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

Consultation response form

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? – Y / N Ofcom's guidance must work *with* the Illegal Content Codes of Practice and Children's Content Codes ("the Codes"), not *against* them. **We are concerned that by including some VAWG measures in the guidance that we argue should be subset VAWG part of the Illegal Codes, Ofcom weakens the regulatory approach in full.** For instance, examples of measures that could be part of the Codes include online misogyny where it is hate speech and/or inciting violence; measures for preventing intimate image-based abuse (now that it is a priority offence), and risk assessments which account for gender-based risks. The latter example is based on women being defined as a demographic with "certain characteristics" as referenced in the Part 1, 2 (a) of the Online Safety Act, as well as a protected characteristic, alongside others such as race, sexuality, religion, disability and gender reassignment, under the 2010 Equality Act.
Question 2: Do you have any comments on the nine proposed actions? Please provide evidence to support your answer.	Confidential? – N Though we are extremely glad to see Ofcom engaging with the "intersectionality" framework, the guidance should more clearly articulate how it should be integrated into practice. As part of this, **Ofcom should put forward recommendations in the guidance that push providers to take the *context* into account during content moderation (reducing circulation), risk assessments, abusability testing, governance and decision making approaches, alongside applying the *intersectionality framework*.** Detecting misogynoir and misogyny

Question	Your response
	So far, social media companies and search service providers' approaches to racialised gender based harms such as misogynoir has been lacking at best - most companies regulated by the Online Safety Act do not have a content policy that mentions misogynoir. **Ofcom's guidance should recommend companies develop policies on misogyny and misogynoir to support the reduction of harmful content towards women and girls.** Also, **Ofcom must call on companies to change policies that allow misogynistic and misogynoiristic content** - most recently, for example, Meta's changes to its content policies to allow that users refer to Black women as property or to Black lesbian women as mentally ill because of their sexuality. These business decisions by companies actively harm Black women and girls and demonstrate a complete disregard for racialised gender based violence. Efforts to detect, stop and/or prevent misogyny and misogynoir online that do exist, tend to focus on the most stereotypical slurs and explicit hate speech, given they are the easiest to detect (and even then, current approaches are often lacking). As found by academics Joseph Kwarteng et al. (2021), misogynoir goes far beyond this, consisting of various forms of nuanced and context-specific tactics including Tone Policing, White Centring, Racial Gaslighting and Defensiveness[1]. Even when companies claim to moderate content that is harmful to Black women and girls, they often fail; our research in 2023, found misogynoir was prevalent across all five major social media platforms in the study. This is partly due to Large Language Models (LLMs) falling short on understanding misogynistic and misogynoiristic comments as "they mostly rely on [...] implicit knowledge derived from internalised common stereotypes about women to generate implied assumptions, rather than on inductive reasoning"[2]. This leads to misogynistic and misogynoiristic content being misclassified as opinions or statements rather than as hateful content, failing to identify a significant amount of misogynistic and misogynoiristic content.

[1] Kwarteng, Joseph; Coppolino Perfumi, Serena; Farrell, Tracie and Fernandez, Miriam (2021). Misogynoir: Public Online Response Towards Self-Reported Misogynoir. In: ASONAM '21: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Association for Computing Machinery, New York, NY, United States pp. 228–235. DOI: https://doi.org/10.1145/3487351.3488342

[2] Muti, A., Ruggeri, F., Khatib, K. A., Barrón-Cedeño, A., & Caselli, T. (2024). Language is Scary when Over-Analyzed: Unpacking Implied Misogynistic Reasoning with Argumentation Theory-Driven Prompts. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 21091-21107). Association for Computational Linguistics, ACL Anthology. https://doi.org/10.18653/v1/2024.emnlp-main.1174

Question	Your response
	As argued by Chelsea Peterson-Salahuddin (2024) there are several complementary mechanisms through which a contextual approach to algorithmic content moderation can be implemented to better identify misogynoir content[3]. A 2023 study by Mullick et al. demonstrates how instead of commonly used single binary classifiers to label content as offensive or not offensive, content moderation can be designed as a cascade of binary multi-layer questions about the content, and it's context, to determine whether it violates platform policies[4]. **This cascading approach should be a recommendation made by Ofcom's to better identify online misogyny and misogynoir and prevent circulation of harmful content.**
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? – N Keeping Black women and girls safe, necessarily demands an engagement with, and response to, the nature of racialised gender-based violence. This is in the context that online "misogynoir is deeply connected to violent extremism and the virulent development of violence towards Black women and girls, and the consumption of misogynoir has deleterious effects on the well-being of Black women and girls"[5]. **For this reason we were glad to see Ofcom's inclusion of a case study on detecting misogynoir in the new guidance for protecting women and girls, though the recommended approach should be more nuanced and practical (see below).** Online abuse is nuanced, multifaceted and contextual, so categorising abuse is always a challenge. *Good* practice means accurately identifying misogynistic and misogynoiristic content *using nuanced, multifaceted and contextual approaches*. As we know, "digital media has served as a home for the public's consumption of anti-Black misogyny as seen in memes, [...] harassment, and [...] videos from the Black

[3] Peterson-Salahuddin, C. (2024). Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation. Big Data & Society. 11. 10.1177/20539517241245333.

[4] Mullick SS, Bhambhani M, Sinha S, et al. (2023) Content moderation for evolving policies using binary question answering. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), Toronto, Canada, July 2023, pp.561–573. Association for Computational Linguistics.

[5] Onuoha, A. 2021. Digital Misogynoir and White Supremacy: What Black Feminist Theory Can Teach Us About Far Right Extremism. https://gnet-research.org/2021/08/09/digital-misogynoir-and-white-supremacy-what-black-feminist-theory-can-teach-us-about-far-right-extremism/

Question	Your response
	manosphere [...] an online community that seeks to encourage Black men, but does so by disparaging Black women"[6]. This includes "Black women [being] discredited, mocked, and disrespected on social media platforms" consistently, whether they are women in the public eye or not[7]. **To effectively promote better practice, it is fundamentally important that Ofcom, and providers, understand the nuances between explicit *and implicit* violence against Black women and girls, as well as other groups impacted by these forms of content, design and decision making.** Equally as important, Black women and girls must have their speech and online experiences protected from inappropriate content moderation, particularly when engaging in counter-speech, intra-community dialogue or when sharing their own experiences in relation to discrimination, violence and harm. **This is currently a gap in Ofcom's guidance.** As demonstrated by the academic studies of Moya Bailey and Brandeis Marshall, Black women and girls on social media have had their content removed and been banned for responding to and defending themselves against racist and sexist posts and incidences, highlighting "how content moderation algorithms reinforce misogynoir or anti-Black sexist logic".[8] Also, hate speech detection systems are more likely to label posts containing dialects of English often spoken in Black communities as hateful, "putting those Black users who engage in this cultural dialect at higher risk of having their content removed"[9]. As Gorwa et al. (2020) suggest, "Even a perfectly 'accurate' toxic speech classifier will have unequal impacts on different populations because it will inevitably have to privilege certain formalisations of offense above others, disproportionately blocking (or allowing) content produced by (or targeted at) certain groups"[10]. This often results in what Chelsea Peterson-

[6] Onuoha, A. 2021. Digital Misogynoir and White Supremacy: What Black Feminist Theory Can Teach Us About Far Right Extremism. https://gnet-research.org/2021/08/09/digital-misogynoir-and-white-supremacy-what-black-feminist-theory-can-teach-us-about-far-right-extremism/

[7] Onuoha, A. 2021. Digital Misogynoir and White Supremacy: What Black Feminist Theory Can Teach Us About Far Right Extremism. https://gnet-research.org/2021/08/09/digital-misogynoir-and-white-supremacy-what-black-feminist-theory-can-teach-us-about-far-right-extremism/

[8] Bailey M (2021) Misogynoir Transformed: Black Women's Digital Resistance. New York: NYU Press. Marshall B (2021) Algorithmic misogynoir in content moderation practice. Heinrich-Böll-Stiftung European Union and Heinrich- Böll-Stiftung.

[9] Davidson T, Bhattacharya D and Weber I (2019) Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 29 May 2019, pp.25–35. Association for Computational Linguistics

[10] Gorwa R, Binns R and Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7(1).

Question	Your response
	Salahuddin refers to as "over-or-underblocking of hate speech", resulting in what Safiya Noble (2018) terms "algorithmic oppression" i.e. algorithms that reinforce oppressive social structures[11]. To effectively keep Black women and girls safer online and create more equity for groups of users who are historically and systemically marginalised, tech companies must engage with design justice principles that acknowledge and account for systemic power differentials that exist between users along multiple axes of oppression, such as race, gender, age and sexuality[12]. <u>Coding misogynoir taxonomy</u> In order to help improve identification, mitigation and prevention of violent content towards Black women and girls, we've built upon the work of scholars to develop a taxonomy for coding (i.e. classifying) misogynoir. This framework can be adopted and embedded into content moderation policies and practices, risk assessments and adaptations to recommender systems. **This coding misogynoir taxonomy is a bid to support Ofcom, and companies, to better understand the online manifestations of this type of hate, and to propose more specific methods that can automatically identify it within the case study in the guidance[13]**. We've specifically built upon the work of Kwarteng (et al.), who define Tone Policing, White Centring, Racial Gaslighting and Defensiveness as core to digital misogynoir. Here are five categories to help identify, mitigate and prevent misogynoir (and misogyny) that should be *understood within context*: **Subjugation (which includes:)** ❖ Gender trolling: Derogatory or discriminatory gender-based insults and stereotypes ❖ Nationalist othering: Dismissing someone's right to live, work or represent the UK because their race and/or gender is not part of the Eurocentric, patriarchal standard

[11] Peterson-Salahuddin, C. (2024). Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation. Big Data & Society. 11. 10.1177/20539517241245333. and Noble S (2018) Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press

[12] Costanza-Chock, S. (2018). Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice . Proceedings of the Design Research Society. https://ssrn.com/abstract=3189696

[13] Kwarteng, Joseph; Coppolino Perfumi, Serena; Farrell, Tracie and Fernandez, Miriam (2021). Misogynoir: Public Online Response Towards Self-Reported Misogynoir. In: ASONAM '21: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Association for Computing Machinery, New York, NY, United States pp. 228–235. DOI: https://doi.org/10.1145/3487351.3488342

Question	Your response
	Shaming ❖ Tone policing: Dismissing someone's idea because of how they expressed it, instead of what they said ❖ Body policing: Criticising someone's appearance for not conforming to Eurocentric beauty standards, or Criticising someone's appearance instead of their views and actions **Marginalisation** ❖ White centring: Prioritising white culture and white people's feelings or perspectives over the needs of people of colour to maintain the status quo or evade criticism ❖ Sexual misogyny and objectification: Language or content that includes non-consensual sexualisation of the person, or doubts someone's professional merit by suggesting they used sex to earn their position, or Emphasising someone's sexual behaviour or treating them as a sexual object, especially to suggest as the reason for their professional success ❖ Dogwhistle racism: Subtle or coded language that suggests diversity and inclusion practices is why someone achieves professional success **Reverse victimisation** ❖ Defensive: Treating criticism as a personal attack rather than taking accountability ❖ Great replacement theory: White nationalist far-right conspiracy theory that white Europeans are being culturally and demographically replaced by people of colour, especially those from Muslim-majority countries, because of mass migration and declining white European birth rates ❖ Racial gaslighting: Dismissing or downplaying someone's experience of racism by for example, accusing them of exaggerating or manipulation **Illegal violations** ❖ Hate speech: Publicly expressing or encouraging hate towards someone or a group based on their race, gender, religion or sexual orientation ❖ Harassment: Unwanted actions or words that create an intimidating, hostile, degrading, humiliating or offensive environment for someone else

Question	Your response
	❖ Stalking: Repeated and targeted unwanted behaviour towards someone, which can have sexual or racial emphasis ❖ Threats: Stating an intention to harm or inflict damage on someone, their family or their property ❖ Encouraging suicide of self harm: Telling someone to inflict harm on themselves or kill themselves. As you will note categories 1-4 are misogynoiristic content which do not fall within the Illegal Content Codes of Conduct, whereas those under category 5 do. We advocate for companies to embed a contextual and nuanced understanding of misogynoir, and how it interacts with their design choices and profit making business models.
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 We are concerned that the VAWG guidance appears to be treated as *separate* to Ofcom's Illegal Content Codes of Practice, rather than complimentary. For example, Ofcom commits to publishing "an assessment of what tech companies are doing – or not doing" around 18 months after the Guidance is finalised, rather than integrating this assessment into Ofcom's existing approach to oversight under the Codes. Professor Lorna Woods of Essex University rightly points out that companies *should* engage with guidance and must "have cogent reasons" beyond a matter of principle, for not complying, in reality many companies perceive guidance as "ignorable"[14]. **It would be far more effective therefore if Ofcom aligns this guidance with other its mandatory assessments, particularly given the crossover of illegal harms mentioned in both the guidance and the Codes. This could also forge positive ways of working internally, preventing the VAWG team from becoming siloed from other relevant teams.** Our coalition campaign to get women and girls included in the OSA also included a policy ask of mandatory codes of practice for violence against women and girls, *not* guidance. Our position on this has not changed; the VAWG aspect of the OSA needs mandatory enforcement mechanics for it to have a chance at ensuring categorised services, and platforms are held accountable for their involvement in facilitating violence

[14] Woods, L. (2025) Ofcom's draft guidance on protecting women and girls https://www.onlinesafetyact.net/analysis/ofcom-s-draft-guidance-on-protecting-women-and-girls/

Question	Your response
	against women and girls. At this stage, this is only possible if the Government amends the Online Safety Act, i.e., re-categorising the VAWG guidance to be part of the Codes, thus handing Ofcom stronger enforcement capabilities. We consider this to be particularly important in light of, at the time of writing, pressure from the US Government to minimise digital regulations on US-based tech companies with UK users, and the outright dismissal from companies such as Gab and Kiwi Farms towards the UK's regulatory measures. If the Government is serious about its mission to halve VAWG in a decade, it would recognise that having legislation that allows Ofcom to enforce changes to platforms is a crucial tactic. Best, better or existing practice? In the consultation guidance, Ofcom articulates its aims as to "summarise the ways different types of content and activity affect women and girls online" and "demonstrate to industry the pressing need" to take action - providing "practical and achievable recommendations" for providers to do so. We strongly believe the industry has all the evidence needed to take action on violence against women and girls, and have done so for years. Companies are well aware of the pressing issues and their inaction has led the UK government to implement parliamentary mandated sector guidance in the first place. Ofcom "hope[s] the draft Guidance will help to secure: a) giving providers a *detailed and holistic* framework for understanding harms to women and girls online; b) setting out *practical and ambitious steps* providers can take to secure women and girls' online safety; and c) *encouraging* service providers to take action to achieve a safer life online for women and girls. As part of these aims, under each action Ofcom includes "foundational steps" and "good practice steps" for providers to take. However the steps included as foundational are more accurately described as "minimum" - this clearly communicates that these steps represent a baseline standard that platforms must meet to be compliant with their duties as set out in the Codes. In a similar vein, "good practice steps" actually refers to existing industry standards. Understandably Ofcom's case studies are 'illustrative' examples however a number of times in the guidance Ofcom mentions providing "good practice steps providers *could take* to further improve the experiences

Question	Your response
	of women and girls"[15]. **Given these are examples of *existing* practice, they do not push the envelope beyond the capacities of the world's largest social media and search services companies - they should therefore be framed as *current* industry standards, rather than stretch '*could do*' targets.** This will allow Ofcom to assess where companies fall behind industry standards in the guidance, and also clearly identify where Ofcom is putting forward steps that go further to *improve* current practices.
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 As mentioned above.
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? – Y / N

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

[15] As mentioned in sections: 2.49, 2.75, 3.26 and case study 24.