WARNING: This consultation response contains language and/or material that may be distressing

Your response

Question	Your response
Question 1: Do you have any comments on our proposed approach to 'content and activity' which 'disproportionately affects women and girls'?	Confidential? – No *Definition of women and girls* We emphasize the necessity of broadening the scope of anti-TFGBV initiatives beyond solely cisgender women and girls. Experienced trust and safety professionals have observed, based on their practical expertise, that men and boys, as well as transgender individuals and other members of the LGBTQIA+ community, encounter gender-based violence online. Extending the purview beyond cisgender women and girls is a recommended practice to ensure a more secure experience for all users. *Definition of harms* To ensure platform accountability and transparency, we recommend that Ofcom defines the harms in this Guidance in terms of user experiences, rather than relying on platforms to determine harm as those definitions will vary platform to platform and be difficult to compare. Standardized metrics should be established to measure instances of users encountering unwanted advances, misuse of their images, or other forms of harassment. Additionally, it is essential that such metrics are comparable across various platforms and over time. In the context of this Guidance, we urge Ofcom to provide a definitive definition of "misogyny," derived from the user-reported experiences, rather than relying on internal platform definitions. This will ensure that platforms are effectively targeting and mitigating a meaningful definition of "misogyny" within the broader context of "online misogyny," as laid out in the Guidance. Establishing a precise definition of "misogyny" would prevent platforms from engaging in either under- or over-moderation, both of which risk alienating particular user demographics. One Integrity Institute member warned that over-moderation of general comments and an overly broad application of the "misogyny" classifier to user conduct could inadvertently drive users to more

Question	Your response
	extreme online platforms. As a result, we recommended that Ofcom carefully consider the varying risk profiles when defining broad content classifiers like “misogyny” while acknowledging that leaving the definition for platforms to decide will lead to varying interpretations and potentially inconsistent reporting. *Pile Ons* We are pleased with Ofcom’s emphasis on pile-ons as this is a meaningful issue facing women in the public eye. We encourage Ofcom to continue to consider harms that are user sided, such as pile-ons. This speaks to the importance of specialized, targeted measures to protect women and girls who are the subject of public scrutiny, such as politicians, journalists, celebrities, and other public figures. We encourage the inclusion of other forms of user-sided harms such as doxxing, spread of deceptive synthetic media, online impersonation, and cyberstalking.
Question 2: Do you have any comments on the nine proposed actions? Please provide evidence to support your answer.	Confidential? – No Generally, TFGBV is treated as an issue of spiked activity rather than a systemic issue that is exacerbated by the design choices of social media platforms that allow small groups of overly active users to disproportionately harm others. We encourage Ofcom to continue treating TFGBV as a systemic presence on platforms and to apply mitigation measures such as expedited reports for chronic harassment or the design suggestions included below. We also urge Ofcom to consider the issue of re-traumatization when platforms investigate reports of harm, and recommend outlining a survivor-centric approach for platforms to adopt. Ideally, better designed platforms would minimize TFGBV on the whole, rather than relying on reactive reporting. It is imperative that TFGBV be recognized not merely as isolated incidents of heightened activity, but rather as an endemic issue within social media platforms. We encourage Ofcom to continue regarding TFGBV as a systematic presence on platforms.

Question	Your response
Question 3: Do you have any comments about the effectiveness, applicability or risks of the good practice steps or associated case studies we have highlighted in Chapter 3, 4 and 5? Are there any additional examples of good practices we should consider? Please provide evidence to support your comment.	Confidential? – No **<u>Action 1</u>** In response to the proposed establishment of an oversight mechanism for trust and safety decisions, we emphasize that many of these platform-wide anti-TFGBV measures are not always decisions made solely by trust and safety teams. We caution Ofcom in suggesting this mechanism as it could actually be a burden to the trust and safety teams internally. Per Ofcom's guidance, if an oversight mechanism were to be created, it should have a <u>broader purview</u> across governance and design decisions that impact TFGBV on the entire platform, not just over the trust and safety teams. Platform governance is cross-functional and, if internal oversight is needed, the trust and safety team should be empowered to step into that role. This is because trust and safety teams can support other internal teams (growth, legal, engineering) that often have conflicting goals to ensure that combatting TFGBV and other harms is not overshadowed by other competing priorities. In summary, to ensure greater protection for women and girls, enhanced transparency in decision-making processes across all parts of the organization - not just trust and safety - is necessary. **<u>Action 4</u>** *Recommender Systems* Many influencers in the "manosphere" who implicitly encourage TFGBV benefit from recommendation systems that drive followers to their extreme messaging. We recommend that Ofcom expand their scope when evaluating recommender system testing. Platforms should evaluate whether their algorithmic choices and systems are likely to increase the production of and user exposure to "harmful," "borderline," or "violative" content, not just illegal content. We also hope that the

Question	Your response
	"additional safety metrics" collected by platforms (in alignment with the Foundational Steps under Action 4) will be made available to external stakeholders. *Red teaming* Regarding the use of red teaming, we recommend that platforms engage in red teaming with cross functional stakeholders within the company. This means incorporating stakeholders from other teams–like policy, legal, trust and safety, user experience researchers, and product managers–as well as the standard participants from engineering and AI safety teams. We also encourage platforms to red team with external stakeholders, so that the exercises incorporate real world user experiences and external data. Upon completion of red teaming exercises, the associated data, including the foundational data and operational assumptions, should be shared with the risk assessment teams. This information should then be integrated into the broader operational and strategic frameworks of the company, as well as trust and safety work and compliance protocols. Ofcom should request information about the methodology and sample size of the raw data used in red teaming. This will reveal the rigor of the red teaming process and whether it adequately addressed the harm in question. Researchers and civil society should be able to access, as appropriate, this red teaming information. For example, researchers should be able to know what internal teams receive the red teaming results and which teams are involved in the exercise. *Personas* Specifying the use of 'personas' for designing safe experiences is not recommended as this is a highly specific tactic that is not universally applicable across platforms. 'Personas' are merely one of many methods that identify common behaviors and desires of the user. Different situations warrant different responses, such as 'Jobs to be Done' (JTBD) and user journey maps. Additionally, 'personas' can often have their own set of

Question	Your response
	biases, as they often simplify a complex audience to an average target user. Research on negative user journeys and bad actor user journeys is crucial and currently underutilized. Ofcom can address the user journey more effectively by requiring such research. This is already a best practice in the field of user experience research. It remains unclear whether the public and civil society will gain transparency into these processes as a result of the Guidance and reporting required by Ofcom. We strongly recommend that the public and civil society have access to these processes, as appropriate. **<u>Action 5</u>** We agree with Ofcom's acknowledgement that privacy defaults are key. However, many of the settings outlined in the Guidance are deemed only appropriate for children, even as features like "Locked Profile" have proven their utility for protecting adult women. Given the volume of women in emerging markets who are likely to join online services, but are unlikely to understand how to change their default settings, having mandatory safe privacy defaults for adult women is key to keeping women safe online. *Customization* "Setting stronger and customisable defaults around interactions, privacy and geolocation" is a strong guideline for safer platforms (noted on page 28 of the consultation). To further strengthen this point, we recommend including that location services should be set to off by default, and, once toggled on, there should be periodic reminders that location services are still on. This has important implications in situations like stalking and IPV. Customization is a very useful tool for users to improve their online experiences. However, Ofcom should be conscious of platforms engaging in "design washing," which means platforms use customization as a catch-all

Question	Your response
	for platform design issues, often without promoting awareness to the user of such options. *Bundling* While bundling allows users more control of their experience, these options are often nested throughout the platform (pop-ups upon opening the app, the account settings page, the privacy tab, etc) making it intentionally difficult for users to opt in. Platforms can gradually introduce default privacy options as users start to engage with features for the first time. To easily access these settings in the future, platforms should keep privacy settings on one page, not dispersed in bundles throughout the platform. Similarly, when a platform introduces new features, the users should be prompted to opt-in or opt-out of the corresponding privacy settings. This nuance is important because defaults for *new* features are not defined by the Ofcom Guidance at this stage. In summary, onboarding users to all privacy settings at once can be overwhelming and cause users to skip through important options. Instead, platforms should be more intentional and slowly introduce default privacy settings as people use features for the first time or introduce the settings as a guided tutorial. These settings should be easy for users to locate. *Good examples of default settings* ● Instagram's Privacy Defaults for Minors: Demonstrates the efficacy of defaulting to safer settings for vulnerable users. ● Real-Time Visibility Snapshots: Provide users with tools to see how their information appears to others, similar to Facebook's "View As" feature. **Action 6** *Recommender Systems* Moderation has failed to reduce the "circulation of content depicting, promoting, or encouraging online

Question	Your response
	gender-based harms." Any limit defined by policies will allow for harmful, non-violating content to remain prominent while simultaneously allowing misogynists to use "censorship" as a way to attract even more attention. Instead of focusing on moderation, platforms must "take responsibility" for the incentives they create for misogynists or harmful actors to attract disaffected men through extreme messages that naturally attract engagement. Engagement based algorithms are inherently vulnerable to such tactics and moving away from engagement based algorithmic systems is essential in any true "safety by design" system. Ofcom correctly identifies recommender systems as a key factor in TFGBV. However, reducing the visibility of harmful content also necessitates examining other platform features that amplify it, such as the reshare button and post-quoting (for example, "quotes" on X) to ensure their safe and responsible use. It also requires transparency of algorithmic choices such as the promotion of content that is likely to be shared or commented on, as such algorithmic choices have been shown to lead to harmful content exposure on platforms. Overall, negative signals to reduce the circulation of harmful content should not be based only on reporting violations, but rather based on the subjective negative judgment of those who have negative experiences. *Rate Limits* Misogyny is often perpetrated by an active, but small group of perpetrators that promote "content with societal risk, targeting others, and misusing other people's images and information" (USC Neely Center Design Code). To combat these bad actors, Ofcom should include rate limits as a mitigation measure. Rate limits, especially for new users, are a key tool to reduce the circulation of harmful content. Imposing rate limits on actions such as friend requests, comments, and messages for new or unverified accounts prevents the exploitation of platform features for harm. Newly created or unverified accounts are often used for

Question	Your response
	spam, harassment, and coordinated abuse campaigns. Specific actions platforms can take include: ● Implementation Structured Rate Limits: Restrict actions for new accounts until they demonstrate authentic behavior through verified activity or longevity. ● Threshold Escalation: Flag accounts that exceed activity thresholds for additional review. ● Gradual Unlocking: Increase interaction limits as accounts gain credibility. Without restrictions, malicious accounts can overwhelm victims with harmful interactions. These limits serve as preventive friction and reduce misuse and abuse by bad actors and bots while creating safer online environments. *Nudges* Platforms can employ machine learning models to detect harmful language in context and trigger tailored nudges. These models can provide a basis for language detection and nudging that can be customized by platforms based on user feedback and contextual needs. Applying nudges across posts, comments, replies, and direct messages can ensure comprehensive coverage. Using A/B testing to evaluate the effectiveness of nudges can optimize them for diverse user demographics and cultural contexts. *Good Examples of Recommender Systems* ● Pinterest: Pinterest does not rely solely on engagement signals for ranking and recommendations. Pinterest uses in-app surveys (where users can directly give feedback about the platform) and independent assessments of content quality (usually generated by manual labeling). ● LinkedIn: Prioritizes content relevance and professional value, using quality metrics to rank posts. ● YouTube & Google Search: YouTube promotes watch time over clicks, incentivizing creators to focus on informative, engaging content. In 2024, YouTube reported that its systems are “trained to elevate authoritative sources higher in search results, particularly in sensitive contexts” –

Question	Your response
	mitigating risks while optimizing high quality information in its ranking framework. Google Search predicts quality using a wide variety of signals, including long established information retrieval signals (the most famous being PageRank, Google's founding algorithm). As a result, users get results from trusted medical organizations and other authoritative sources when using Google Search, especially around sensitive topics. Replacing engagement-based ranking with trust-based models offers a clear pathway to reducing the amplification of TFGBV while maintaining platform integrity. *Good Examples of Rate Limits* ● Reddit's Post Restriction Tools: Communities on Reddit can limit posting frequency for new users to reduce abuse and spam. Similarly, under Reddit's Karma System, certain subreddits can choose to automatically remove new posts from users who haven't met specific engagement criteria, even if the content isn't spam. ● X's Rate Control Systems: Prevents misuse of features like direct messaging and follows. ● Instagram's Comment and DM Limits: This feature enables users to restrict comments and DM requests during periods of heightened attention. It helps protect individuals from potential abuse by automatically hiding comments and messages from users who don't follow them or have only recently started following them. *Good Examples of Nudges* ● Snapchat: The expanded in-app warning feature warns teens when they are receiving messages in chat from someone who has been blocked or reported by others, or is from a region where the teen's network isn't typically located – signs that the person may be a scammer or otherwise suspicious.

Question	Your response
	• Instagram: Instagram has reported that, over the course of one week it sent approximately one million nudges to users, 50% of the time users deleted or amended their comment as a result. The reduction in hurtful comments posted is also long lasting, according to Instagram's research on what it calls "repeat hurtful commenters" — people who leave multiple offensive comments within a window of time. • X: In 2022, Twitter displayed embedded prompts for users to reconsider potentially harmful tweets before posting. Matt Katsaros studied the effect of this feature and found that 9% of users decided not to post, and 22% edited their tweet. More importantly, the nudge had a lasting impact, as recipients were less likely to post offensive content in the following weeks. "The nudge changes the behavior in the moment, but more importantly, it has a lasting impact. People are more likely to rethink their approach in future interactions," explained Katsaros at a 2024 Symposium on Comment Section Research & Design. • X has also reported that nudging users to reconsider replies containing harmful language "resulted in people changing or deleting their replies over 30% of the time when prompted for English users in the U.S. and around 47% of the time for Portuguese users in Brazil." **Action 7** Filters give users control over their online experiences by enabling them to block specific keywords, topics, or other triggers. These tools reduce exposure to harmful content and help users create safer, more positive environments. Filters do not just hide harmful content but allow users to define the boundaries of their online interactions, creating a protective buffer against TFGBV and other potential harms. Because filters do not prevent a user from posting, but rather are used to tailor the visibility of content, they do not run afoul of free speech considerations. By creating safer environments, filters

Question	Your response
	also encourage greater participation from users who might otherwise disengage due to harassment. One useful filter is a sensitive content quarantine, which establishes a separate dashboard where filtered content is stored, mirroring email spam folders. We recommend that Ofcom encourage platforms to launch awareness campaigns around available filters, so that users are educated about the tools at their disposal during onboarding–beyond the standard privacy settings. Platforms can also send periodic prompts to ensure widespread adoption of these mechanisms. *Good Examples of Filters* ● Instagram's Hidden Words Feature: One year after the Hidden Words feature launched, users with large followings (10,000+ followers) saw 40% fewer comments that might be offensive after turning on the feature. It effectively mitigates harmful interactions by empowering users to define their personal boundaries. ● Perspective API and Coral offer a toxicity filtering for comments sections, aimed at fostering a healthier online discourse. FACEIT, one of Europe's largest gaming platforms, experienced a 20% decline in toxic messages since employing Perspective API's filtering. ● Discord introduced a new Ignore feature that allows users to ignore rather than block other users. The feature "allows a person to hide any new messages, DMs, server notifications, profiles and activity from selected users without alerting them...In practice, DMs received from an Ignored person will appear in the inbox with an icon and a grayed-out name, so they are available if the ignorer chooses to look at them." ● TikTok offers a feature that allows users to reset their 'For You' feed, providing an opportunity to start afresh and tailor content recommendations to their current preferences. This action resets the feed to display a new set of popular videos, similar to the experience of a new user, and as users interact with these videos, the platform's algorithm begins to personalize the feed based on the new interactions.

Question	Your response
	• X's Safety Mode helps reduce unwanted interactions by temporarily blocking accounts that use potentially harmful language or repeatedly send uninvited replies and mentions… • Bluesky's Stackable Approach to Moderation: Users can install different moderation services onto the platform to filter out specific content they do not want to see, from disturbing images to harmful language. If a piece of content has evaded the moderation service, users can report to that service directly from their Bluesky feed. **Action 8** Platforms should build or integrate existing user-friendly reporting systems that provide feedback on abuse reports and enable users to document and store evidence easily. These measures improve trust and accountability while empowering victims to seek resolution. Beyond allowing users to track reports, we recommend that users be able to view all submissions in a central dashboard where users can report abuse, track submission status, and view outcomes. Platforms should allow users to engage in automatic evidence capture–users should have the capacity to save flagged content in encrypted formats, including metadata and timestamps, for use in legal or institutional actions. Platforms should adopt interoperable documentation tools, enabling users to collect and share evidence of abuse—such as screenshots and metadata—across multiple platforms. These measures reduce retraumatization, empower victims to take action, and encourage platforms to improve cross-platform accountability for abuse. *Good examples of reporting procedures:* • Instagram's Reporting System offers contextual guidance and feedback for flagged content. • Instagram Report Status: Allows for users to see the status of the reported content or account

Question	Your response
	that they submitted to the platform and also allows users to see the full history of their past reports. ● Pirth.org: Allows users to report online threats across any social media platform. Upon submission, it instantly generates a personalized action and resource list, including digital safety tools, helplines, legal aid, mental health support, and more. With user consent, the Pirth.org team reviews reports and escalates threats to platforms, easing the burden on victims.
Question 4: Do you have any feedback on our approach to encouraging providers to follow this guidance, including our proposal to publishing an assessment of how providers are addressing women and girls' safety? Do you have any examples or suggestions of other ways we could encourage providers to take up the 'good practice' recommendations?	Confidential? – N/A
Question 5: Do you have any comments on our impact assessment, rights assessment, or equality impact assessment? Please provide any information or evidence in support of your views.	Confidential? – N/A
Question 6: Do you agree that our draft Guidance is likely to have positive effects on opportunities to use Welsh and treating Welsh no less favourably than English? If you disagree, please explain why, including how you consider the draft Guidance could be revised to have positive effects or more positive effects, or no adverse effects or fewer adverse effects on opportunities to use Welsh and treating Welsh no less favourably than English.	Confidential? – N/A

Please complete this form in full and return to OS-Section54@ofcom.org.uk.