complexity of the Online Safety Act brings, especially in the context of the collective desire from the Government, Ofcom and other key stakeholders to quickly and effectively implement and enforce the legislation and improve online safety in the UK.

However, as outlined in this response, ISD shares the key overarching concerns expressed in the collective Online Safety Act Network statement, as well as the feedback provided on the Codes of Practice. This includes the risk that the proposed approach could serve to entrench existing industry practices and expectations, especially for the largest services, rather than require a more ambitious and proactive approach to improving online safety. We also have concerns that the proposed thresholds, requirements and measures for smaller platforms are too prescriptive and risk a 'one-size-fits-all' approach to many of the smaller but severely high-risk platforms ISD encounters in our work on online terrorism, extremism and hate.

Volume 2: The causes and impacts of online harm

Question 6.1: Do you have any comments on Ofcom's assessment of the causes and impacts of online harms? Do you think we have missed anything important in our analysis? Please provide evidence to support your answer.

The overall assessment of a wide range of online harms within the scope of the Act demonstrates considerable depth. However, potential gaps in the assessment process include considerations related to the business models of services and their intersections with different harms. Separating functionalities from the overall business models of many services

is challenging, as most features that are not safety-specific are primarily designed to maximise user engagement and time spent.

This interconnectedness impacts various forms of harm, including terrorism. For instance, account recommendation systems designed to grow users' networks can lead to resilient terrorist networks of accounts, making individual account-level takedowns less effective. This, combined with the ease of setting up new accounts, allows users to quickly create new accounts and rejoin the network. Similarly, in the context of online abuse, harassment and hate speech issues arise with content recommendation systems, such as misogynistic content on YouTube, which may direct users towards increasingly hateful or inciting content. Trending terms and hashtags may also be manipulated to encourage targeted hate.

There is a danger that these potential gaps could lead to superficial risk assessments, undermining the systems-focused intent of the Act, as outlined by the government and during Parliamentary debates. Addressing these gaps is crucial for ensuring a comprehensive and effective regulatory framework.

Additionally, there remain significant gaps in independent understanding of risk areas between different sizes and types of services. Whereas harmful activity has been relatively well documented and analysed across major social media platforms, gaming platforms have typically been under-analysed (in part due to challenges in accessing data). The same applies to a range of emerging platforms, and new technologies increasingly being incorporated into major services (such as generative AI or alternative- or virtual-reality). Accordingly, it is essential that risk assessments on these services are comprehensive.

When considering risk factors associated with hate and harassment, Ofcom notes that "*different types of services can contribute to this risk factor. Social media services and online gaming services pose a particular risk of hate offences, and have therefore been included in the risk profiles. Video-sharing services and private messaging services have also been identified as spaces that can be used to commit or facilitate offences related to hate, targeting minorities and other protected groups*." It should be noted that ISD research has found the widespread use of hate speech targeting Muslims and Jews in YouTube comment sections, and identified the central role that video content plays in activating online hate speech. Accordingly, there is a case to be made for counting video-sharing services as having a comparable magnitude of risk to social media services and gaming services.

The hate offences laid out by Ofcom note race and ethnicity, religion and sexual orientation as risk factors in exacerbating users' experiences of hateful content. And Ofcom notes "a user's experience of hateful content may also be influenced by their age and gender." We would recommend expanding the scope of this language to incorporate other protected characteristics than can impact user experiences, including disability and transgender identity.

Question 6.2: Do you have any views about our interpretation of the links between risk factors and different kinds of illegal harm? Please provide evidence to support your answer.

The UK's Independent Reviewer of Terrorism Legislation referenced in their submission evidence suggesting that young internet users are at risk of radicalization online, interacting with a wide range of communities, some of which ISD noted in evidence provided to the Intelligence and Security Committee of Parliament which are themselves youth led. This includes ISD analysis which provided evidence of minors engaging with illegal terrorist material on Discord. In some cases exposure to potentially radicalising content is itself facilitated by platform recommendation systems, as evidenced by an ISD investigation into the promotion of misogynistic content on YouTube to boys and young men. Accordingly, the risks posed to young people, and risks associated with terrorist content need to be considered in combination.

In Ofcom's (summary) assessment of relevant functionalities, ISD also has the following suggested clarifications around how terrorism offences manifest online:

- **Hyperlinking**: While this is explained in more detail at 6B.52, it should be clearer in the summary that 'hyperlinking' also refers to general outlinking on services (which, as noted, is highly relevant to terrorist activity online).
- **Recommender systems**: While outlined at 6B.69, recommender systems are missing from the functionalities summary, where it would be helpful to reference the risks of services providing those already engaging with terrorist content with more of it through recommender systems.
- **Audio sampling functionalities:** It would be important to mention the potential risk of audio functionalities (both live and otherwise) in this summary, which ISD provides concrete examples of in our research referenced below.

Additionally, Ofcom notes that "*hate offences can be committed via direct responses or comments on posted content. This functionality can in particular enable the amplification of hate, known as 'cybermobbing' and/or 'dogpiling'*". This 'cybermobbing' is itself frequently coordinated by communities of online trolls or extremists in channels on other, typically smaller platforms. In particular, ISD research has noted this activity on Discord in a number of contexts. We would recommend recognising the risks posed by these inter-platform user behaviours when discussing hate and harassment risks.

Furthermore, ISD also recommends including the following additional relevant evidence in Ofcom's assessment of how terrorism offences manifest online:

- **6B.13** – ISD research for Ofcom, which included outlink analysis of terrorist accounts, found that alt-tech platforms like Bitchute, Odysee, Gettr and Rumble were linked to more often than Facebook, Instagram or Reddit.

- **6B.14** – ISD research has repeatedly shown the significance of social media outlinks to networks of terrorist websites in the online efforts of terrorist actors.
- **6B.23** – ISD's research has shown how so called so-called 'news sites' are used by online terrorist actors as part of a suite of a moderation evasion tactics on social media platforms.
- **6B.25** – ISD research has shown the importance of gaming adjacent services in violent extremist mobilisation.
- **6B.26** – ISD's analysis of the 'Caliphate Cache' shows the importance of file sharing sites in terrorist mobilisation.
- **6B.28** – ISD's research on gaming services shows the relevant of such functionalities for terrorist mobilisation.
- **6B.43** – Beyond 'live audio', ISD research has shown the relevance of functionality for sharing snipped audio being used for terrorist content on TikTok, for example speeches promoting the ideology of Anwar al-Awlaki.
- **6B.61** – ISD has evidenced the tenacity of the cross-platform Salafi-jihadist ecosystem.
- **6B.66** – ISD research has evidenced the use of evasion tactics by terrorist actors online.
- **6B.69** – ISD research has provided evidence in the role of recommender systems in surfacing increasingly extreme and violent content, and we have also provided recommendations for how services can better mitigate these risks.

Related to these online terrorism dynamics, Ofcom notes the relationship between online hate speech and violent attacks, explicitly mentioning the 2018 Pittsburgh synagogue shooter as an example. However, it is worth noting that radicalisation to violence is not the only vector by which hate speech can inspire other offences. Of a far more significant magnitude is growth of cultures which reward or promote the spread of hate speech (e.g. forum culture associated with imageboards like 4chan). In this instance it can be observed that the growth of a culture which is built around the offences detailed here, can inspire individuals to commit more hate speech of related violent offences.

Volume 3: How should services assess the risk of online harms?

Question 8.1: Do you agree with our proposals in relation to governance and accountability measures in the illegal content Codes of Practice? Please provide underlying arguments and evidence of efficacy or risks to support your view.

ISD's experience in dealing with tech companies reveals a common trend where, from an external perspective and without access to internal structures and reporting details, Trust & Safety functions often seem to be deprioritised compared to other key functions like engineering, features design and user experience, commercial, and government affairs. Moreover, these functions often operate separately rather than being integrated with Trust

& Safety, with online safety expertise brought in late in the process of designing or developing new products or features, if considered at all.

In the case of smaller or growing companies, Trust & Safety is frequently treated as a late addition, essentially 'bolted on' rather than being embedded into the company structure and integrated throughout the development process. Best practice would dictate that online safety should be regarded as a key cross-cutting function from the outset. This is further evidenced by the prioritisation of teams or functions during company layoffs in recent years, for example at Twitter/X following Elon Musk's takeover of the platform in 2022, across Meta, Alphabet and Twitter in 2023, and from various other companies over the past several years.

Given these observations, it is crucial that the responsibility for online safety resides at a senior or leadership level internally. Moreover, it should be well-integrated across different functions and levels within the company in a holistic manner. This integrated approach would help to ensure that online safety is not an afterthought, but an intrinsic consideration throughout the development and operational processes of in-scope services.

Question 8.2: Do you agree with the types of services that we propose the governance and accountability measures should apply to?

The response to Question 8.1, emphasising the integration of online safety expertise throughout corporate structures from an early stage with clearly defined responsibilities, should be a standard practice, even for smaller companies. The Act aims to establish a regulatory regime centred on systems and processes, necessitating the creation of structures to support this objective. This emphasis on online safety should not be deprioritised relative to commercial interests or growth, even if a service is small or currently unprofitable. This approach aligns with established practices in safety regulation across various sectors such as health and safety, product safety, and financial services.

There are legitimate concerns regarding how governance and accountability apply to small high-risk services, particularly those designed with implicit or explicit intent to facilitate illegality or harm. Addressing these concerns will be crucial to ensuring that online safety measures are effectively implemented across the digital landscape, irrespective of the size or profitability of the service. Examples of such activity identified in ISD research include the proliferation of hate speech on 4chan, and the spread of violent extremist and hateful communities on smaller platforms - such as Telegram - globally, including analysis evidencing this phenomenon in the UK, Germany and Canada.

Question 8.3: Are you aware of any additional evidence of the efficacy, costs and risks associated with a potential future measure to requiring services to have measures to mitigate and manage illegal content risks audited by an independent third-party?

The failure of self-regulation within the tech industry has served as the impetus for the implementation of regulations not only in the UK but also in numerous other jurisdictions such as the EU, Australia and more recently Canada. Tech companies have demonstrated their inability to effectively design products and manage their services safety, often highlighted in the testimonies and internal evidence provided by whistleblowers.

ISD's experiences of working with tech companies, for example through multi-stakeholder initiatives like the Global Internet Forum to Counter-Terrorism (GIFCT) and the Christchurch Call, have often yielded only marginal improvements at a slow pace. A notable lack of transparency, evidenced by self-defined or self-serving metrics, intentional gaps in transparency disclosures, and evasive behaviour before Congressional or Parliamentary committees, underscores companies' reluctance to be subject to independent scrutiny and oversight. There are also numerous examples of companies restricting access to data for independent researchers, including Twitter/X revoking free access to its API and taking legal action against the Centre for Countering Digital Hate (CCDH), and Meta withdrawing support for CrowdTangle and shutting down NYU's Cybersecurity for Democracy project's access to Facebook's Ad Library. In light of these challenges, it is suggested that independent third-party audits should be mandated from the outset, particularly for the highest risk services.

Risk Assessments

Question 9.1: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

Question 9.2: Do you think the four-step risk assessment process and the Risk Profiles are useful models to help services navigate and comply with their wider obligations under the Act?

The strong emphasis on proportionality related to potential costs incurred by services, as opposed to the impacts and costs resulting from online harms, needs to be reevaluated. This is essential for ensuring comprehensive and balanced risk assessments that take into account the severity of harms for individuals and wider society. Additionally, there is an over-reliance on reactive measures, such as content moderation and user reports, compared to proactive safety-by-design approaches, such as testing products and features throughout the design process and lifecycle. Notably, smaller services often escape the requirement for these proactive approaches.

The reliance on reactive measures becomes particularly problematic as user reporting processes are often ineffective, with services frequently failing to respond appropriately. Many users refrain from using these options due to a lack of faith in their efficacy or the overwhelming volumes of illegal or harmful content encountered, compounded by the absence of bulk reporting options on many services.

Services should also not solely base their assessments on the risk profiles produced by Ofcom. Instead, they should be mandated to incorporate insights from their own internal data and understanding of user engagement, along with input from and consultation with external experts. Recognising that the tech sector has a track record of resistance to efforts to improve safety and has withheld internal evidence of risks and harms, even when superficially cooperative, the overall regulatory regime needs to incentivise a significant shift in the tech sector's approach to and prioritisation of online safety.

The structure of the proposals for risk assessments often considers harms separately or in isolation. However, it should ensure that services also consider how risks can intersect or combine in practice, potentially becoming more severe cumulatively, or impacting different users or groups differently. For example, ISD's cross-harm analysis in the UK explores the inter-connection between terrorist and violent extremist communities online, and those involved in the spread of hate speech and conspiracy theories. Similarly, the same principle applies to product features, which can combine and intersect in ways that increase the levels of risk.

Forward-looking and future-proofed measures are imperative to address emerging risks from the start of the regulatory regime. For instance, although evidence is still evolving on how generative AI is used to perpetrate various online harms by hate, extremist or terrorist actors, or its use in foreign interference campaigns (as identified by two ISD investigations into the use of AI technology in Russian information operations), significant evidence already exists on its use in online violence against women and girls (VAWG).

Record keeping and review guidance

Question 10.2: Do you agree with our proposal not to exercise our power to exempt specified descriptions of services from the record keeping and review duty for the moment?

Yes, at least until services can consistently demonstrate low levels of risk, effectively mitigated.

Volume 4: How to mitigate the risk of illegal harms – the illegal content Codes of Practice

Question 11.1: Do you have any comments on our overarching approach to developing our illegal content Codes of Practice?

Please refer to the collective Online Safety Act Network statement and responses to Questions below.

Question 11.2: Do you agree that in general we should apply the most onerous measures in our Codes only to services which are large and/or medium or high risk? Please provide the underlying arguments and evidence that support your views

Question 11.3: Do you agree with our definition of large services?

We have concerns regarding the use of a blunt user number threshold to define services as 'large', following a similar approach to the EU's Digital Services Act (DSA) in applying a threshold based on 10% of the population. Size, while a significant factor, is not necessarily an accurate proxy for levels of risk. Moving beyond the largest platforms (e.g. VLOPs or VLOSEs under the DSA), it is difficult to assess which specific platforms that would approach or meet the threshold.

However, many platforms boasting a substantial user base likely in the millions in the UK, such as Telegram and Discord, cannot be considered 'small' and play a crucial role in the online ecosystem concerning terrorism, extremism, and hate, as evidenced by ISD investigations into terrorist, extremist and hateful communities globally, including in the UK, Canada New Zealand, and Germany. There are also a large number and wide range of even smaller platforms that are likely to pose significant and severe risks, as indicated by other examples from ISD research outlined above. Consequently, ISD recommends either establishing a lower threshold for 'large' platforms (1-2 million+ users) or introducing a 'medium' category to create a more graduated scale of requirements between 'small' and 'large.'

Additionally, ISD would also advocate for increased transparency, currently lacking under the DSA, regarding how services calculate their user numbers, the regularity of required reviews, and how Ofcom plans to independently verify them. This is particularly important given the potential risk of companies claiming exemptions for certain aspects of their services or attempting to sub-divide their user numbers to evade meeting higher thresholds.

Question 11.4: Do you agree with our definition of multi-risk services?

There is a potential danger in the system creating an incentive for services to downplay risks in order to evade triggering additional requirements linked to the 'multi-risk' categorisation. Certain platforms may also pose a specific single risk but to an exceptionally severe extent, such as platforms designed to perpetuate specific harms like terrorist platforms/apps or tools utilising artificial intelligence for generating non-consensual sexual 'deepfake' images or videos. To address this, requirements and obligations for services should be determined through a comprehensive risk assessment process. This process would enable a more nuanced evaluation of the nature and severity of risks posed by different platforms, ensuring that regulatory measures are proportionate to the actual risks presented by each service.

Question 11.6: Do you have any comments on the draft Codes of Practice themselves?

Please refer to the responses to the questions above, and the collective Online Safety Act Network statement.

Content moderation (user-to-user / search)

Question 12.1: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

Question 13.1: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

Content moderation is a long-standing approach employed by the majority of services to address online harms. However, the variety, extent and prevalence of online harms demonstrates that content moderation measures, and associated user reporting tools, are necessary but not sufficient in effectively mitigating risks. Therefore, as a primarily reactive measure, content moderation should not be over-relied upon over proactive safety-by-design efforts to prevent or discourage illegal content or activity from occurring in the first place.

Services' content moderation efforts have also too often been characterised by a lack of genuine transparency, both for individual users and at a more macro level to enable independent external assessments of their proportionality, consistency and effectiveness. In our view, many services' existing transparency reports often rely on self-selected metrics and measures of success that do not present an objective assessment. We would therefore recommend Ofcom introduces baseline expectations and consistent measures to assess their impact in mitigating risks and reducing harms, and allow for cross-industry comparisons.

Automated content moderation (User to User)

Question 14.2: Do you have any comments on the draft guidance set out in Annex 9 regarding whether content is communicated 'publicly' or 'privately'?

We strongly agree with A9.19 – "*The fact that content (or any parts of the service on which that content is generated, shared or uploaded) is labelled as 'private'*" – and would recommend that factors such as size, purpose, accessibility and the nature of relationships between users of a channel or a community should be taken into account when making assessments about public or private spaces online.

Currently the Guidance does not specify a user number threshold to determine whether an online space is public or private, and leaves this up to services to determine. This could lead

to greater inconsistency across services, confusion among users, and could provide a potential loophole for services to designate significant proportions of their platforms as private in order to avoid requirements to ensure risks are effectively mitigated. It also may not be possible to accurately determine location of users able to access content on certain services, meaning that it may be difficult for them, Ofcom and independent researchers to understand the availability of content to users in the UK.

For example, ISD's research has demonstrated how larger 'private' groups on Telegram, with invite links openly shared in public channels, are often used for calls for violence, the spreading of harmful conspiracy theories and organising (sometimes violent) offline mobilisation. Telegram's rules currently do not apply to either public or private groups (that can have up to 200,000 members) or to private channels, which makes these spaces an attractive online venue for harmful actors. Telegram remains a platform without a comprehensive system of moderation. Although the platform sporadically removes illegal content or content promoting violence, these actions remain inconsistent and insufficient. Private groups therefore provide an additional loophole for circumventing legal requirements to remove illegal content.

If Ofcom is reluctant to set a blanket threshold, it should require companies to set clear and reasonable limits based on the nature of their platforms and risk assessments that consider a platforms' specific features or functionalities (e.g. encryption), risks and potential vulnerabilities. Ofcom should also encourage companies to make clear to users which aspects of their platform are more public or more private, as well as the consequences of this for user privacy and the enforcement of Terms of Service. Ofcom should then assess, based on the risk assessment, whether the limit set by the platform is appropriate and sufficiently mitigates any risks or harms identified.

User reporting and complaints (U2U and search)

Question 16.1: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

Content reporting by users alone is not an effective means to ensure the safety of users, and the responsibility for identifying illegal content should not be shifted from services onto users. Instead, it should be considered as complementary to other more proactive safety-by-design measures.

It is also crucial that users are allowed to provide contextual information along with their reports and complaints. This becomes particularly significant in cases of persistent abuse or harassment, where services often fail to grasp the broader context of a targeted user's experience.

Transparency in decision-making processes regarding content moderation is also often lacking. Ofcom should require services to provide sufficient information explaining why a particular decision has been made, either when a user has their content moderated, or when a service does not take action on a users' report. Given the lack of consistency in many services content moderation decisions, Ofcom should also require services to allow for appeals for users when their reports are not actioned.

Recommender system testing (U2U)

Question 19.1: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

We support proactive measures to assess the potential risks associated with recommender systems as a key part of an overall safety-by-design approach, and to ensure that these are conducted prior to their use. A similar approach should be applied to any changes made to such systems, and repeated regularly to ensure new or emerging risks are fully considered and effectively mitigated at an early stage.

Question 19.2: What evaluation methods might be suitable for smaller services that do not have the capacity to perform on-platform testing?

In addition to our response to Question 19.1 above, services should also consider whether the presence of untested recommendation systems is appropriate, or whether they can be safety deployed without the ability to test the impact of any future changes, either to optimise for business or safety-related objectives.

Enhanced user control (U2U)

Question 20.1: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

While user tools can offer important additional options and flexibility for users, it is imperative not to depend solely on them to mitigate risks. The responsibility for risk mitigation should not be shifted onto users. Instead, these tools should complement a broader approach for online safety. Services should also provide bulk reporting, blocking, and muting tools for victims of online harassment or abuse, especially for high-profile, vulnerable, or marginalised users.

Question 20.3: Do you think there are situations where the labelling of accounts through voluntary verification schemes has particular value or risks?

Labelling accounts can serve as a useful tool, providing users with additional cues to assess the veracity of information or the trustworthiness of an account. However, services must ensure transparency in explaining how verification works and apply it consistently across their platforms.

It is important to note that the misuse or inconsistent application of account labels can pose risks. Overly broad categories may be unhelpful, obscuring variations in the trustworthiness of different sources. For example, if all government or journalistic sources are labelled in the same way, it may overlook the fact that different media sources operate to varying editorial standards, or that certain (often undemocratic) government entities or accounts may have a demonstrated track record of disseminating disinformation.

User access to services (U2U)

Question 21.1: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

From an external perspective it is not possible for civil society researchers to understand or assess the way in which services' account strike policies are implemented or enforced. However, the prevalence and extent of online hate and abuse suggests that while there is a place for such measures, they do not appear to significantly disincentivise certain users from perpetuating online abuse or harassment.

Similarly, in ISD's research on terrorism online, we have observed the ease with which users are able to create new accounts (and this is often communicated as a point of pride) when previous accounts have been suspended or banned. Services should therefore not rely on such measures to deal with persistent networks of harmful accounts, and should instead take a more proactive approach. Ideally this should include specialist harm-specific teams within services assessing an accounts' wider network for similar illegal activity.

We would also suggest that account strikes are not an appropriate response to the sharing of highly harmful illegal content, such as terrorist content or CSAM. To mitigate risks to freedom of expression associated with the erroneous takedown of content wrongly assessed to be illegal terrorist content (and permitted with a services Terms), often as a result of automated content moderation systems, services could consider an expedited appeals process.

Overall Assessment

Question 23.1: Do you agree that the overall burden of our measures on low risk small and micro businesses is proportionate?

Question 23.2: Do you agree that the overall burden is proportionate for those small and micro businesses that find they have significant risks of illegal content and for whom we propose to recommend more measures?

Question 23.3: We are applying more measures to large services. Do you agree that the overall burden on large services proportionate?

Please refer to the collective Online Safety Act Network statement. and responses to Questions 9.1 and 9.2.

Statutory Tests

Question 24.1: Do you agree that Ofcom's proposed recommendations for the Codes are appropriate in the light of the matters to which Ofcom must have regard? If not, why not?

Please refer to the collective Online Safety Act Network statement.

Volume 5: How to judge whether content is illegal or not? (Illegal Content Judgements Guidance)

Question 26.1: Do you agree with our proposals, including the detail of the drafting? What are the underlying arguments and evidence that inform your view.

Overall we would support the feedback provided by Professor Lorna Woods for the Online Safety Act Network. However, ISD also has the following harm-specific suggestions for more detailed guidance for illegal content judgments related to terrorism offences:

A) In the guidance on '*Proscribed groups*', it is stated that content that is explicitly supportive of a proscribed organisation and is posted for example, to a far-right chat group will not necessarily fall outside the definition of illegal content merely because the user adds "lol" or "joking" to the end of it. However, ISD notes that identifying this context is often challenging in online subcultures where a culture of 'shitposting' exists around expressions of extreme violence is commonplace (e.g. 4chan). It would be helpful to have better guidance on the boundaries of this, for example in the case of the use of the phrase 'in Minecraft' after calling for acts of terrorism, to introduce plausible deniability.

B) In the section on '*Information likely to be of use to a terrorist*', Ofcom states that services should first consider whether the content is information that is, of its very nature, designed to provide practical assistance to a person committing or preparing an act of terrorism. It is an offence to collect, make a record of, possess, view or access

such information. It would be useful to have more guidance on the thresholds for *"of its very nature designed[…]"*. There is very often a grey area of material which can be useful for terrorist purposes, but where it is not clear if it 'by its very nature' terroristic. ISD's analysis of post-organisational terrorism has found that a number of instructional publications have been used as permissible evidence in terrorism trials, which are not themselves inherently terroristic (e.g. The Anarchist Cookbook).

C) The guidance states that a potential reasonable excuse may be if the service has evidence that the content is being used for journalistic or academic research purposes, and that its dissemination is properly and effectively limited to persons with that purpose. It would be helpful to provide further guidance on the limits of academic research in this context, and how dissemination could be effectively limited. For example, whether this includes a broader swathe of researchers, including civil society groups or journalists covering on counter terrorism or extremism topics.

D) When considering the dissemination of terrorist publications, the following example is used – "*Publishing on the internet publications authored by known terrorists, e.g. 'shooter manifestos' such as Anders Behring Breivik's manifesto*". However, it would be useful to have more guidance on how 'known terrorists' is defined – for example, would this include lesser-known terrorists, or those often regarded as terrorists (especially outside the UK) prosecuted under, for example, federal hate crime legislation (e.g. Elliot Rodger). There is also a question about the case of ideological manifestos which do not include overt encouragement or guidance, but are still associated with terrorism, e.g. the proliferation of Ted Kaczynski material within violent extremist online communities.

E) Another usage example of dissemination of terrorist publications is "*publishing on the internet publications known to be distributed by terrorist network (for example Siege by James Mason)".* It would be helpful to have more guidance on how 'terrorist networks' are defined in this context. While this is clearer for proscribed groups like Atomwaffen or National Action, would this, for example, include the Hard Reset by Terrorgram, 09A's Iron Gates, or the White Resistance Manual (upon which convictions have been based).

F) For the guidance around Encouraging Terrorism, a usage example is provided of *"Content calling on others to emulate or follow today, the acts of historical figures who used violence for political ends".* Our concern is that this could be interpreted very broadly to encompass 'legitimate' violence towards political ends (e.g. Suffragettes, Malcolm X, Syrian resistance) alongside examples of calling on others to emulate or follow contemporary figures who have used violence for extremist political ends ('Saints Culture' phenomenon of far-right terrorists).

G) In Ofcom's guidance on the '*Preparation of terrorist acts'* it is stated that "*reasonable grounds to infer that an account is operated by or on behalf of a proscribed group may also arise where a significant proportion of a reasonably sized sample of the content recently posted by the user amounts to a proscribed group offence*." However ISD research has shown examples of aggregator accounts promoting branded terrorist content following Hamas' attacks in Israel 7th October 2023 – which were clearly not operated by or on behalf of a proscribed group. This potentially makes it challenging to determine the '*reasonable grounds to infer'* or '*reasonably sized sample of content'* required to determine that an account is meaningfully associated with a proscribed group.

Additionally, ISD has harm-specific suggestions for more detailed guidance for illegal content judgments related to hate and harassment offences:

A) Crucially the only types of hatred specifically covered by the offences outlined by Ofcom are religion, race or sexual orientation. However, hate offences also impact upon individuals from other marginalised groups in society. Accordingly, it should be noted that some of the offences outlined above may also be hate crimes if they are motivated by hostility based on race, religion, disability, sexual orientation or transgender identity. Although the approach laid out here suggests that services should take action on the basis of the offence which is easiest to prove, it is recommended that these factors (e.g. hate speech targeting a wide range of communities) are explicitly given consideration in any risk assessment.

B) Ofcom notes that "*In considering the offences in this section, services will need to be particularly mindful of context and nuance. Content is not illegal merely because it is offensive, shocking or disturbing; nor because it is rude. Lawful content may express unpopular or unfashionable opinions about serious or trivial matters. Banter and humour, even if in poor taste to some or painful to those subjected to it, is not necessarily unlawful.*" Whilst it is absolutely right that offensive or shocking humour is protected, here it should be considered that extremist movements and subcultures specifically leverage transgressive humour and irony, including in hateful targeting of minority communities. This ISD Explainer details the ways in which right-wing extremists use humour to obfuscate hateful narratives. It would be helpful here to have more detail around the thresholds of protected speech. This point about humour is also pertinent to considerations around violent threats, including hate.

C) Ofcom notes that the only specific defence to threatening behaviour is *"that the threatening behaviour was 'reasonable' in the particular circumstances in which it happened."* Could Ofcom provide examples of when threatening behaviour may be reasonable? Would, for example, a video of drill music shared online which specifically calls for violence be protected?

D) Ofcom notes that "*judgements about whether content is likely to cause harassment or distress, and whether the defence of reasonable behaviour is available, are likely to be particularly difficult when services are considering content that has political or religious purposes and relates to religion, sexual orientation or gender identity*." Is there any guidance around the thresholds of acceptability within certain contexts? As this is currently worded it suggests that hateful rhetoric inspired by religious or political beliefs could be more acceptable than non-religious or politically inspired hateful rhetoric.

Question 26.2: Do you consider the guidance to be sufficiently accessible, particularly for services with limited access to legal expertise?

While as a civil society organisation it is not possible for us to respond from the perspective of services, especially smaller services and/or those based outside of the UK (which may be more likely to be exploited for terrorist or other illegal or harmful purposes), we suspect the depth and legal complexity of the guidance will be challenging for many to engage with. As a result, we suspect many platforms will instead adopt a Terms of Service based approach that covers all the forms of illegal content covered by the guidance. In many cases, this would likely require significant changes to their existing Terms.

Volume 6: Information gathering and enforcement powers and approach to supervision

Question 28.1: Do you have any comments on our proposed approach to information gathering powers under the Act?

It is vital to the effective implementation and enforcement of the Act that Ofcom uses its information powers effectively. As online safety regulation is only starting to be implemented in many jurisdictions, and as a result of longstanding restrictions on data access and services lack of transparency, the current evidence base surrounding online harms and the impacts of services design and functionalities is often patchy and incomplete.

While proportionality is an important consideration, both for smaller services and for Ofcom's ability to make effective use of the information requested, all platforms that pose significant risks to online safety should be expected to respond in full to information requests. Ofcom can also play a role in supporting services to comply.

It is additionally important that Ofcom makes effective use of its' information gathering powers in the absence of stronger research access to data provisions for researchers in the Act (in comparison to the EU's DSA for example), especially in an environment where many

services either do not offer or have decided to restrict options for independent access to data, as we have outlined in previous research on data access challenges facing researchers across the online ecosystem.

Question 29.1: Do you have any comments on our draft Online Safety Enforcement Guidance?

Overall we support Ofcom's proposed approach to enforcement outlined in Volume 6 and the draft Online Safety Enforcement Guidance.

From a strategic perspective, we would also emphasise the importance of A3.9.b.i. - "*whether enforcement action would help clarify the regulatory or legal framework for other stakeholders*" - in the context of small services that pose severe risks (whether single or multi-risk) and that may be designed and/or operated with an implicit or explicit intent to facilitate harm. Where such services are unresponsive or uncooperative, we would encourage Ofcom to take swift, precedent-setting action to help ensure other similar services understand their obligations under the Act.