Your response

Question	Your response
Overview: Glitch welcomes the opportunity to submit to this call to evidence ahead of the Online Safety Bill returning to the Commons after its passage through the House of Lords. While we have responded to relevant questions below, we are keen to point out that during the Lords stages on this topic, Baroness Nicky Morgan raised concerns around categorisation and her amendment (245) was passed at Report Stage in the Lords - which makes categorisation decisions for Ofcom relating to Category 1 services based on either size or functionality, rather than both. We know that this call to evidence was published before this amendment was agreed, and that this will be reflected in the future version of the Online Safety Bill. We are also aware that following an amendment from the government for guidance for women and girls - in response to the Baroness Morgan supported amendment based on the draft violence against women and girls Code of Practice that Glitch drafted alongside civil society colleagues, academics and Carnegie Trust - we would appreciate more clarity on how and whether this guidance will impact categorisation of platforms, as well as other elements of the Bill. that it is unclear how this future guidance will relate to all elements of the implementation of the Online Safety Bill, including the task of categorisation in question. Glitch also published new research into Digital Misogynoir in July, which is highly relevant to categorisation. The Digital Misogynoir Report: Ending the dehumanising of Black women on social media examines almost one million posts concerning women across five social media platforms: Instagram, Facebook, Twitter, Gab and 4chan. In this research, we compare abuse on large established platforms to those on small high harm platforms, highlight how specific jargon and conspiracy theories move from small platforms known to be havens of extremist views to the mainstream, where these ideas are further amplified and contribute to the shifting of social norms towards extremist views, such as those of white-supremacists. For this reason, we are concerned that high harm platforms that contribute to discourse not just within their own services but also in the mainstream, leading to harms both online and offline, may well fall outside of robust levels of regulation.	

Question	Your response
Question 6: Do you have evidence of functionalities that may affect how easily, quickly and widely content is disseminated on U2U services? • **Are there particular functionalities that enable content to be disseminated easily on U2U services?** • **Are there particular functionalities that enable content to be disseminated quickly on U2U services?** • **Are there particular functionalities that enable content to be disseminated widely on U2U services?** • **Are there particular functionalities that prevent content from being easily, quickly and widely disseminated on U2U services?** Confidential? – N	
<u>Enable: Do you have evidence of functionalities that enable content to be disseminated easily, quickly and widely on U2U services?</u> We would like to raise concerns around users-to-users services without basic safety functions, e.g.: • Whether or not safety settings are on by default • Whether the platform has a 'block' function: It is worth noting that Elon Musk announced the removal of the 'block' function on X, the platform formerly known as Twitter, yet such a tool is deemed necessary for user-to-user platforms in guidelines for Google Play and the App store. In 2020, Glitch and the End Violence Against Women Coalition's Ripple Effect Report into gendered online abuse during the first national lockdown found that the most common behavioural change from survey respondents on the impact of experiencing online abuse was blocking (76% of respondents), followed by 64% muting, 53% reporting. The 4th most common response was spending less time online - with 48% of Black and minoritised respondents and 41% of white respondents indicating this approach. Twitter/X's move to remove the block function is deeply concerning. On the contrary, Mozilla has recently claimed a campaign win regarding changes from Slack for a 'hide messages from another member' feature to address what Mozilla calls the lack of the "most basic of messaging features: a block button". Other important functions include:	

Question	Your response
• **Reposting functionality** used to generate harassment for example in Twitter 'pile ons'; the functionality to forward messages easily has also been linked to the spread of harmful disinformation on WhatsApp • **Whether platforms restrict who can see and interact with posts** and whether the user has any control over that • The **reporting mechanism and approach** - whether posts are deleted, reach is reduced through algorithmic changes, whether there is an appeals process • **Trending themes**: for example we saw in the case of #BBCPresenter that not only were presenters who were not involved in the scandal subjected to libel but the presenter in question was effectively 'outed' on social media through this function. A former Twitter employee spoke in the press about how this was a result of the systems that should have been in place to stop this from happening were not functioning correctly, possibly due to the heavy tech staff layoffs that the company had experienced. - **Responsible recommendation systems and content moderation**: including algorithms that promote the amplification of hateful or illegal content based on the level of attention that it is receiving. Responsible content moderation includes filtering harmful content but doing so transparently, e.g. publishing criteria for amplification/demotion and statistics on what content is demoted/promoted. This is to prevent abuse as well as "shadow-banning" - which is something Black creators have experienced disproportionately, and something explored by Rakin Creative in THE ~~UN~~SEEN project that documents censorship being applied inequitably, protecting those with large audiences but enforcing shadow bans, content removals, promo bans and full account removals on smaller creators. **Prevent: Are there particular functionalities that prevent content from being easily, quickly and widely disseminated on U2U services?** **Block function**: While many may consider the 'block' function as standard safety protocol, until regulation it is within the remit of company decision makers to go against commonly agreed guidance, thus increasing the risk of harm on platforms at short notice. **Direct Message functions**: Instagram is trialling new direct messaging features that would change the functionality of direct messaging to text-only DM requests (rather than video, photo or voicemail/calls) from someone you did not follow. This consent-based approach involves a step that sends an invite to get permission to connect with someone. New changes to

Question	Your response

Instagram's direct messaging functionality may have a dramatic impact on the level of gendered abuse that women and others receive on the platform - the specific changes are a systematic change that will drastically decrease the aberrant levels of unsolicited cyberflashing images that have been reported widely in the media recently, including in the BBC documentary by Emily Atack (Emily Atack: Asking for It?), featuring Glitch's founder and CEO Seyi Akiwowo - a behaviour common in the daily lives of many women and girls in the UK.

Nudge tactics and **adding friction (**a term used often by 'Facebook Whistleblower' Frances Haugen when given evidence to legislatures, including the UK Parliament) into U2U systems has been reported to slow down and even prevent content from being easily, quickly and widely disseminated on U2U platforms. For example:

- nudge notifications that encourage users to pause and reflect - for example nudges that question whether a user wants to post a message that appears to be harmful, and whether it will meet certain strict rules (as is the case with some subreddits)
- feature changes that stop service users from reposting links such as new articles that they have not read – which has an impact on the proliferation of disinformation
- users embracing a chronological timeline

As highlighted in the draft violence against women and girls code of practice, drawn up by Glitch and sector colleagues at Carnegie Trust, the End Violence Against Women Coalition, Refuge, 5Rights, NSPCC and Profs Lorna Woods and Clare McGlynn:

- 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. The provider should consider how this might affect a person's right to receive or impart ideas.

Many providers aim to ensure communication is as frictionless as possible, which means that people can share content even without opening it and therefore not considering the content (and similar points may be made about 'like' buttons and similar features). These features support the virality of certain sorts of content. This is potentially problematic given the bias towards content expressing discriminatory or abusive content. Regulated services should therefore consider the constitutive role of these features in the spread of VAWG-related content.

Design choices and product functions of online services can facilitate the escalation and amplification of content that may be seen by millions of other users in a short space of time. This could be through user-to-user reshares or via algorithmic amplification.

Question	Your response
Question 7: Do you have evidence relating to the relationship between user numbers, functionalities and how easily, quickly and widely content is disseminated on U2U services? Confidential? – N	
Glitch's research The Digital Misogynoir Report: Ending the dehumanising of Black women on social media looks at both mainstream platforms with high numbers of users and much smaller, high-harm platforms. The report looks at Facebook, which has 3.03bn monthly active users, compared to the estimated active membership of 100,00 on Gab. The interconnected nature of discourse moving from the small, high harm platform to the mainstream is important, as shifting narratives around what is acceptable societal speech is practised in the margins and enacted, and distributed to the masses on the larger platforms. Since this research has completed, Meta have amassed more than 30 million new users within the first day of launching their new platform, Threads, which stresses the importance of speed when it comes to a regulator reacting to the pace of the shifting technological landscape. The functionality of not only the larger platforms, but also the smaller platforms is important when it comes to reacting to this interconnected nature of discourse and harms across social media platforms though in relation to high-reach, fast dissemination of harmful content, the robust nature of larger social media platforms cannot be understated, and particularly in relation to not only policies and terms of service clearly stating what is and isn't permissible on platforms, but also enacting those policies in a way that demonstrates to both perpetrators and victims of abuse that perpetrating harms on platforms will lead to clear consequences and efforts to minimise and eradicate harm before it reaches potential victims will be enacted. The Digital Misogynoir Report examines almost one million posts concerning women across five social media platforms: Instagram, Facebook, Twitter, Gab and 4chan. The research found that hateful rhetoric and jargon is trickling from the alternative platforms (the small, high harm platforms Gab and 4chan) to the mainstream platforms (Twitter, Instagram and Facebook). We found that misogynoir underpins hateful narratives like white supremacy, antisemitism and great replacement theory. The key findings were that: ● On mainstream social media (Instagram, Twitter, Facebook), we mostly found stereotyping, body-shaming and fetishising. In contrast, on alternative platforms Gab and 4chan, we found more white supremacist and antisemitic themes	

Question	Your response
	• Hateful jargon from alternative platforms - like 'gorillion' and 'globohomo', and conspiracy theories like 'the great replacement' - are steadily trickling into the mainstream • Even abuse aimed at white women is often based on demeaning other races e.g. racist vitriol against mixed race couples in which white women are seen as 'betraying the white race' The interactions between platforms is also an important element as to how harmful they are. For this reason, we call for legislation and regulation that holds tech companies accountable for misogynoir, including comprehensively including smaller platforms in online harms, with a particular focus on platforms which act as a conduit to white supremacy. Glitch also conducted research in partnership with the End Violence Against Women Coalition in 2020 to document the increase in online abuse during the first UK Covid-19 lockdown: the Ripple Effect Report, which looked at online gender based violence experienced on Facebook, Twitter, Instagram, Snapchat, Zoom, Slack, Microsoft Teams, Google Hangouts, ASANA, Email, Facebook Messenger, other messaging services such as WhatsApp and iMessage. The research documents that 92% of survey respondents increased their internet use during the pandemic, and the unplanned nature of the increased time online, which in many cases also included the switch to a virtual-office, increased levels of online abuse with 46% saying they had experienced abuse since the beginning of the pandemic and 29% reporting it was worse than before Covid-19. While most abuse took place on mainstream platforms (65% of respondents experienced abuse on twitter, 27% on Facebook and 18% on Instagram), other platforms such as Zoom, Email, MS Teams, Hangouts and Slack - predominantly associated with the virtual office - were recorded in the research as sources of online abuse for all women. Further details of the 27 recorded types of abuse can be found in the report (p. 25), which highlights the way that work platforms can be co-opted by abusers and platforms proved unprepared for dealing with gender based violence. This will remain relevant for Ofcom to consider both in relation to past incidences of online harm and in response to the fast pace of digital development with emerging and future circumstances that platforms - old and new - may be under-prepared for.
Question 8: Do you have evidence of other objective and measurable factors or characteristics that may be relevant to category 1 threshold conditions? Confidential? – N	

Question	Your response

While the Online Safety Bill amendment for a VAWG Code of Practice was not passed in the House of Lords, we believe elements related to violence against women and girls should be incorporated into the Ofcom drafted guidance on the protections of women and girls - and should ensure that online gender-based violence is factored into the categorisation threshold of category 1 platforms.

Likewise, we wait to see how Baroness Morgan's amendment to the language of categorisation (245) as passed at Report Stage in the Lords - which makes categorisation decisions for Ofcom relating to Category 1 services based on either size or functionality, rather than both - ensures that smaller, high-harm platforms, like Gab and 4chan - as discussed in Glitch's research The Digital Misogynoir Report are captured.

As highlighted in the draft VAWG Code of Practice**,** some features may not be derived from a formal business relationship but be through third party independent software such as services that allow a user to post to multiple social networks. Social media providers should also consider the risks of harms arising from VAWG arising from such software. For example in relation to:

- deep fake or audio-visual manipulation materials
- nudification technology.
- other new technology
- bots and bot networks
- content embedded from other platforms and synthetic features such as gifs, emojis, hashtag

Each service is designed to allow and incentivise a user to create content in a different way. How content creation is designed can affect the risks of VAWG being created and disseminated. Features such as metrics or financial incentives based on popularity should be considered in relation to the motivation(s) of the creator. Outrage and content that plays on the biases of users (including sexism and misogyny) seemingly drive engagement (as clickbait headlines show), and there is a risk of cycles of ever increasingly outrageous content to drive likes and upvotes

Accessible and transparent user mechanisms must be in place for adult users to also implement such features that protect them from exposure to harm. This could include:

- features to prevent the direct messaging of accounts that do not follow a user;
- messages from unknown contacts reviewed by moderators; and
- control features around who can search for a profile, what content is visible for example features

Question	Your response
which filter harmful content and words appearing, and how personal content can be shared or re-distributed online. A service may decide to introduce barriers to stop people sending unsolicited nude pictures without consent. This could include blurring the picture or stopping the message from being sent and warning the intended recipient. Service providers should consider how to ensure that their recommender features are auditable including considering and documenting the questions of what was considered when setting up the features and what the operation of the features show. In this, providers should pay particular regard to special guidance on algorithmic accountability and auditing.	
Question 9: Do you have evidence of factors that may affect how content that is illegal or harmful to children is disseminated on U2U services? • **Are there particular functionalities that play a key role in enabling content that is illegal or harmful to children to be disseminated on U2U services?** • **Do you have evidence relating to the relationship between user numbers, functionalities and how content that is illegal or harmful to children is disseminated on U2U services?** Confidential? – N	
As above, we believe the guidance for women and girls should be written in such a away as to ensure that a gendered lens is applied to the provisions around the dissemination of content that is harmful to children - Girlguiding's annual Girls' Attitude Survey annually documents the impact of harmful experiences of girls and young women online - where negative experience increase as the girls grow in age.	
Question 10: Do you have evidence of other objective and measurable characteristics that may be relevant to category 2B threshold conditions? Confidential? – N	

Question	Your response
While we are confident that Baroness Morgan's amendment on categorisation will broaden the scope of what the threshold for a category 1 platform will be, we are still very aware of issue around small, harmful platforms and their relation to other platforms, including category 1 platforms in terms of cross-posting and the migration of ideas from the margins to the mainstream - as highlighted in The Digital Misogynoir Report. As detailed in more depth above, the draft VAWG Code of Practice highlights that some features of category 2B platforms may not be derived from a formal business relationship but be through third party independent software such as services that allow a user to post to multiple social networks.	
Question 13: Do you have evidence of other objective and measurable characteristics that may be relevant to category 2A threshold conditions? Confidential? – N	
While Glitch does not focus our research on category 2A platforms, we are aware of issues around **autosuggest** and **autocorrect functionalities.** For example, as in the #BBCPresenter case included above, with category 2A, there were issues in relation to autosuggestions where the presenter who was at the heart of the scandal was 'outed' by search services autosuggesting his name when BBC presenter was inserted into search, as highlighted by The Guardian journalist Chris Stokel-Walker. Autosuggest has also been flagged as highly problematic in relation to hateful content, including antisemitic content, as raised by the Antisemitism Policy Trust both in their briefing on the House of Lords Stages of the Online Safety Bill which highlights the issue and that changes to Google and Bing's algorithms can reduce antisemitic searches - and more broadly in their work, including this briefing paper 'Hidden Hate: what Google searches tell us about antisemitism today' (2019). As highlighted in the draft VAWG Code of Practice we believe that reporting features for problematic autocompletes should be clearly visible and easy to use. Where problems arise, providers should verify that the issue is solved. See further Guideline 5 on complaints. Some of these problems can be avoided if service providers are clear about their values and ensure that their recommendation and curation features embody those values Autocorrect can also rely on biassed AI technology to make problematic autocorrections.	

Please complete this form in full and return to os-cfe@ofcom.org.uk.