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* Glitch is a UK charity (no. 1187714) that exists to end online abuse and to increase digital citizenship across all online users. We believe that our online community is as real as our offline one, and that everyone should work together to make it a better place. We work to promote good digital citizenship and address online harms such as online abuse, online hate speech and information disorders, and have developed bespoke training programmes covering Digital Citizenship, Online Active Bystanders and Digital Self Care and Self Defence. As part of this, we have delivered training to women in public life. We are submitting evidence to Ofcom's inquiry because we believe that the Online Safety Bill regime has the potential to make a significant difference to the prevalence of online abuse experienced by internet users in the UK. However, for it to appropriately serve those disproportionately affected by online abuse – women, and especially Black women, and racialised and minoritised people – Glitch believes that the implementation of the Online Safety Act will need to reflect the experiences of Black women and other marginalised communities subjected to high levels of online abuse.
Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services? **IMPORTANT: Under this question, we are not seeking links to or copies/screenshots of content that is illegal to hold, such as child sexual abuse. Deliberately viewing such images may be a criminal offence**	*Is this response confidential? N* ● As the regulator, Ofcom has an important role in ensuring that tech platforms understand that online gender-based violence is an important subcategory of illegal harm that occurs on their platforms and affects the human rights of their users. ● Much of the online abuse that affects women and girls, including misogynistic racist abuse i.e. misogynoir, either is not illegal, relies on incomplete laws that do not adequately protect women and girls today, or would not reach the criminal threshold. ● Though not disaggregated between illegal and legal content, we know from research that women are 27 times more likely to be harassed online than men, with Black women 84% more likely to be targets of abusive tweets than white women and 60% more likely to receive problematic tweets. ● Significantly, technology-facilitated abuse is included in the Domestic Abuse Act 2021, with perpetrators of domestic abuse increasingly using existing and emerging technologies to continue to perpetuate abuse, for example both during and after intimate partner relationships. Home Office Domestic Abuse statutory guidance states that 'Perpetrators can use technology, including social media to abuse victims';

and will be reported to the police.	• Refuge's research on tech abuse shows that while much of the domestic abuse related online abuse (tech abuse) taking place on tech platforms like social media is illegal (e.g. harassment, stalking, non-consensual sharing of intimate images/videos), over half of women who reported domestic abuse to the police said they had their report handled badly. 56% of women reporting abuse from a partner or former partner to social media platforms said their reports were handled badly. • Refuge estimates that almost 2 million women in the UK have faced online abuse from a partner or former partner. **As per the VAWG Code of Practice:** **Enforcement of criminal law** (1) Service providers must have in place a point of contact for law enforcement authorities in theUK. The contact is responsible for giving information about potentially criminal content to law enforcement authorities under para 2. This includes – (a) information about the content; (b) the details of the user, including location; (c) details of enforcement action on the content undertaken by the provider; and (d) other materials relevant to criminal investigations. (2) Information requested by government and law enforcement authority in accordance with UK law should be delivered within the time frame specified by national rules or no later than one month of receiving the request. In exceptional circumstances this can be extended, with written approval from the relevant authorities placing the request, with a full expected time frame set out. (3) Effective protections should be put in place by service providers to ensure flagging and court orders are not used for malign purposes by Government agencies or law enforcement of any kind to remove content they find objectionable, which is neither illegal nor harmful. **Transparency Reports** 1. Online services must publish transparency reports in line with Ofcom's guidelines. These must be easy to access and understand. Online services must be prepared to answer questions on the findings. 2. On request, online service must provide individuals with easy to digest data that the online services hold on them. 3. Online services must uphold individuals' right to be forgotten and rights under GDPR. 4. Online services must respond to requests for information by any Government or Ofcom appointed user advocate in the required time. 5. Online services must proactively share information with third sector organisations where it is relevant for the organisation to safeguard the citizens that they represent on a regular basis such as quarterly meetings. For instance, this could include sharing intelligence on – (a) Themes and categories relating to VAWG moderation and user reporting.

	(b) Scale and dimensions of online risk experienced by women and girls. (c) Data relating to emerging risks and new trends in online harm perpetration. (d) Effectiveness of risk mitigation tools and protective measures in place relating to VAWG online. 6. Online services must maintain effective channels of collaboration and communication with civil society organisations with expertise in these areas. 7. Online services must consider whether decisions on gender-based harms would benefit from consultation with civil society including VAWG specialist services. For instance, when risk assessing new technology.
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* ● It is important that platforms use plain, accessible language, including local languages (official languages of UK and languages spoken in UK) ● They must alert users to simplified terms of service and policy statement updates both when accounts are set up and periodically as they are updated and make them easy to access for users searching out such terms ● Issues around clarity not only applies to the language but also the confusion caused when users believe that they see clear violations of the stated terms of service and either receive an automated/machine generated response to reporting abuse or human moderation that states that terms of service have not been infringed. ● Accountability mechanisms for these decisions are important. For example, reviewing case decisions and appeal processes. ● Media literacy and awareness around terms of service are also very important. There must be strong education campaigns to help users understand policies. Tech companies should work with organisations like Glitch to ensure education programmes and trainings are relevant, up-to-date, and aligned. **As per the VAWG Code of Practice:** 1. Regulated services should have in place Terms of Service which are clear and accessible by all likely users; this includes being age-appropriate and accessible for those with disabilities and different access needs. The terms of service should include how the service responds to VAWG, including actions taken to prevent VAWG, and be visible to would-be users before they sign up to the service. Community standards should also be visible and should, where relevant, cover the content of advertising. 2. Regulated services should undertake regular, systemic reviews of their Terms of Service and Community Guidelines to ensure that they remain up to date, effective, and proportionate.

	3. To ensure Terms of Service and Community Guidelines are effective, regulated services need to review how they are operating and how they are enforcing them.
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* ● Tech companies must publish transparency reports and be prepared to answer questions on the findings. ● A major issue around transparency, accessibility and user's willingness to report is that many people on social media platforms feel that reporting either to the police or to platforms is futile. ● Trust has broken down between platforms and users in relation to platforms failing to stop online gender based violence or respond to it when it is reported . ● UltraViolet's Report Card for social media platforms from December 2021 demonstrates issues across the different major social media platforms. ● It is essential that platforms make it clearer whether a human moderator or AI system has been used in a decision relating to reporting and complaints mechanisms ● Platforms should allow trusted research institutions and civil society organisations to access appropriately anonymised and aggregated data, including relating to reporting and complaints mechanisms, which includes the type of action taken, the time it takes to review reported content ● Platforms need to understand online gender based violence and how it is occurs on their platform, paying for the expertise of the organisations such as Glitch working to end online abuse and the specialist violence against women and girls sector. If tech companies have clear policies on gender-based violence, women and those with monoritised genders are more likely to report abuse as they know that action should be taken, based on these policies and the enforcement encouraged by regulation ● Increased transparency around appeal processes is needed when action has or has not been taken after a report or complaint is made. ● More transparency is needed from tech companies about their policies related to dehumanising language based on gender, race and other protected characteristics, as well as additional forms of discrimination not recognised in the Equality Act 2010. ● Policies should be regularly reviewed and updated to address new trends, patterns and manifestations of online abuse including forms of gender-based violence against women and people with intersecting identities ● Companies should be constantly monitoring their platforms for new, emerging trends as well as those that remain pervasive and common ● Tech companies should be far more transparent about who within companies is setting the agenda relating to these systems, and how changes are decided, designed, who is involved or consulted and what measurable changes have taken place. We should not rely on whistleblowers to understand harmful practices within tech companies.

	• Education and awareness raising campaigns, as well as resources, are incredibly important in relation to raising awareness of why reporting exists, why it's important to report and what happens at each step so that users can be confident in the process and the regulator can hold companies to account, based also on user-expectations. **As per the VAWG Code of Practice:** **Reporting Mechanisms:** 1. Users must be able to effectively report content that is illegal or harmful to regulated services through clear and transparent flagging mechanisms. Regulated services are obligated to have effective and easy to use reporting functions and must use them to triage content for both human and automated moderation. 2. Service providers should have reporting processes that are fit for purpose for reporting VAWG content and wider harms, that are clear, visible and accessible and age-appropriate in design. Thought should be given to reporting avenues for non-users such as teachers or family friends and support services, who are able to report without the victim needing to engage further with the harm. 3. Service providers should have in place clear, transparent, fair, consistent and effective processes to review and respond to content reported as VAWG content. Users must be given the ability to submit third-party content to the companies' intelligence systems in relation to specific cases of content violation. 4. Reporting processes should set out clear time frames and should inform the user directly of any decision made. Reporting processes should include a specific point of contact that is provided to users so users are able to follow up on decisions made.
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* • Tech companies need to invest more in human moderation and ensure moderation considers local context, including (but not limited to) linguistic, social, cultural, historical, racial and gendered context. • Human moderators should work in holistic environments which appropriately support their wellbeing, proportionate to the level of upsetting and harmful material they are moderating • Diversity within teams is incredibly important, as is the training that is offered, for example training on recognising and responding appropriately to racism; online gender based violence; anti-transphobic content etc. • Human moderators should be paid well in recognition of the heavy burden of a difficult job. • Comprehensive training for moderators about online gender-based violence and different tactics of online abuse, and how abuse specifically targets women, Black and minoritised communities and users with intersecting identities is paramount - without this moderation risks being ineffective, inequitable and/or discriminatory • Tech companies need to be transparent about their investment in and resourcing of content moderation

- It is essential that users understand when content has been moderated by human moderators and when it has been moderated by other means.
- Platforms should acknowledge internet biases in machine learning and AI systems and aim to eliminate biases through trusted partner interventions, for example through 'bias bounty challenges' and open access for researchers if these approaches are deemed to be effective by Ofcom.

As per the VAWG Code of Practice:

Moderation:
(1) Regulated services must have in place sufficient numbers of moderators, proportionate to the online service size and growth, and to the risk of harm, who are able to review VAWG content. This may include moderators who work exclusively on VAWG issues.
(2) Regulated services must put in place appropriate, updated education and training on VAWG for all staff and subcontractors involved in the content production and distribution chain. This includes senior executives, designers, developers, engineers, customer support and moderators, designed in consultation with independent VAWG experts. The moderators must be appropriately trained, supported and safeguarded.
(3) Regulated services must consider assigning moderators to specific types of VAWG content to ensure the correct moderators, trained in their specialist subjects and on related language and cultural context considerations are able to review the content in a consistent fashion.
(4) Regulated services must have in place processes to ensure that where machine learning and artificial intelligence tools are used, they operate in a non-discriminatory manner and that they are designed in such a way that their decisions are explainable and auditable. For instance, technology to remove sexualised pictures must not remove photos of breast feeding. A platform provider should consider the way in which AI and machine learning systems and/or human moderators will distinguish between hateful and harmful content, reclaimed terms used by particular groups, and that of 'counter speech', minimising the risk of blocking or limiting legitimate use of terms within certain online communities and counter speech.
(5) Users must be informed of the use of such automated tools. Machine learning and artificial intelligence tools cannot wholly replace human review and oversight.
(6) If the VAWG content involves a person protected by UK law, regulated services must review the content taking into account the terms of service and UK law.
(7) Regulated services must have clear timeframes for action against flagged content, in line with the good practice outlined in the previous section. Awareness begins at the time flagged content, by means of email, in-platform notification, or any other method of communication, is received.
(8) Regulated services must act, proportionate to risk, on content which is not deemed to be illegal but is considered to break their Terms of Service, Community Guidelines, or is considered a new form of VAWG, as soon as it is identified. Acceptable actions on a piece of content which violates a provider's Terms of Service can include –

(a) removal of content;
(b) labeling as inaccurate/misleading/contrary to the rules;
(c) demonetise content;
(d) suppress content in recommender tools;
(e) termination of account;
(f) suspension of account;
(g) geo-blocking of content;
(h) geo-blocking of account;
(i) issuing a strike, if a strike system is in place;
(j) instituting delay in posting content or otherwise adding friction to the communication process;
(k) limiting number of posts over a given time period; and
(l) adding friction to mechanisms by which content may be shared.

(9) Regulated services must have systems of assessment and feedback to the initial reporter and the owner of content that has been flagged and actioned to ensure transparency of decision making. Users must be kept up to date with the progress of their reports and receive clear explanations of decisions taken.

(10) Provide holistic support for moderators who are exposed to harmful content in recognition
of psychological impacts of what they are exposed to (examples may include mental health
support or clinical supervision).

(11) Online services must consider putting in place an appropriate trusted flagger programme that maintains independence from the online service and from governments. The programme must include UK based non–government organisations and other experts, including the specialist VAWG sector, who will be vetted, to inform on policy development and report on new trends in harmful and illegal content. It is recommended service providers have a Trusted Flagger Policy that includes –

(a) trusted flaggers are not used as a sole provider of flagging content;
(b) trusted flaggers are appropriately compensated and incentivised for work provided to companies to ensure their compliance while not compromising their independence and impartiality;
(c) regular meetings held (with members of the trusted flagger programmes) to review content decisions and discuss any concerns;
(d) provision of support for trusted flaggers who are exposed to harmful content, as per the support provided to the companies' own moderators, whether directly employed or working for out-sourced companies;
(e) a specific Trusted Flagger reporting email address;
(f) a specific trusted flagger escalation route if no / unsatisfactory response received;
(g) clear criteria for what can be reported and what cannot;
(h) clear limited and reasonable expectation for additional information on escalation;
(i) commitment to an expectation on response times of 24 hours. Responses should include details of action taken or reasons for rejections and should include links to policies or Community Standards as relevant;

(j) willingness to reopen a case and review if additional information comes to light; and

(k) adoption of automatic suspension of content reported via Trusted Flagger route pending review.

(12) Where online services use civil society organisations for significant undertakings, they must
consider remunerating them for their time and expertise.

Dispute resolution

(13) Regulated services have an obligation to instigate dispute resolution functions which allows users to raise a complaint against decisions made by the platform.

(14) Regulated services are obligated to put in place a right of appeal on all decisions made concerning illegal or harmful content, or content that has been flagged as illegal or harmful content. All users must be given a right to appeal any measures taken against them. Users must be able to present information to advocate their position.

(15) Regulated services must acknowledge an appeal request within 24 hours of receipt. If more time is needed to assess the content the user must be informed.

(16) Regulated services must have appeals systems which must take no longer than seven days to assess appeals, except in exceptional circumstances. Exceptional circumstances could include a major disaster, or an event or incident of the same magnitude.

(17) Regulated services must explain the outcome of a dispute in clear and simple language.

(18) Complaints related to VAWG must be reviewed by a professional trained in VAWG issues for example by a VAWG specialist service.

(19) Dispute resolution procedures must be fair, transparent, and easy to use. They must not discriminate between users, introduce bias, or be applied inconsistently

(20) Regulated services must remain conscious that children may not be able to access dispute resolution procedures and offer alternative mechanisms for children to raise issues.

Discovery and navigation

(1) Regulated services should review their recommender systems, especially their automated systems, so that they do not cause foreseeable harm, including VAWG, through –

(a) promoting VAWG content;

(b) suggesting groups or other users to follow that endorse or positively view VAWG or
misogyny; and

(c) rewarding controversy with greater reach, causing harm both by increasing reach and engagement with a content item.

(2) Consideration must be given, in line with child-related duties, as to how to protect children to a greater degree.

(3) Platforms must consider how easily, quickly, and widely VAWG content may be disseminated by means of the service and respond appropriately.

(4) Regulated services should consider the impact of autoplay functions, especially in the context of content curated or recommended by the provider.

	Where the service provider seeks to take control of content input away from the person through autocomplete or autoplay (see below). The provider should consider how this might affect a person's right to receive or impart ideas. (5) Regulated services should consider the need for explainability or interpretability, accountability and auditability in designing AI and machine learning systems, particularly with regard to the representation of women and girls, especially those from minority groups, in their data sets. (6) A platform provider should consider the speed and ease of transmission, for example methods to reduce the velocity of forwarding and therefore cross-platform contamination. (7) A platform provider should consider the way in which AI, machine learning systems and/or human moderators will distinguish between hateful and harmful content, reclaimed terms used by particular groups, and that of 'counter speech', minimising the risk of blocking or limiting legitimate use of terms within certain online communities and counter speech. (8) A platform provider should be responsible for ensuring that algorithms do not suggest material that is in contravention of the site's own Terms and Conditions.
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* ● Glitch is aware of emerging tools from TikTok and Twitter as we sit on their Trust and Safety Boards. We believe that social media platforms benefit from the expertise of such boards who can feed into the development of such tools and early emerging evidence related to them. ● Nudge behaviour changes is sometimes suggested as a solution to long lasting change and can be implemented on platforms. For example, Twitter has developed a feature that asks whether a user is sure they want to send a potentially harmful Tweet, leading to many users (up to 30%) deciding not to post, or amend their message. ● Other social media platforms have also built in reminders for users to take time away from their devices, or increased consentual parental controls - see Instagram's parental controls and TikTok's built in screen time manager ● Bumble uses 'Privacy Detector' message screening to counter cyberflashing and other 'lewd images' ● Other interventions, such as Twitter circles (Introducing Twitter Circle, a new way to Tweet to a smaller crowd) and turning off replies from select accounts has been positive too. ● More user tailoring tools are key to enhancing user's experiences online. ● Closing private messaging options are also important. **As per the VAWG Code of Practice:** (1) Regulated providers must implement appropriate "safety by design" technical and organisational measures, including but not limited to those detailed in these Guidelines. The intended outcome is to

(a) minimise the risk of those harms arising from VAWG content and practices
(b) mitigate the impact of those that have arisen,
(c) enhance women and girls' freedom online
taking into account the nature, scope, context and purposes of the online platform services and the risks of harm arising from the use of the service.
(2) Companies must ensure and be able to demonstrate their systems are safe by design, including addressing the following concerns:

(a) Taking an appropriate and proportionate approach to the principle of knowing your client [KYC] to address VAWG harms spread by those using multiple, false, or anonymous identities.
(b) Ensuring that young users' settings are set to safety by default.
(c) Ensuring algorithms used on the service do not cause foreseeable harm through promoting hateful content, for example by rewarding misogynistic influencers with greater reach, causing harm both by increasing reach and engagement with a content item.
(d) That speed of transmission has been considered, for example methods to reduce the velocity at which intimate images can be non-consensually shared and therefore the risk of cross-platform contamination.
(e) Actors cannot take advantage of new or emerging tools to cause harms to women and
girls. For instance –
– deep fake or audio-visual manipulation materials.
– nudification technology.
– bots and bot networks.
– content embedded from other platforms and synthetic features such as gifs, emojis, hashtag.
– other new technology
(f) Consideration of the circumstances in which targeted advertising may be used and oversight over the characteristics by which audiences are segmented.
(g) Account security systems which enable survivors of abuse, who are hacked and locked out, to recover their accounts.
(h) Systems for cross-platform co-operation to ensure knowledge about forms of offending that may present a foreseeable risk of harm in relation to attacks of those with protected characteristics.
(i) Use of tools including, but not limited to, prompts which clarify or suggest an individual's intended search.
(j) Policies concerning advertising sales in respect of promoting harmful content or for malicious intent in respect of those with protected characteristics.

Settings and Tools
(1) Regulated services must empower users by providing tools which, in addition to content and behaviour reporting tools, allow users to improve control of their online interactions and to improve their safety. These could include –

(a) controls over recommendation tools, so a user can choose for example to reject personalisation. Examples include –
– user-set filters (over words or topics)

	– tools to limit who can get in touch/follow a user, or to see a user's posts. – tools to allow users to block or mute users, or categories of user (for example anonymous accounts); (b) tools for adapting privacy settings and setting privacy options as default for young and vulnerable users; (c) controls for the user over who can and cannot redistribute their content or username/identity in real time; (d) the ease of use of these tools and their prominence such that users are aware they exist; – including ease of use for children and those with accessibility issues (e) specific tools in place for users under 18. This could include – – Tools to stop children from receiving unsolicited messages from adults – Measures which are targeted at the adults doing this – Notifications to make an adult messaging a child aware of the policies of the service in relation to communication with children – Notifications to ask a child if they know who is messaging them and to explain what children can do if they are confused or made to feel uncomfortable by it
Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	*Is this response confidential? N* ● There is a huge data gap related to online gender based violence across tech companies. Data needs to be disaggregated by characteristics of users where known or assumed by tech companies ● Access for independent researchers and civil society organisations ● Data disaggregated by protected characteristics ● Data that related explicitly to VAWG-related harms
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? N* Glitch strongly recommend the adoption of: ● A **VAWG Code of Practice** ● Intersectional approach that acknowledges the disproportionate impact on Black women and marginalised communities online ● Further data and recommendations can be found in The Ripple Effect: Covid-19 and the Epidemic of Online Abuse, published by Glitch and the End Violence Against Women Coalition (EVAW) in September 2020. **As per the VAWG Code of Practice:** 1. Regulated services should have a specific policy commitment to prevent and take action to combat VAWG arising on their service. This commitment should be endorsed by the UK leadership of the organisation and a board member, or person reporting into the board, appointed to be accountable for delivering it. The policy should be informed by specialist VAWG expertise. It should clearly set out the values of the regulated service.

2. (a) Regulated services should carry out a suitable and sufficient assessment as to the risk of VAWG-related harm, taking into account international human rights standards, obligations and best practice. Risk assessments must take into account and mitigate potential harms arising from intersecting inequalities. This means the particular risks of harm to people with more than one or overlapping characteristics that typically experience discrimination and oppression, arising from the operation of the service or any elements of it. The risk assessment should be accompanied by a mitigation plan that addresses the issues raised in this Code.

   (b) The risk assessment should not solely consider individual risks to individual users but also consider broader social and cultural harm, such as the ways in which all women are affected by the threat of violence and harm even if they have not directly experienced it themselves.

   (c) The risk assessment should be carried out before any new service or any new feature is made available. It should include consideration of how different types of content are shared and practices carried out on the platform, and by whom.

3. Service providers should identify suitable metrics to assess the appropriateness and success of the mitigation plan overall, and in relation to each set of risks and use them to assess effectiveness of the mitigation plan regularly (at least annually) and revise the mitigation plan accordingly.
4. The risk assessment should be reviewed by the service provider on an ongoing basis or, if there is reason to suspect that it is no longer adequate or complete; or there has been a significant change in the matters to which it relates. Where as a result of any such review changes to a mitigation plan are required the service provider should make them.
5. Risk assessments and mitigation plans should be recorded, retained for a period of no less than three years and published on the service provider's website in an accessible manner.
6. All measures taken in the following guidelines, including the metrics at (4), should feed back into the risk assessment as it evolves.