BILETA's Response to Ofcom's Draft Guidance: "A Safer Life Online for Women and Girls"

This response is led by the **British and Irish Law, Education and Technology Association (BILETA)**. Formed in April 1986, BILETA aims to promote, develop, and disseminate high-quality research and knowledge on technology law and policy to organizations, governments, professionals, students, and the public. BILETA also encourages the use of and research into technology at all stages of education. This response was prepared, approved, and funded by BILETA, with contributions from Professor Kim Barker, Reader Edina Harbinja, Miss Ellie Colegate, and Dr Felipe Romero-Moreno.

Response to Question 1: Comments on Proposed Approach to 'Content and Activity'

The Draft Guidance identifies four categories of harm that disproportionately affect women and girls, collectively referred to by Ofcom as 'online gender-based harms'. This term seems to be selected and used over broader, established labels including e.g., VAWG (Violence Against Women and Girls) online or 'technology facilitated gender-based violence (TFGBV) which are the more common acronymic labels adopted in other contexts.

The reduction in women's and girls' participation in public life, democratic debate, and economic opportunities is recognized as a consequence of such abuse. This has led to recognition at high levels that online violence against women and girls (VAWG) is an obstacle to gender equality and a violation of women's human rights. The Council of Europe's Istanbul Convention (2011), an international treaty on violence against women signed and ratified by the UK, defines violence against women, including psychological abuse and harassment, as "a violation of human rights and a form of discrimination against women". The Convention requires states to take measures to prevent and address such violence, whether it occurs in private or public life. While the Istanbul Convention was drafted before social media became widespread, its principles apply online, and its monitoring body (GREVIO) has stressed that Parties must address "expressions of violence against women… perpetrated with the help of technology" as part of their obligations. Similarly, the UN Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) and its General Recommendations have stated that gender-based violence, including digital forms, is discrimination that states must address. It is therefore particularly good to see that there is recognition and capture of these forms of violence within the proposed four categories in the draft guidance.

The Ofcom approach intentionally looks at women and girls' online experiences "in the round," acknowledging that gendered harms cut across the legal categories used in the Online Safety Act (illegal content, content harmful to children, etc.) Indeed, certain illegal harms (like many forms of harassment, stalking, or threats) and certain types of content harmful to children (such as misogynistic content that could be seen by teenage girls) intersect in the lived experience of women and girls online, so it is appropriate that these are explicitly captured here.

Notably, Ofcom limits the guidance to those harms "in relation to which [regulated] service providers have duties" under the Act. This means the focus is on user-to-user services and search services (Part 3 and 4 services under the Act) and on content either illegal or harmful to children. Ofcom rightly acknowledges that girls face particular risks of grooming and sexual abuse online but proposes to differentiate between those (illegal) risks, and others here that fall within the proposed four categories. This approach means it will be important for Ofcom to ensure that insights about gender (for instance, the fact that girls are disproportionately targeted in certain illegal activities) inform the implementation of the illegal content and child protection codes as well.

The guidance echoes this by treating online harms as part and parcel of the same continuum as offline abuse. It is important that this is recognised specifically within the guidance, because the digital sphere amplifies and extends the reach of misogyny. As the UN Special Rapporteur on Violence Against Women observed, the internet is operating "in a broader environment of widespread and systemic structural discrimination and gender-based violence against women and girls." Ofcom explicitly notes that women and girls who have multiple protected characteristics – such as women of colour, LGBTQ+ women, or those with disabilities – often face "additional harms" online – another significant point but one which is not necessarily specifically captured within the four categories. Ofcom's draft guidance, and the four categories rightly indicate that women and girls are diverse and that some face greater risks online. These harms (misogyny, harassment, sexual abuse, etc.) can be compounded by other prejudices like racism, homophobia, transphobia, ableism, or age discrimination. That said, the current proposed guidance in relation to the content and activity seems to group women and girls together as one group.

In conclusion on Q1, BILETA finds that the proposed approach is an improvement of the previous lack of consideration of specific harmful content that particularly affects women and girls, and which poses threats to their online participation. Ofcom's framing of the "content and activity" of concern is appropriately expansive and evidence-based. We consider that this is a step in the right direction and offers a useful position from which to start, albeit with the caveat that while there is progress here, there are still aspects here that are not captured. Moreover, the fundamental capture across the four proposed categories of 'women and girls' does not differentiate between the different groups, even within this. It will be important to track how these four categories at the heart of the approach are addressed and followed across the first 18 months of implementation of the guidance, especially in terms of the impacts on the different groups within 'women and girls.'

Response to Question 2: Proposed Actions (1–9)

Ofcom's draft guidance identifies nine actions for online service providers, grouped under three pillars – Taking Responsibility, Preventing Harm, and Supporting Users. We address each action's relevance, sufficiency, and practical impact, with particular regard to providers' human rights obligations (especially freedom of expression and privacy).

Action 1: Ensure Governance and Accountability Processes Address Gender-Based Harms

Assessment: This action is highly relevant. It calls for a top-down commitment to women and girls' safety, integrating gender-based harm prevention into corporate governance. We agree that leadership accountability and diverse representation in decision-making are essential, as evidence shows tech firms have been slow to tackle online misogyny absent clear internal accountability. Measures like dedicated policies on misogynistic content, consulting subject-matter experts (including women's safety NGOs) in drafting terms of service, and training staff on online gender-based violence can embed a safety-by-design culture. This aligns with the UN Guiding Principles on Business and Human Rights, which urge companies to conduct human rights due diligence (including a gender lens) and commit at the highest levels to respect rights. It also reflects The Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) call for institutions to take violence against women seriously in all arenas (including online).

Human Rights Considerations: Governance changes must be implemented consistently with freedom of expression and privacy duties. The Online Safety Act 2023 requires all services, when deciding on safety measures, to "have particular regard to" users' rights to freedom of expression and privacy. We therefore encourage that any new content policies under Action 1 define prohibited "online gender-based harms" clearly (e.g. distinguishing illegal threats from merely offensive speech) to avoid overreach that could chill legitimate expression. Likewise, training and oversight should emphasise *consistent*, proportionate enforcement, protecting women from abuse while not unduly censoring debate. We commend the draft guidance's suggestion of an *oversight mechanism* (such as an external appeals ombudsman or independent experts) to review content decisions related to abuse. An external review process can enhance accountability and provide due process for users, ensuring that mistaken removals or biases can be corrected. This will help balance women's safety with freedom of expression by preventing excessive takedowns and giving users a remedy. Privacy should also be respected in governance measures: for instance, any data collected for diversity or accountability purposes must comply with UK data protection law. In sum, Action 1 is a strong foundation, it is necessary and, with proper implementation, compatible with human rights obligations. The key will be genuine commitment (avoiding a "tick-box" approach) and ongoing evaluation of how these governance reforms impact online discourse and user trust.

Action 2: Conduct Risk Assessments that Capture Harms to Women and Girls

Assessment: We find this action to be relevant and necessary. Risk assessments are a core duty under the OSA, and Ofcom rightly emphasises making them gender sensitive. Historically, women's online experiences (e.g. volume of harassment, subtle forms of coercive abuse) have been overlooked in platform risk analyses. By requiring platforms to specifically evaluate risks of online misogyny, harassment, domestic abuse, etc., Action 2 ensures these issues are not subsumed under generic categories. This aligns with emerging international standards, for example, the EU's Digital Services Act (DSA) now *mandates* that very large platforms assess "actual or foreseeable negative effects… on gender-based violence" as a systemic risk. Ofcom's guidance is consistent with this global trend of treating online gender-based harms as a distinct risk category requiring targeted mitigation.

The proposed steps (beyond the legally required "foundational" risk assessment duties) are practical and commendable. Using external expert assessors to monitor new threat patterns and engaging with survivors of online abuse to inform risk analysis, are highlighted as good practices. We agree, survivors and frontline NGOs can offer insights into how abuse manifests (e.g. coded misogyny, tactics to evade filters) which internal teams might miss. User surveys to understand women's safety preferences and experiences are another valuable tool. For example, a platform working with a survivor-centric organisation (like Chayn, noted in the guidance) used trauma-informed surveys to capture women's experiences of abuse, helping shape safer feature design. These approaches reflect a human-rights-based methodology: they centre the voices of those affected (consistent with CEDAW) and form part of companies' human rights due diligence (UNGP Principle 18 recommends consulting affected groups in impact assessments).

Human Rights Considerations: It is crucial that risk assessments not only catalogue harms to women but also evaluate potential *impacts of safety measures on rights*. Under the OSA, Category 1 services must assess how their policies might affect freedom of expression and privacy. We suggest that gender-focused risk assessments integrate this step: for instance, if a mitigation strategy involves automated filtering of "misogynistic language", the assessment should weigh the risk of algorithmic overblocking (and how that might silence legitimate speech, including women's own expression). Similarly, if addressing online domestic abuse leads to greater monitoring of private messages, privacy implications should be examined. By explicitly incorporating freedom of expression and privacy into gender-harm risk assessments, providers can choose interventions that maximize safety while minimising rights intrusions. We also caution that engagement with survivors should be handled carefully: participants' privacy and well-being must be protected (e.g. anonymizing survey data, offering support resources), and companies must avoid "ethics-washing" (i.e. consulting survivors without real follow-through). Overall, Action 2 is sufficient as a framework and its impact will depend on rigour. Ofcom should ensure that the forthcoming risk assessment reports meaningfully reflect gendered realities (for example, reporting the prevalence of gendered abuse incidents, not just aggregating all harassment). A transparent summary of findings could even be shared, which would tie into Action 3 on transparency and build public confidence that platforms understand and are tackling these specific risks.

Action 3: Be Transparent about Women and Girls' Online Safety

Assessment: We strongly support greater transparency as an enabler of accountability and user empowerment. Action 3 presses providers to disclose more information about how women and girls are affected on their services and what is being done in response. Evidence suggests transparency drives corporate responsibility and allows users and regulators to monitor progress. Key good practices mentioned include publishing data on the prevalence of different forms of abuse, outcomes of user reports, and how automated tools flag (or fail to flag) content targeting women. For example, a platform might report what percentage of misogynistic posts were detected by AI vs. by user reporting – such granular data would indicate where systems fall short and encourage improvement. This kind of openness aligns with the EU DSA's transparency regime, which requires detailed reporting from large platforms on content moderation actions and errors, and it mirrors the trend of *transparency reports* becoming industry standard. It also resonates with human rights principles: transparency is a prerequisite for the public and stakeholders (including civil society and academia) to evaluate whether companies are upholding rights like non-discrimination and free expression online.

Human Rights Considerations: Done correctly, transparency enhances the protection of rights. Publishing clear policies and enforcement statistics can help promote freedom of expression by showing users that restrictions are applied even-handedly and for legitimate aims (and revealing if they are not). BILETA cautions, however, against transparency that might expose personal data or enable abuse. The draft guidance wisely urges *caution in sharing personal data* when being transparent. For instance, if a company publishes case studies of abusive incidents, they must anonymize identities so as not to retraumatize victims or inadvertently "name and shame" individuals without due process. Another risk is that overly detailed disclosure of automated moderation rules could allow malicious actors to game the system. We advise a balanced approach: share enough information to illuminate the effectiveness and fairness of safety measures (e.g. percentages of content removed, median response times to abuse reports, existence of appeals), but not so much as to undermine those measures. Additionally, transparency should extend to explaining *user options*: women and girls should easily find information on how to adjust settings (tie-in with Action 7) or how to get help if abused. In sum, transparency is a powerful tool to ensure the other actions are genuinely working, it should be implemented in a privacy-conscious way and complemented by independent oversight (as per Action 1) so that data released can be trusted and verified. We find Action 3 relevant and likely impactful; its sufficiency will depend on robust metrics and consistency across the industry. Ofcom could encourage a standardised reporting framework for gender-based harm, possibly coordinating with international efforts.

Action 4: Conduct Abusability Evaluations and Product Testing

Assessment: BILETA considers this a crucial forward-looking action to *prevent* harm before it occurs. "Abusability testing" means proactively probing how features could be misused by bad actors. In our view, this is both highly relevant and practically impactful, as many online harms exploit unintended product loopholes. The guidance notes this is especially important for features that might be weaponized in contexts like domestic abuse or pile-on harassment. We agree – for example, a live location-

sharing feature could be misused by a stalker, or a group chat function might enable mass harassment if not properly moderated. By stress-testing features (through techniques like red teaming, where internal teams or external experts simulate attacker tactics), companies can identify and fix vulnerabilities early. Good practice examples include working with experts who understand perpetrator behaviours and using "personas" to model diverse user experiences. BILETA also notes that Ofcom references its own research on red teaming for generative AI misuse (e.g. deepfakes) and various academic studies on designing against abuse – indicating a solid evidence base for this approach. In short, Action 4 encourages a safety-by-design ethos: much like security testing is standard for preventing hacks, abuse-mode testing should become standard for preventing harassment and exploitation.

Human Rights Considerations: Abusability evaluations largely occur behind the scenes in product development, so they generally pose low risk to user rights. In fact, they are meant to *protect* users (especially vulnerable groups) proactively. BILETA emphasises that such evaluations should include perspectives of *diverse* users to avoid discriminatory design. For instance, teams should test whether algorithms flag slang or dialect used by women of colour as "abusive" at higher rates – a known issue where automated systems have misclassified Black women's speech as harassment. Including intersectional abuse scenarios (e.g. "misogynoir", the specific abuse faced by Black women) in red teaming will ensure mitigation measures do not inadvertently silence marginalised voices. The draft guidance indeed cites the need to account for intersectional harms, noting that treating categories like racism and misogyny in isolation fails to capture compounded abuse. We support this nuance and recommend it be built into all testing. Another consideration: smaller platforms might find comprehensive abusability testing resource intensive. The guidance acknowledges that full red-team exercises may be most feasible for high-risk, high-reach services. We suggest Ofcom develop or share lightweight tools (perhaps drawn from the *"Trust and Abusability Toolkit"* cited in the guidance) that lower the burden for smaller companies – ensuring Action 4's benefits reach a wider array of services. Overall, we believe Action 4 is sufficient and innovative. Its practical impact could be significant in reducing "downstream" harms – users may never experience certain abuses because they were anticipated and designed out. In regulatory terms, this is a *feasible* ask: it leverages companies' own technical capacity and creativity to solve problems, rather than imposing blunt external controls. We recommend Ofcom monitor and share examples of successful abusability fixes (for instance, if a platform prevents a new form of gendered abuse via design change, publicize this as a model for others).

Action 5: Set Safer Defaults

Assessment: BILETA strongly agrees that default settings should favour safety and privacy. Research confirms that many users (especially younger or less experienced users) do not change default settings, so making those defaults protective can "bake in" a baseline of safety. In the context of women and girls, *default configurations* can reduce exposure to abuse. For example, a social media platform might by default limit incoming messages to only friends, hide a user's online status, or obscure precise location data. The draft guidance cites evidence that women are more susceptible to abuse when features are permissive by default, and that stronger default privacy settings make it easier for users to keep themselves safe. We concur. A notable illustration is the problem of "cyberflashing" (sending unsolicited explicit images via features

like Bluetooth/AirDrop): Apple's recent move to tweak default AirDrop settings to limit contact from strangers was a direct safer-default solution to curb abuse. Similarly, the guidance's examples – such as automatically removing metadata (like location tags) from images uploaded or requiring explicit opt-in to location sharing – are sensible defaults that mitigate stalking or doxxing risks. Setting stricter defaults for interactions, privacy, and geolocation (with the option for users to loosen them if they choose) is a practical measure with immediate impact. It does not remove any content or capability; it simply starts every user off in a safer position. We also appreciate the emphasis on *customisability*: users should be able to adjust settings, but the onus shouldn't be on a vulnerable user to discover and enable a dozen safety features, the service should activate core protections by default.

Human Rights Considerations: Prioritising privacy and safety in defaults inherently supports the right to privacy (Article 8 ECHR) and can indirectly support freedom of expression by preventing certain harms (for instance, women who feel safer are more likely to speak out without fear of harassment). There is little risk to free expression here, since defaults typically govern *exposure* and *visibility* rather than banning speech. However, one possible concern is ensuring that safer defaults do not unduly *isolate* users or hinder their equal participation. For example, if a default hides all a new user's posts from public view, women seeking to build a public platform might remain invisible unless they know how to change that setting. Thus, we recommend defaults that protect against non-consensual intrusions (like unwanted messages, tracking, tagging) but do not silence users' outgoing expression. The guidance's suggestions seem well-calibrated: e.g. removing geolocation metadata by default (which protects users from inadvertent location leaks) clearly has no speech downsides; bundling easy "safe mode" settings for those experiencing harassment is another example that empowers without censoring. We also note that safer defaults should be attentive to **accessibility and inclusion** – some women (such as those with disabilities or less tech-savvy individuals) might need extra support to understand and adjust settings. Clear prompts and education (perhaps a quick tutorial on safety settings during onboarding) can enhance the effectiveness of Action 5. Overall, we find this action sufficiently addressed by the guidance, and its practical impact could be high, as even passive users will benefit. It is also legally sound: nothing in UK law prevents services from choosing privacy-protective defaults; indeed, data protection principles (GDPR/UK GDPR) encourage data minimisation and privacy by design, which aligns with this action. We encourage Ofcom to coordinate with the ICO as needed, to ensure consistency between safety by design and privacy by design principles.

Action 6: Reduce the Circulation of Online Gender-Based Harms

Assessment: This action addresses how to limit the *spread* and viral amplification of harmful content. It is a vital complement to removing content (Action 9) – even when content cannot or has not yet been taken down (for legal or practical reasons), platforms can often intervene to reduce its reach. Ofcom correctly notes that there is no "one size fits all" solution here and that a mix of approaches will be needed. BILETA considers this action relevant but also one of the more complex areas to implement in a rights-compatible way. Good practices highlighted include using nudges and carefully tuned automation to prevent content from going viral or being algorithmically promoted if it's likely to be abusive. For instance, platforms might tweak recommender systems to avoid promoting posts that resemble known misogynistic tropes, or limit

reshare functions on a post that's accumulating mass abuse (a tactic some platforms have explored during "pile-on" attacks). Another example is interstitial warnings or " click-through" screens on content that may be lawful but harmful – slowing down how easily such content circulates. The guidance cites an eSafety Commissioner Safety by Design report suggesting **hiding content pending review** if it's likely harmful. We find that suggestion practical: if a piece of content is flagged (by automated detection or user reports) as potentially containing gender-based hate or intimate abuse, temporarily suspending its visibility until a human can verify it prevents further harm in the interim. Many platforms already do this for suspected child sexual abuse material; extending it to egregious misogynistic content could be effective, provided review is timely. We also encourage the positive use of counter speech and education. For example, if certain misogynistic narratives are trending, platforms could inject authoritative content or feminist perspectives to dilute hateful narratives (though this ventures beyond the guidance's scope, it's worth noting as a "circulation" countermeasure that supports women's voice).

Human Rights Considerations: Reducing circulation implicates freedom of expression because it often involves deprioritising or constraining content that is not necessarily illegal. The OSA notably does not impose removal of legal content that is harmful to adults, but it does allow and encourage systems that give users more choice and control over what they see. Action 6's measures must be careful to respect this balance. BILETA agrees with Ofcom's balanced framing – the goal is to minimise harm while balancing users' rights. Measures like subtle algorithmic de-amplification or user prompts likely pass muster, as they do not outright censor content but nudge behaviour. For example, a prompt asking a user to reconsider posting something potentially misogynistic (one of the good practices mentioned) can reduce harassment without banning the speech entirely. However, we caution against opaque "shadow moderation." If content is systematically down-ranked or hidden, transparency (Action 3) becomes critical – users should have some awareness or recourse. The EU DSA, for instance, will require that recommender systems allow users to adjust algorithms, and that any content moderation decisions (including reduction of visibility) be communicated to the content provider with reasons. UK law will similarly expect fairness in how content is treated. Therefore, any automated limitation on circulation should be coupled with explanations or user control. One idea is to leverage user empowerment tools mandated by the OSA for Category 1 services: users can be given settings to opt-in to stronger content filters for hate or abuse. If a user (perhaps a woman who is often targeted) enables a strict filter, the platform could then aggressively down-rank likely abusive content from her feeds – an approach that respects individual choice. On the flip side, platforms should avoid paternalistically suppressing content *about* women's issues. For instance, discussions of sexual violence (for awareness or solidarity) might trigger filters for "graphic content" – a nuanced approach is needed so that attempts to reduce harmful content don't silence conversations *led by women*. Intersectionality is key: as noted in the guidance, treating misogyny in isolation could fail to detect content that targets women of specific races or sexual orientations. BILETA recommends testing circulation-reduction algorithms for unintended bias – e.g. ensure content by feminist activists or women of colour isn't disproportionately caught in broad throttling of "heated" threads. In summary, Action 6 is sufficient in outline and technologically feasible, but its implementation must tread carefully on free expression. It should focus on genuinely abusive or contextually harmful spread

(such as dog-piled harassment or non-consensual images) and be transparent and adjustable. If done right, reducing the virality of abuse will protect targets (often enabling *their* freedom of expression by preventing silencing through intimidation) while still allowing general discourse to flow.

Action 7: Give Users Better Control Over Their Experiences

Assessment: We endorse this user-empowerment approach. Empowering women and girls to curate their online experience is fundamental to a safer internet. The draft guidance identifies features like the ability to easily block or mute multiple accounts, to filter content recommendations, and to delete or limit the visibility of one's own past posts. These are practical tools that put power in the hands of users. For example, the concept of an "account safety mode" (bundling several protective settings) could allow a user facing a harassment campaign to quickly lock down her account (restricting comments, turning off public visibility, etc.). Some platforms have begun offering such modes or bulk-block features (e.g. Twitter's "Safety Mode" or third-party tools that block followers of a harassing account). The guidance's mention of retroactively changing visibility of content is also important. Women may share content at one time but later face abuse for it; giving them the option to mass-hide or delete old posts (or images) can reduce opportunities for abusers to exploit past information. Essentially, these controls recognise that context changes – what was safe to share yesterday might feel unsafe today after an incident of abuse, and users should have agency to respond. Additionally, greater control over algorithmic recommendations (choosing to see content chronologically or to mute certain keywords/topics) can help women avoid harmful material (for instance, avoiding misogynistic commentary that is trending). All these measures are highly relevant and can have immediate positive impact on users' daily life online. They shift some burden off the platform's policing and onto user preference – which is appropriate for legal-but-harmful content. Under the OSA, Category 1 services will be required to offer user empowerment tools for adults to reduce the likelihood of encountering certain types of content (like abuse, hate, or self-harm content). Action 7 is very much in line with that regulatory approach, reinforcing it with specific gender-based use cases.

Human Rights Considerations: Enhancing user control is generally positive for rights. It *advances privacy* (users decide who can interact with them and who sees their content) and *advances freedom of expression* in a nuanced way: it allows women to continue expressing themselves online while minimizing the exposure to hostile audiences. It's crucial, however, that these controls are easy to find and use, otherwise they exist only on paper. The right to freedom of expression includes the right *not* to be harassed into silence; tools like block/mute are the modern means by which users can exercise a degree of *private censorship* of what they receive, without impinging on the harassers' ability to speak in their own corner. There is minimal concern that providing such tools could harm others' rights, one person's decision to block does not censor the blocker's antagonist in general, it only curates her personal space. One risk to flag is that if platforms push user controls as a panacea, they might under-invest in systemic fixes. User empowerment should complement, not replace, provider responsibility. Many vulnerable users may not be aware of controls or may feel overwhelmed managing them (for instance, blocking hundreds of accounts one by one during a coordinated pile-on is burdensome). So, while we champion better controls,

we stress that this action should go hand-in-hand with platform-led enforcement (Action 9) and design changes (Actions 4–6) that reduce the *need* for constant user vigilance. Another consideration is equity: all users should have access to these tools regardless of platform or device. Simpler interfaces (like on some gaming consoles or older mobile devices) should still allow robust blocking/reporting. We see no direct legal compliance issues – indeed empowering users to shape content is explicitly envisioned by UK lawmakers as an alternative to state or platform-imposed content removal, thus supporting free expression within the law. We recommend services also provide **education or prompts** about these features (for example, if a user receives a sudden influx of abuse, the app could proactively suggest: "It looks like you're getting a lot of replies – would you like to restrict who can reply or use our bulk block tool?"). This kind of contextual help would maximise uptake. In conclusion, Action 7 is highly sufficient in concept and, if broadly adopted, will significantly promote women and girls' autonomy online. It embodies user-centric design, which is a positive evolution for online governance.

Action 8: Enable Users Who Experience Online Gender-Based Harm to Make Reports

Assessment: BILETA views this action as a cornerstone of effective redress. Even the best preventative measures cannot stop all harm, so it must be easy and safe for users to report abuse or seek help when they are targeted. The guidance encourages platforms to design reporting and flagging systems with a *trauma-informed* approach. This is particularly relevant for sensitive harms like domestic abuse or intimate image abuse, where victims may fear retribution or feel distress in reporting. Good practice here includes *user-friendly reporting tools* (easy to find, simple to use, with clear instructions) and offering real-time support options. For instance, an app might have a "panic button" or quick report feature within each message or post, and upon reporting certain severe abuse (like a threat or non-consensual image), the user could be provided with on-screen information about specialist helplines or the option to contact law enforcement. The guidance itself notes that Ofcom deems it *"core to women and girls' safety online that they have greater control … and the information about them visible to others,"* and one foundational step is *"user-friendly reporting tools"* that are accessible and intuitive. We agree: a convoluted or hidden reporting system effectively denies users a remedy and can violate their right to seek recourse. Some platforms have already improved in this area (e.g. Facebook and Instagram's interfaces allow reporting specific categories like "hate speech -> targets women" with a few clicks and offer updates on the report status). We also support the idea of in-platform support: perhaps a chat with a support agent or an AI assistant that can guide victims of online violence on what steps to take (while maintaining privacy). The draft guidance references *trusted flagger programmes* as well, which could allow expert organisations (like domestic violence charities) to assist in flagging harmful content or users on behalf of victims. This can be very effective for users who are traumatized or face systemic abuse (e.g. women journalists under coordinated attack) – NGOs with trusted flagger status could help get the content reviewed faster.

Human Rights Considerations: Effective reporting mechanisms could further freedom of expression for victims. When women and girls know they can report abuse and have it dealt with, they are less likely to be bullied into silence. It also ties into the right to an effective remedy (a general principle in human rights law): while private platforms

aren't courts, providing a channel to complain and seek action is part of the wider ecosystem of remedy for online harm. That said, reporting processes must be designed to avoid *bias or abuse*. One risk is false or malicious reports – bad actors might spam the reporting system to take down a woman's account (a known tactic to silence activists, sometimes called "mass reporting"). Platforms should mitigate this by, for example, detecting bulk report patterns and reviewing them carefully, or not penalizing users without human review if the content is borderline. The guidance's mention of *constructive feedback loops* – allowing users to give feedback on the reporting experience – is a good way to refine these systems and ensure they are working as intended. Privacy is another consideration: when a user reports abuse, they might need to share sensitive info (like why content is abusive, which could reveal personal context). Platforms should handle such data securely and only use it for the purpose of addressing the report, in line with data protection laws. Anonymity options can be explored too: some victims might wish to report content without revealing their identity to the perpetrator or public. As for the *applicability* of good practices, we think even smaller services can implement basic accessible reporting (e.g. a simple form or email hotline), though larger services will be expected to provide more sophisticated in-interface tools. The guidance sets a reasonable expectation that this action is universal. We recommend also incorporating *user feedback on outcome*: users should be informed of what was done in response to their report (even if only a generic note). Lack of feedback discourages reporting and leaves victims feeling voiceless. In summary, Action 8 is strongly sufficient and legally compliant (indeed, failing to offer reporting could put a service at risk of not meeting OSA duties to address illegal content). It carries minimal risk beyond operational burden. To ensure practical impact, enforcement of this action should check not just that a "report button" exists, but that it functions well (e.g. high uptake, satisfaction, and resolution rate). Including some case studies of successful reporting systems (perhaps referencing platforms that have hotline partnerships or trauma-informed designs) would be useful; we note the consultation invites more case studies here, which we support.

Action 9: Take Appropriate Action When Online Gender-Based Harms Occur

Assessment: BILETA considers this the follow-through to reporting – it covers enforcement and remedial actions by platforms against offenders and harmful content. The guidance encourages more proactive and consistent enforcement, suggesting measures like "strike-based" policies and escalating sanctions for repeat violators, up to outright bans. This is very much in line with standard content moderation models (e.g. a first offense earns a warning, repeated harassment leads to temporary suspensions, and persistent abuse can result in account termination). We agree that a clear, graduated enforcement scheme can deter would-be abusers and signal that misogynistic harassment has consequences. For example, if users know that sending threats will lead to an immediate suspension and a record on their account, they may think twice. The guidance also implies looking beyond individual accounts, if certain parts of a service (say, an unmoderated forum) breed gender-based hate, "appropriate action" could include restricting or better moderating those sections. Additionally, we appreciate the reference to proactive enforcement: rather than waiting solely for user reports, platforms should utilise automated detection and staff review to identify egregious abuse (such as networks promoting image-based abuse) and act on it. Some platforms have "community operations" teams that hunt out coordinated harassment

campaigns; this practice should be expanded, especially for high-profile incidents (e.g. a woman public figure receiving thousands of abusive comments – the platform could sweep and clean that up en masse). Importantly, cooperation with law enforcement falls under "appropriate action" in severe cases. The guidance is focused on platform actions, but we note that where online abuse constitutes a crime (threats to kill, stalking, etc.), platforms should have channels to refer such cases to police or assist victims in doing so. This ties back to human rights: states have a duty to protect individuals from violence, and that sometimes means ensuring private actors (platforms) cooperate in prosecuting serious abusers.

Human Rights Considerations: When it comes to punitive action like content removal or account bans, freedom of expression is directly engaged. However, not all expression is protected – threats, harassment, and non-consensual intimate images are clearly unlawful or unprotected speech. In those cases, platforms not only may remove content consistent with Article 10(2) ECHR (protecting the rights of others) but arguably have a responsibility to do so as part of the state's positive obligation to safeguard victims' rights (e.g. the right to private life under Article 8). That said, even enforcement against lawful-but-harmful content must be careful and proportionate. We advise that "appropriate action" should always include an element of due process: users who are penalized should be informed of what rule they violated and have an opportunity to appeal if they believe it, was a mistake. The draft guidance's nod to an external oversight mechanism (Action 1) and transparency around enforcement (Action 3) buttress Action 9's compatibility with fairness. For example, a strike system works best when users know the rules (first strike warning, second strike 1-week ban, etc.), and an appeals process exists for wrongful strikes – this ensures that while women are protected, users are not arbitrarily deprived of their accounts. Another consideration is consistency and non-discrimination in enforcement. Platforms should ensure that rules on misogyny/harassment are enforced equally regardless of the perpetrator's profile or the victim's status. There have been problematic instances historically where high-engagement users or advertisers were given leeway despite abusive behaviour, which undermines trust. Conversely, there have been cases of overzealous enforcement – e.g. women of colour being banned for calling out racism using terms that algorithms misinterpret as hate speech. To avoid discriminatory impact, enforcement tools (like automated filters or word lists) must be continuously evaluated for bias (recalling Action 4's role in testing). We are pleased that Ofcom emphasises continuously *improving systems and minimising bias*. This acknowledges that enforcement algorithms should not disproportionately silence the voices of marginalized women or over-censor legitimate content.

The OSA already makes platforms liable for failing to remove illegal content (with hefty fines for non-compliance) and requires terms of service enforcement. So, Action 9's push is largely in service of meeting those legal duties in a gender-aware way. One challenge is enforcement *capacity*, i.e. smaller services might struggle to actively police content at scale. Ofcom could consider encouraging collaboration (for example, sharing hash databases of known abusive images via initiatives like the UK's StopNCII.org for intimate image abuse). That kind of collective effort can lighten individual burdens and has been successful in removing violent or sexual abuse imagery across platforms. Finally, we note a potential gap: persistent abusers often re-offend by creating new accounts. "Appropriate action" might include measures to prevent

easy re-registration by banned harassers (within privacy limits). Some sites use device fingerprinting or phone/SMS verification after bans – though these raise privacy and inclusion concerns, they can be tools to discuss in extreme cases. Any such measure should undergo privacy impact assessment and consider exceptions (e.g. not locking out someone who was falsely flagged). On balance, we believe decisive action against violators is warranted and can be done consistent with human rights. If anything, this is an area where platforms have historically under-delivered (allowing prolific abusers to return repeatedly). Robust enforcement, transparently carried out, will send a message that online violence against women will not be tolerated, aligning with international human rights resolutions.

In conclusion on Q2, BILETA finds the nine proposed actions to be relevant and ambitious, covering critical interventions to make online life safer for women and girls. We consider them generally sufficient, with the caveat that ongoing dialogue and possibly future refinement (or additional specificity via codes of practice) will be needed as we see how these measures work in practice. We emphasise the importance of measuring practical impact. It will be essential to track outcomes (e.g. reduction in prevalence of abuse, user satisfaction with safety tools, etc.) to ensure these actions deliver tangible improvements. If any action area proves ineffective or too difficult to enforce, Ofcom should revisit it in future guidance updates. Similarly, if new risks arise (for instance, harms in emerging platforms like the metaverse or AI chatbots used for harassment), the framework should expand. The commitment to periodic update of the guidance is reassuring in this respect. Finally, we appreciate the human-rights-based approach evident in the draft (explicitly or implicitly) and encourage Ofcom to continue consulting with rights experts, women's groups, and industry to fine-tune these recommendations for maximum positive impact with minimal adverse effects on fundamental rights.

Response to Question 3: Good Practice Steps and Case Studies (Chapters 3–5)

Chapters 3, 4, and 5 of the draft guidance illustrate each action area with good practice steps and case studies. We assess the effectiveness and applicability of these examples, noting any risks, and suggest additional good practices where appropriate.

A. Good Practices in "Taking Responsibility" (Governance, Risk Assessment, Transparency)

Key Good Practices Identified**:** In Chapter 3, Ofcom outlines steps that providers can take to embed women's safety into governance (Action 1), perform gender-informed risk assessments (Action 2), and increase transparency (Action 3). Notable good practices include setting specific policies on gender-based abuse, ensuring decision-making includes intersectional perspectives, consulting with subject-matter experts or survivor groups, conducting user surveys, and exercising caution in data sharing for transparency. A case study hinted for Action 1 is establishing an external oversight board or ombudsman to review moderation decisions related to abuse. For Action 2, a case study example is *Chayn* working with a platform to create a trauma-informed survey of women's experiences, demonstrating how user input can shape risk under-

standing. Although the guidance did not include a concrete case study for transparency (Action 3) due to limited industry examples, it invites input on how transparency can be implemented.

Effectiveness: These practices are grounded in proven strategies. Having clear policies on online misogyny and harassment (e.g. explicitly banning "misogynoir" – hate against Black women – in terms of service) is effective because it gives both users and moderators a clear standard. The intersectional decision-making point is particularly effective: if a platform's trust & safety team includes women, minorities, and individuals attuned to abuse dynamics, they are more likely to spot and prioritise issues that a homogenous team might miss.

For Action 2, the use of surveys and external assessments has high potential. The case study of Chayn's trauma-informed survey indicates that when done carefully, platforms can gather deep insights (e.g. how safe women feel using certain features, or whether victims of abuse find support on the service). A trauma-informed approach (designing questions that are sensitive and not re-traumatizing) is key to effectiveness. The risk assessments under DSA (EU) similarly require consulting civil society; early reports suggest platforms are engaging with women's groups to fulfil those obligations. This cross-pollination of knowledge is leading to risk mitigation steps like better content moderation in languages or contexts where women face distinct harassment (e.g. women politicians in some countries being targeted by coordinated disinformation campaigns – platforms identified this risk after NGOs flagged it). As for transparency good practices, even though an industry gold standard hasn't emerged yet, some effective examples exist: YouTube's quarterly reports now break down harassment and hate content stats. These helped stakeholders gauge progress. We expect the EU DSA transparency regime (applicable from 2024) to greatly enhance data availability on how different user groups (including women) are affected and how platforms respond. That regulatory push will promote the effectiveness of transparency as a tool – companies might as well implement similar reporting in the UK via Ofcom's guidance, to stay ahead of compliance.

Risks and Challenges: One risk is policy over-reach – if a platform sets a very strict anti-harassment policy without nuance, it could lead to over-removal. The guidance's own example lists "terms of service can clearly describe harms such as 'misogyny' (hate directed at women and girls)". Clarity is good, but platforms must avoid vague terms that moderators struggle to interpret. For instance, defining "gender-based harassment" too broadly might result in removing content where women reclaim slurs or engage in critical discussion of gender issues. The good practice of consulting experts can mitigate this risk by fine-tuning definitions. Another challenge is resource allocation – smaller companies might find it hard to convene expert panels or run user surveys. The case studies mostly draw from larger platforms or collaborations. To apply this good practice across the board, some scalability is needed: perhaps industry associations (like techUK) could facilitate collective workshops where multiple small platforms hear from abuse survivors at once.

Transparency also carries a risk: too much detailed transparency might expose platform methodologies to adversaries. However, the guidance's advice to consult ICO and follow anonymisation for data sharing is apt – privacy-preserving transparency

(aggregated stats, anonymous examples) can avoid exposing personal data. A potential risk in transparency is interpretation: raw data could be misinterpreted by the public (e.g. a spike in reports might indicate worse abuse *or* better user trust in reporting). We suggest platforms accompany data with context/explanations to mitigate misunderstanding.

Additional Good Practices: One practice we suggest in the "taking responsibility" domain is gender-responsive training and metrics. The draft mentions training staff on gender-based harms – we fully support that. An addition could be requiring moderators to undergo periodic training on unconscious biases (so they don't dismiss abuse reports due to personal biases) and on cultural contexts of abuse (so global platforms can handle, say, misogynistic slang in different languages). Another emerging good practice is internal audits or assessments of algorithmic bias. For transparency, companies could conduct audits of whether their algorithms (for content recommendation or moderation) treat content by or about women unfairly. If issues are found, companies should be transparent about addressing them.

Case Study Additions: BILETA proposes two additional case studies to illustrate good practice: (1) The Wikimedia Foundation's approach to harassment – Wikipedia, while not a traditional social media, has faced harassment issues, especially targeting women editors. In response, Wikimedia created an *Ombuds commission* and enhanced reporting pathways, and importantly, developed an open-source machine learning tool ("Revision Filter") to flag toxic comments on discussion pages. This multifaceted approach (community governance + tech tool) reduced visible harassment and could inspire other community-based platforms. (2) The Microsoft Xbox gaming safety team – they implemented proactive content moderation with human review for gaming communities, and they publish transparency reports including data on harassment enforcement in Xbox Live. They also partner with organisations on digital civility campaigns targeting youth (many girls experience harassment in gaming). These examples would show that the good practices aren't just limited to social networks, but also apply in knowledge and gaming platforms, respectively – broadening relevance.

B. Good Practices in "Preventing Harm" (Design & Technical Measures: Abusability, Defaults, Content Spread)

Key Good Practices Identified: Chapter 4 offers concrete design interventions: red teaming for abuse scenarios, working with perpetrator-behaviour experts, using "personas" in design, adhering to evaluation principles (for features), strong default privacy/settings, prompts and nudges, filtering and blocklists updates, removing certain content from training data (e.g. nudity). Case studies include a detailed look at how red teaming can expose vulnerabilities (the guidance references Ofcom research on red teaming generative AI features, presumably to prevent deepfake intimate images). Another case example is the use of "personas" testing: creating fictional profiles of different types of women/girls (e.g. a teenage girl, a public woman figure, etc.) to test how a new feature might be experienced by each. For safer defaults, the guidance doesn't list a specific company case, but it draws from known practices (World Wide Web Foundation and others have advocated for strong default privacy in social media). Also mentioned are behavioural insights to empower users (a DSIT/Public report is cited) – likely referring to experiments where user interface changes led to more thoughtful sharing or reporting.

Effectiveness: Many of these practices are drawn from emerging "safety tech" and design research, showing promising results. Red teaming has been effective in cybersecurity for years and applying it to content abuse is novel but logical. A well-known example: a few years ago, a "Stalker's Paradise" study (cited in the draft) uncovered how features like location sharing or calendar syncing could be misused by abusers, prompting companies to patch those holes. Working with perpetrator experts might seem unconventional, but understanding how trolls or abusers think can greatly improve defences (similar to how banks work with former fraudsters to test their systems). Personas and user journey mapping are standard UX practices; integrating safety into that (e.g. considering "How might an abusive ex-partner exploit this feature?" for each persona) is highly effective in catching issues early.

Safer defaults have proven effectiveness as well. For instance, when WhatsApp implemented default end-to-end encryption and set new groups to "invite-only" by default for non-contacts, it reduced unwanted contact significantly on the platform. Instagram's move to default teen accounts to private reduced random adult follow requests. These real-world changes show that defaults can indeed deter or eliminate certain harm avenues. Nudges and prompts have a growing evidence base. Filter and blocklist improvements – one example: Google updated its autocomplete and search recommendation algorithms after learning sexist or violent suggestions were showing up. This reduced the autosuggestion of harmful queries. Removing "nudity content from training datasets" as mentioned likely refers to ensuring AI vision algorithms don't sexualize or erroneously censor women's content (which can happen if the AI is trained on biased data). By removing irrelevant nudity data, an algorithm might better distinguish, say, breastfeeding images (allowed) from pornography. That fine-tuning is effective to reduce wrongful takedowns of women's content while still catching true sexual abuse imagery.

Reducing circulation tactics like throttling shares of flagged posts or limiting pile-on visibility have also been field-tested. Reddit, for example, found that by delisting hate communities from search and recommendations (without outright deleting them at first), the growth of those communities stalled, and some died out – a harm reduction short of full removal. This indicates that visibility control works. Facebook's "Gender-Based Violence Hub" internally rolled out some years ago gave moderators a dashboard to see when an individual (often a woman) was being mass-tagged or attacked across the platform, allowing rapid response – effectively reducing the spread and piling on.

Risks and Challenges: For these technical measures, a key risk is over-reliance on automation and false positives. Red teaming might identify patterns that lead to automated rules (like "auto-delete any image that our detector flags as nudity"). If done naively, this could yield many false positives (e.g. flagging art or health content) and disproportionately affect certain users. However, the guidance's approach of combining human insight and iterative testing can mitigate that – the goal is not to let AI run rampant, but to continuously refine it. Another risk is smaller services' capacity, as noted, not all have in-house teams for AI or red teaming. They might interpret "use red teaming" as too advanced. To address this, perhaps industry collaboration or third-party safety audits can help smaller companies implement these practices.

Default settings changes might face user resistance in some cases (if users are used to open settings). But since this is about new improvements, likely users will appreciate more privacy by default. There is a subtle risk that stricter defaults could reduce user engagement (some companies worry if things are too locked down, users might share less or connect less, affecting growth metrics). This is more a business concern than a safety risk, but it can be a barrier to adoption. Ofcom's guidance, by framing it as good practice, provides air cover for safety teams inside companies to argue for these changes despite any engagement hit – safety should trump shallow engagement metrics.

For nudges, a risk is users finding them annoying or paternalistic. The key is to calibrate frequency and allow some bypass ("don't show this again" option if needed). Additionally, blocklists (e.g. lists of slurs or banned terms) need constant updating and cultural context – a risk is relying on them without input; involving community feedback (some platforms invite users to suggest new slurs that abusers invent) can keep them effective.

Additional Good Practices: In this category, one can look to innovative safety-by-design features emerging in the industry. For example, temporary user freezes: some platforms are experimenting with automatically limiting a user's ability to post further if their recent posts are getting lots of reports (a cooling-off period). This is a step between no action and banning that might prevent a spiral of abuse.

Also, user education pop-ups can be seen as a design feature: e.g. if someone uses a term that is often considered derogatory toward women, a gentle pop-up could explain why that term is harmful (not just prompt to reconsider but actually educate). This turns a nudge into an educational moment – aligning with broader goals of cultural change.

Case Study Additions: We suggest adding a case study on Discord's approach to harassment – Discord (a chat platform) introduced a combination of user controls (like who can DM you) and auto-mod tools that server admins can use to filter out content (with pre-built filters for slurs including misogynistic terms). Discord also has an interesting bot that uses AI to flag possible harassment in chats. This multi-layered approach in a private messaging environment could illustrate how to prevent harm in more closed networks.

C. Good Practices in "Supporting Users" (User Control, Reporting, Response)

Key Good Practices Identified: Chapter 5 focuses on empowering and helping users when harm happens. Good practices highlighted include: bundled safety modes (for account control), easy mechanisms to block/mute multiple abusers, visibility settings to retroactively protect content, strengthening account security (e.g. 2FA to prevent stalking via account compromise), accessible reporting tools (with categories matching gender-based harms), trauma-informed design of report flows (not forcing the victim to repeatedly see the abusive content while reporting), offering feedback on reports, and escalating enforcement like strike systems or device bans for ban evaders. A specific bullet in the guidance references feedback loops on reporting and trusted flagger partnerships (from eSafety). The case studies likely include the "user survey" by Chayn already noted, and possibly one on a platform providing real-time sup-

port (the draft mentions real-time support in context of reporting tools – perhaps a reference to something like Tinder's Noonlight panic button integration). Also, the case study for Action 9 might include an example of a strike policy or innovative enforcement: e.g. Twitch (the streaming service) has a practice of banning users not just by account but by device or IP for severe offenses and even acknowledges user behaviour off-platform (if a streamer engages in egregious harassment elsewhere, Twitch might act). These could be implicit in Ofcom's thinking on "persistent abusers".

Effectiveness: Many of these supportive measures directly address known pain points for users. The ability to block or mute multiple accounts is extremely effective against brigading. Third-party apps like Block Party (which let users share block lists) have shown that community-driven bulk blocking significantly reduces harassment visibility for targets. Incorporating that into platforms (as a standard feature) is indeed best practice. The concept of "safety mode" or one-click lockdown is effective for those in crisis: Facebook has something akin to this with its "lock profile" feature (mostly used in regions to prevent non-friends from seeing anything) – women journalists under threat have used it to halt abuse. If mainstream platforms had a button "I'm experiencing a pile-on" that instantly turns a suite of settings to private/high security, that could prevent escalation, we consider that an effective idea that should be adopted widely.

Strengthening account security (like requiring two-factor authentication or alerting users of login attempts) can prevent a scenario where an abuser hacks or takes over a woman's account, which can be devastating (there have been cases of ex-partners hacking social media to impersonate or humiliate victims). Good security defaults and features (e.g. backup codes kept safe, login alerts) are effective at stopping this kind of harm.

Reporting tools and response: A streamlined, empathetic reporting process greatly increases reporting rates and therefore removal of harmful content. The trauma-informed approach – for example, giving the option "do you want to attach a personal note to explain context?" or "we will not ask you to view the content again, it's been logged" – can reduce the burden on victims. Some effectiveness evidence: Instagram introduced an option to report for "nudity > someone is in this photo without their consent" specifically for image-based abuse; reports under that category go to a specialised team. This specialisation makes handling more effective (staff trained in that harm type). Also, automation can help here: using AI to detect when a reported image might be an intimate image and fast-tracking it for removal (some companies do this, recognising time is of the essence).

Enforcement actions like strikes and bans absolutely help by removing perpetrators from circulation. The trick is consistent enforcement. If a platform actually bans the worst offenders and keeps them out, the overall level of abuse drops – for instance, Reddit's ban of certain hate subreddits and users in 2020 led to a documented decrease in hate speech site-wide. The effectiveness of enforcement can also be measured by recidivism: if someone is permabanned and doesn't return, that's effective; if they do, then maybe stronger identity verification is needed.

Risks and Challenges: A challenge in user empowerment tools is discoverability and usability. A sophisticated control is only good if users know how to use it. Some women (especially younger girls or less tech-literate individuals) may not navigate complex

settings. The draft's focus on user-friendly design is important. Services should user-test these features with target demographics to ensure they are indeed helpful.

For reporting, the biggest risk is that it may not lead to meaningful action, causing frustration. If users frequently report abuse and get the response "we found it doesn't violate our standards" when it clearly was abusive, they lose trust. This can happen if policies are too narrow, or moderators are overworked. So, a good practice beyond just the tool is adequate resourcing of moderation teams, and possibly moderator training specifically for gender-based issues. This ensures the reports are correctly adjudicated.

Another risk: retaliation. If a perpetrator is banned, they might seek revenge on the victim elsewhere. Platforms might consider notifying users of enforcement in a careful way; maybe not telling the reporter "We banned X user because of your report" if there's a risk X could guess who reported them. This is tricky – anonymity of reporting can protect victims. Thankfully, most platforms do keep reports anonymous, but smaller tight-knit communities might still deduce reporters. So, trauma-informed practice might mean offering tips to reporters like "Consider not engaging further with this user; here are resources if you feel unsafe offline." This merges online enforcement with offline safety advice – a holistic approach that could be considered good practice.

Additional Good Practices: We suggest a few additions: Integration with law enforcement or legal support – e.g. for severe cases, platforms can have a process to help victims preserve evidence or file police reports. Some platforms already do: Facebook has a "law enforcement portal" and will provide data to police with proper requests; a good practice is informing users of how they can get help from police if threatened. Even providing *general guidance* like "if you are in immediate danger, contact 999" or "here's how to report to police and what information to include" in the safety centre can be useful. Another addition is mental health support: experiencing online abuse can be traumatic; platforms might consider offering links to counselling or at least moderation teams trained in mental health first aid who can handle escalated cases sensitively.

Another supportive practice: Community moderation and support networks – for example, some online forums have volunteer "aunties" or peer mentors who new users can talk to if they face harassment. While not directly in Ofcom's purview, encouraging a culture of peer support can complement official reporting. For instance, the game *League of Legends* instituted a tribunal system (players reviewing reports) and an honour system to encourage positive behaviour, which somewhat improved their notoriously toxic environment. Empowering users to help each other (with proper safeguards) could be mentioned as a soft good practice.

Additional Observations: "Taking Responsibility"

The emphasis on senior leadership and training is well-founded. Active, educated leadership ("safety champions" at the board or VP level) must ensure that these issues are taken seriously. We note that a coalition of NGOs and experts (has already developed a VAWG Code of Practice (2022) recommending many similar governance steps, indicating broad consensus that such measures are effective.

Ensuring transparency (Action 3) is especially critical: publishing data on the prevalence of abuse, content removals, and user reports – broken down by gender and other characteristics – will help identify systemic problems and biases. For instance, transparency could show if certain types of abusive content are under-moderated or if reports by women are less likely to result in action, enabling necessary course corrections. Ofcom's planned use of annual transparency reports and possible issue-specific requests (e.g., data following a major "pile-on" incident) is an innovative tool to keep companies accountable.

To mitigate risks associated with transparency, we agree with the guidance that services should not publish granular details that could identify victims or aid offenders in evading detection. Instead, transparency should operate at a statistical and policy level. If executed as intended, these governance and transparency measures will improve regulatory oversight and drive a race-to-the-top on safety practices without unduly infringing rights.

Summary of Good Practices

We appreciate that Ofcom's guidance draws on a wide range of sources: academic research, civil society toolkits, industry pilots, and regulatory innovations.

Ensuring Feasibility and Avoiding Unintended Consequences: Most good practices listed are technically and operationally feasible, especially for larger platforms. Smaller platforms may need more support – perhaps through industry forums or an Ofcom "toolkit" summarising best practices and vendors/tools that can help implement them. The risk of unintended burden on free expression exists mainly if measures are too blunt (like poorly tuned filters or lack of appeal in enforcement). To counter that, we reiterate that many good practices (like transparency, oversight, user control) are themselves safeguards against such burdens. Where we identified potential issues (algorithmic bias, vague policies, etc.), the remedy lies in inclusive design and continual review, which the guidance encourages by calling these "iterative" and seeking stakeholder feedback.

There's also a risk of **discriminatory impact** if, for instance, enforcement focuses on certain languages or communities more than others. Companies should monitor whether their actions disproportionately affect women of certain backgrounds. For example, a platform might notice most bans for "harassment" are against users in one region – is that because that region has a harassment problem, or because the algorithm misinterprets their slang as harassment? These are deeper questions that good data (transparency) can help answer. We encourage Ofcom and researchers to analyse the data that comes out of these implementations for any such patterns.

Enforcement Viability: The guidance itself is not enforceable law, but many foundational steps tie to legal duties. Where good practices go beyond, Ofcom's planned approach (publishing a progress report 18 months after final guidance) can create a pseudo-enforcement via public accountability. If uptake is poor, Ofcom might later consider integrating some of these into Codes of Practice, which could then be enforced. One challenging area is how to verify that a platform is *truly* conducting, say, abuse red teaming or surveys. Ofcom may need to rely on audit rights under the OSA – possibly reviewing internal documents or interviewing staff during assessments. That's doable for big firms but hard for numerous smaller ones.

Response to Question 4: Comments on Ofcom's Approach to Encouraging Providers to Publish Guidance Addressing Women and Girls' Safety and 'Good Practice' Recommendations

BILETA welcomes Ofcom's proposal to publish assessments of how providers are addressing the safety of women and girls online. This mechanism can foster accountability and incentivize the adoption of recommended practices. Furthermore, the non-statutory nature of the guidance provides necessary future-proofing, given the evolving nature of online harms targeting women and girls, including increasingly sophisticated AI-generated content like deepfakes. For example, research indicates that deepfake pornography overwhelmingly affects women, who constitute 99% of victims. The ease and low cost of generating these 60-second videos mean that celebrities, politicians, and everyday individuals are all susceptible to this form of abuse security hero; see also Romero-Moreno.

However, we emphasize the critical need to consider the challenges faced by smaller or niche-topic platforms in implementing this guidance due to potential limitations in resources, knowledge, and capacity. While the consultation's acknowledgement of a proportionate approach is appreciated, recognizing that a 'one size fits all' model is inappropriate for platforms of varying size, risk levels, and functions, the pragmatic risk of non-compliance for smaller entities remains significant without adequate support. Recent reports of smaller discussion forum platforms closing due to the Online Safety Act's regulations, risk assessment requirements, and potential financial burdens underscore this concern. This is particularly pertinent for online 'spaces' intentionally created to provide safe discussion environments for women, such as the FLOW (Feminist Leadership, Organising and Witnessing) forum and the Women's Migration and Asylum Network (WoMAN), which serve women and girls facing risks of discrimination, exclusion, and abuse online.

We strongly advocate for proportionality particularly considering the ECHR and relevant ECtHR case law not only in the guidance and measures themselves but also in compliance expectations and encouragement mechanisms. Publicly casting lower-resource platforms negatively for lacking the capacity of global tech companies would be counterproductive. Instead, tailored guidance, partnerships to support implementation, and incentivized complementary measures are needed to boost the uptake and sustainability of 'good practice' recommendations. The consultation's acknowledgement of potential compliance costs (Table 3) suggests that these pragmatic financial realities should be considered when assessing a platform's 'achievement' of good practice standards.

Beyond the publication of assessments, which according to CJEU case law (such as C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117[49], [50], [58], [65], [66], [77]) should be accessible, intelligible, concise, and transparent, alternative and additional encouragement mechanisms include public recognition and safety or trust schemes that have proven successful in the age assurance space. Ofcom should also avoid punitive framing of assessments, incorporating supportive and developmental approaches alongside compliance determinations. Public acknowledgement of platform efforts for women and girls' safety can extend beyond assessment publications. Certification via safety and trust schemes, akin to the Age Check Certification Scheme, could provide a visible marker of reliability

for users. A ‘safety badge’ for women and girls would demonstrate compliance, enabling informed choices about service engagement, a primary goal in Paragraph 2.88. Ofcom could review and award these certifications, maintaining a public register of compliant platforms, similar to the illegal harms register. This offers a tangible reputational and business incentive for platforms, akin to BSI Kitemarks or compliance badges, supported by the UK ICO while aligning with regulatory requirements.

Response to Question 5: Comments on Impact Assessment, Rights Assessment, or Equality Impact Assessments?

The complexities inherent in balancing freedom of expression with online content moderation are longstanding and largely unresolved. BILETA welcomes the explicit consideration of these rights and their assessment in this guidance, specifically noting the importance of ensuring women and girls can exercise these rights safely. The acknowledgement that Ofcom has strived for a fair balance (A2.10), rather than a perfect one, reflects the ongoing debate between moderation and free speech principles. Paragraph A2.11 rightly notes the reluctance of women and girls to actively participate online and exercise their free speech due to concerns about potential abuse, harassment, and harm, including the increasing threat of malicious deepfakes like non-consensual intimate imagery and targeted online harassment. BILETA encourages Ofcom to provide further clarification and detail on how the guidance intends to overcome these concerns, especially when women and girls are avoiding or withdrawing from platforms entirely due to these specific threats. Strengthening this aspect of the guidance to empower women and girls to exercise their free speech online and contribute to their chosen platforms would significantly benefit the overall guidance and support the mission of ensuring online safety for women and girls as active digital citizens. Furthermore, we recommend further guidance and expansion on how moderation practices and free speech considerations should interact with automated content moderation systems where women and girls require protection, given the potential for bias in these systems to either fail to detect harmful content targeting this demographic or unfairly censor their speech.

BILETA expects Ofcom to carefully reflect on the recent Supreme Court ruling in *For Women Scotland Ltd v Lord Advocate* [2025] UKSC 16 and incorporate its implications, alongside subsequent Equality Commission comments and guidance, into any final guidance for technology companies on the protection of Women and Girls. This ruling, which determined that the definition of “woman” under the Equality Act 2010 is based on biological sex, excluding individuals who identify as transgender women with or without Gender Recognition Certificates, is likely to have significant ramifications for the assessment previously made and presented in the consultation, particularly regarding the scope and application of safety measures for all individuals identifying as women and girls. These implications for the scope and pragmatic application of measures by platforms should be duly considered in the formulation of final codes. Anecdotal accounts suggest that the aforementioned ruling has sparked increased online discourse regarding the judgment, potentially impacting both trans and cis women depending on their perspectives on the ruling and trans rights Centre for Mental Health . We note that this is a highly contentious issue and exemplifies the delicate balance that must be struck between competing rights and free speech considerations, both offline and online, a balance that remains under debate. This reality appears to be significantly acknowledged in the consultation document, points A2.9 to A2.12. The

absence of definitive conclusions regarding where this balance will ultimately lie is consistent with ongoing discussions in both academia and practice as content moderation initiatives evolve. We strongly advocate for a thorough consideration of how this ruling will impact the scope and direction of guidance in this area. BILETA would welcome and encourage engagement with relevant NGOs and charities to gain a deeper understanding of how these simultaneous events will affect each other and the online experiences of users, particularly women and girls.

BILETA's strong advocacy for thorough impact assessments underscores the critical need to proactively evaluate the potential consequences of the proposed guidance and AI-driven safety tools on the fundamental rights, data privacy, and equality of women and girls. This stance aligns with UN Resolution A/78/L.49 on Safe, Secure, and Trustworthy AI, which globally emphasizes the necessity of safeguards and impact assessments for all AI systems. For instance, in the context of deepfake detection—a significant threat involving non-consensual intimate imagery and online harassment predominantly targeting women and girls—a Fundamental Rights Impact Assessment like the one required under Article 27 of the EU AI Act would scrutinize whether the deployment of detection technologies could infringe upon rights such as freedom of expression, non-discrimination, data protection, or privacy.

The increasing reliance on extensive datasets for content moderation and deepfake detection introduces substantial privacy and bias concerns. Data Protection Impact Assessments become crucial under legal frameworks like GDPR (Article 35) to analyse the risks to the personal data of women and girls used in training these AI models. For example, if datasets disproportionately feature certain demographics, AI models trained on this data may exhibit bias, leading to failures in accurately identifying harmful content targeting specific groups of women and girls or, conversely, unfairly flagging their content. Equality Impact Assessments are vital to identify and mitigate such discriminatory outcomes. Research has indicated that facial recognition technology, often a component in deepfake detection, can exhibit racial bias, with higher error rates for individuals with darker skin tones Ju et al. This illustrates the potential for AI-driven tools to perpetuate existing societal inequalities if bias is not proactively addressed.

BILETA's call for Ofcom to encourage proactive scenario planning and stakeholder engagement, as advocated by the World Economic Forum, emphasizes the importance of anticipating potential negative consequences and involving diverse perspectives in the development process WEF. This could involve workshops with women's rights organizations, technology developers, and legal experts to identify and address potential pitfalls early on.

Existing legal frameworks, including UN Report A/73/348 which addresses AI systems, freedom of expression and the rule of law and technology, the EU AI Act (Article 27) requiring Fundamental Rights Impact Assessments for high-risk AI systems, and GDPR (Article 35) mandating Data Protection Impact Assessments, provide a robust foundation for responsible AI development. Furthermore, the EU Digital Services Act (Articles 34, 35) obligates very large online platforms and search engines (VLOPs and VLOSEs) to assess and mitigate systemic risks associated with deepfakes, implicitly promoting the use of Explainable AI (XAI).

BILETA stresses that Algorithmic Impact Assessments (AIAs) for deepfake detection must prioritize explainability alongside accuracy to ensure compliance with GDPR, (UK DPA), which grants individuals the right to meaningful information about the logic involved in automated decision-making, as seen in CJEU case law such as Case

C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 and Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957. Understanding how AI systems make decisions is crucial for building trust among women and girls and ensuring that harmful content is effectively addressed without unfairly censoring their speech. The UK case of *R (on the application of Edward Bridges) v the Chief Constable of South Wales Police* [2020] EWCA Civ highlighted the importance of transparency and justification in the use of automated facial recognition by law enforcement.

Similarly, in the context of deepfake detection, understanding why certain content is flagged as harmful is essential for accountability and fairness. AIAs must incorporate rigorous bias assessments, evaluating demographic disparities in deepfake detection. Research has shown statistical breakdowns in the accuracy of deepfake detection models across different racial groups. AI systems exhibit societal biases disproportionately affecting women and girls, including:

- **Higher facial recognition errors for darker-skinned women** [1], [1].
- **Gender stereotypes in NLP models** [2].
- **Age-related differences in deepfake detection accuracy** [3].
- **Underrepresentation in deepfake avatars, raising risks of targeted campaigns** [4].

Addressing these disparities is essential for identifying, mitigating, and rectifying bias to ensure fairness and prevent harmful outcomes for all women and girls, including those from marginalized groups who are often disproportionately affected by online abuse and the misuse of their likenesses. For example, if a deepfake detection model is less accurate in identifying manipulated images of women from certain ethnic minority backgrounds, they could be left more vulnerable to this form of abuse.

Therefore, BILETA's recommendation for Ofcom's strategy to explicitly include guidance on conducting thorough impact assessments, prioritizing transparency and explainability in AI-driven safety tools, and proactively addressing bias is crucial for fostering a safer and more equitable online environment where all women and girls can participate without fear of disproportionate harm or censorship.

Response to Question 6: Comments on Guidance Regarding Welsh Language

Note: The BILETA contributors to this consultation do not speak Welsh and therefore cannot provide informed comments on the guidance regarding the Welsh language.