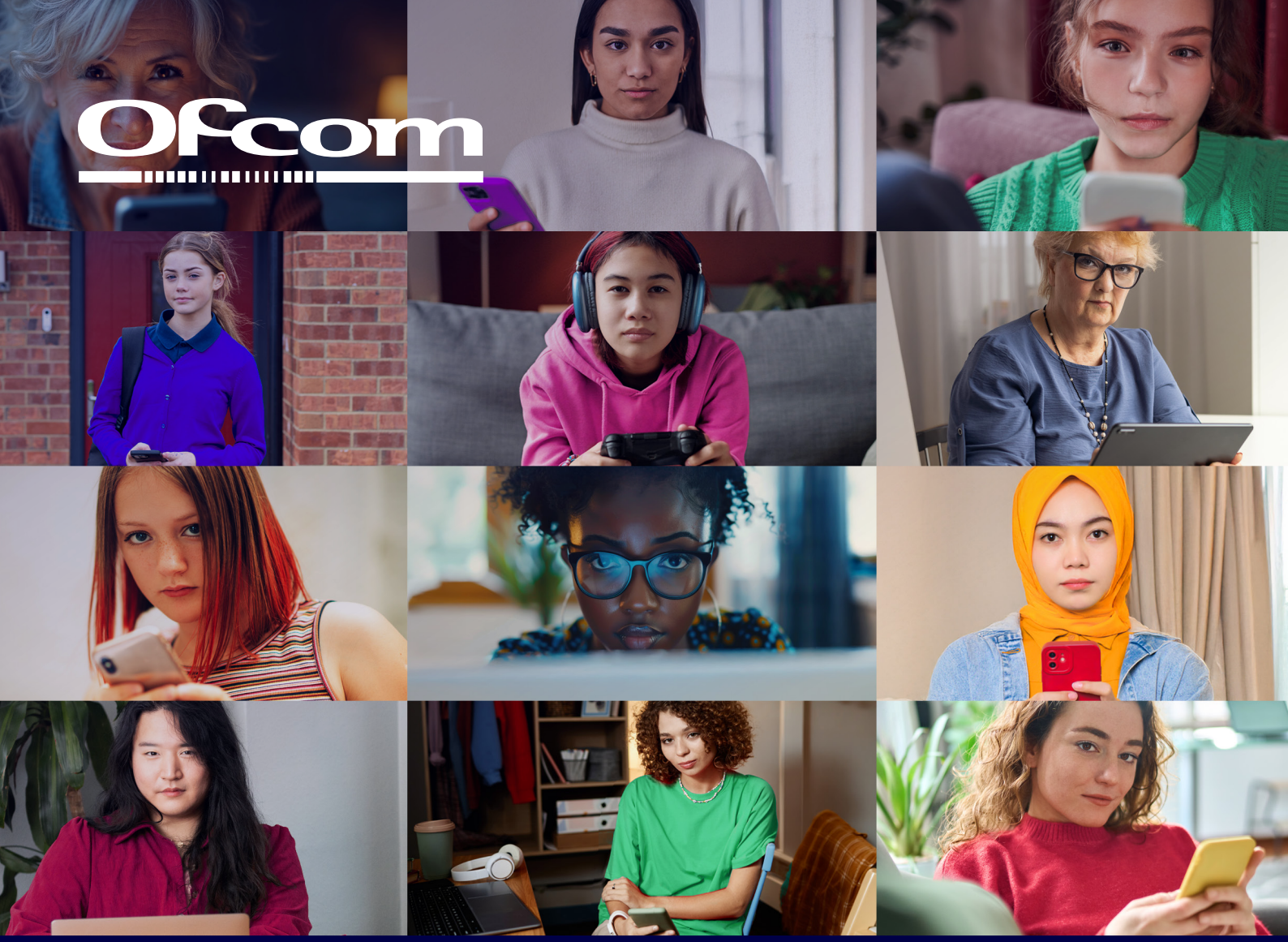

A Safer Life Online for Women and Girls

Practical Guidance for Tech Companies

Published 25 February 2025

Welsh version available

Contents

Section

Annex

1. Introduction

Warning: this chapter contains content that may be upsetting or distressing.

1.1 The Online Safety Act 2023 ('the Act') places clear requirements on online service providers ('service providers')[1] to address illegal gender-based harms such as intimate image abuse, and to protect children from gender-based harms like misogyny and to shield them from pornographic content. Ofcom has already published Codes and risk assessment guidance on how we expect online service providers to tackle illegal content and protect children.[2] Once the duties are in force, Ofcom's role will be to hold providers to account, using our robust enforcement powers as needed. The foundational protections that are set out in our Codes and risk assessment guidance will benefit women and girls as well as all UK users of online services.

1.2 In addition to these safety duties, when passing the Act, Parliament included a requirement on Ofcom to produce additional, dedicated Guidance on women and girls' safety. This Guidance sets out how service providers, including dating apps, social media, gaming, pornography sites[3] and search services, can address content and activity that disproportionately affects women and girls, in recognition of the specific risks they face online.[4]

1.3 Evidence clearly shows that women and girls experience unique and serious risks online. Women and girls are more at risk of image-based sexual abuse, are disproportionately targeted by misogynistic pile-ons and harassment and are the main survivors and victims of online domestic abuse.[5] Throughout the development of this Guidance, we have heard from organisations on the front line of supporting women and girls, experts, and most importantly from survivors and victims themselves that they want to see providers make meaningful, practical changes to their services so they are safer for the millions of women and girls across the UK who use them daily.

1.4 This Guidance sets out nine actions for improving the safety of women and girls online. We illustrate these with practical examples of changes providers could make. To do this, we include relevant 'foundational steps' drawn from the measures set out in our Codes and guidance on Illegal Harms, Protection of Children and Transparency. But we also go further.

---

[1] Such services are defined under Part 3 of the Act and include user-to-user and search services. Throughout this document, we refer to the online platforms themselves as 'services', and the legal entity that provides the service as a 'service provider' or 'provider'.

[2] We published our statement on Illegal Content Codes and Risk Assessment Guidance in December 2024, setting out how services must approach their new duties relating to illegal harms. In April 2025, we will publish the Protection of Children Codes and Risk Assessment Guidance, looking at how services should approach their new duties relating to content that is harmful to children. The draft proposals on the Protection of Children, published for consultation in May 2024, are available here. [accessed 13 February 2025].

[3] The Guidance does not apply to Part 5 services (providers that publish or display pornographic content themselves, with no user-to-user interactions or search content). Ofcom has produced separate guidance for these services.

[4] Section 54 of the Act says the Guidance must focus on 'content and activity….in relation to which such providers have duties set out in [Part 3] or Part 4 of the Act' and 'which disproportionately affects women and girls'.

[5] In Chapter 2 of the Guidance, we provide an overview of key evidence sources on harms to women and girls online. For a more detailed overview of how harms manifest online, including risks to women and girls, see our Illegal Harms Register of Risks, and our draft Children's Register of Risks. [accessed 13 February 2025].

We highlight additional good practice – features, tools and processes – that providers can implement to deliver ambitious and meaningful changes towards a safer life online for women and girls. This strikes a balance between difficult issues. It recognises that some of these harms are not illegal and that providers must have discretion to ensure their users are still able to express themselves freely. It also reflects the real experiences of women and girls and gives voice to their calls for better protections and more choice from the online services they use.

1.5 As the online safety regulator, we plan to closely monitor and shine a light on how service providers choose to protect their users, including how they have applied this Guidance, in order to help users make informed decisions about their online experiences.

1.6 Our Guidance covers the following:

- Chapter 1 introduces concepts used throughout the Guidance.
- Chapter 2 explores how harms that target women and girls manifest women online, and the severe and wide-ranging impacts they have, including on women and girls' ability to freely express themselves and safely participate in life online.
- Chapter 3, Chapter 4, and Chapter 5 set out nine action areas for online service providers to tackle these harms.
- Annex 1 includes a glossary with definitions of the terms we have used throughout the Guidance.

What harms do we focus on in this this document?

1.7 In requiring us to produce this Guidance, the Act sets an expectation that Ofcom focuses on women and girls' experiences online in the round. This includes understanding how gendered harms overlap and how they cut across illegal content and activity,[6] and content and activity harmful to children.[7]

1.8 This content and activity include the harmful ways online services are used to control, exploit, monitor, silence, humiliate, abuse, and threaten women and girls because of their gender. Research shows that those with multiple protected characteristics, such as LGBTQ+ communities and women from ethnic minority backgrounds, face additional harms.

1.9 The impacts are severe and wide-ranging, from inhibiting the safety and participation of women and girls in online spaces – ultimately affecting their ability to engage and express

[6] For a full list of illegal content set out under the act, see Overview of Illegal Harms which defines and categorises over 130 'priority offences'. See our Illegal Harms Register of Risks for detailed discussions of on how illegal harms manifest, including how certain kinds of harm such as intimate image abuse, controlling and coercive behaviour disproportionately affect women and girls. [accessed 13 February 2025].

[7] For a full list of content set out in under the Act which is harmful to children, see Overview of Protecting Children from Harm Online. There are three types of content specified by the Act, some of which disproportionately affects women and girls. Primary priority content refers to pornographic content, content encouraging self-harm, and content encouraging suicide. Priority content refers to a range of harms including abuse content, content inciting hate, content encouraging violence, and bullying content. Non-designated content refers to content, not included in the Act, that presents a material risk of harm to an appreciable number of children. Under the Act, services are required to prevent children from encountering primary priority content that is harmful to children, and to protect children in age groups judged to be at risk of harm from priority content and non-designated content. detailed discussions of content harmful to children, see our draft Children's Register of Risks. [accessed 13 February 2025].

themselves freely online – to the normalisation of misogynistic attitudes and behaviours. Ofcom's research found that women are significantly less likely than men to believe that the benefits of being online outweigh the risks, and to say they can share their opinions and have a voice online. Women and girls are also more likely to be adversely impacted by potential harmful content online. Online cultures that promote actions and attitudes that demean, oppress and otherwise harm women and girls have a broader influence on the offline world. [8]

1.10 For the purpose of this Guidance, we focus on four overlapping forms of harm covering content and activity set out in the Act.

- **Online misogyny:** This describes the circulation – or promotion – of content that actively encourages or cements misogynistic ideas or behaviours, including through the normalisation of sexual violence.
- **Pile-ons and online harassment**: This describes cases where groups of coordinated perpetrators target a specific woman or girl, or groups of women and girls, often with abuse and threats of violence. While pile-ons can happen to any user, they often target women in public life, such as journalists and politicians.
- **Online domestic abuse**: This describes using technology for coercive and controlling behaviour in the context of an intimate relationship.
- **Image-based sexual abuse**: This refers to intimate image abuse (the non-consensual sharing of intimate images) and cyberflashing (sending explicit images to someone without their consent).

1.11 We refer to these collectively as **online gender-based harms**.[9] We provide further detail on how these harms manifest and their impacts in Chapter 2.

1.12 Identifying such a broad range of content and activity which is relevant to the experiences of women and girls online does not mean that we expect all such content to be taken down or heavily policed. However, we consider it is right for providers to take a holistic view of the experiences of women and girls online when making decisions about their policies, tools, and features they offer users, and how those choices may impact women and girls' safety.

What actions are we asking service providers to take?

1.13 The Act states clearly that services should take a safety-by-design approach. This includes making improvements to existing systems on longstanding services or features, as well as ensuring new services or features are designed with safety in mind from the outset.

[8] See our Illegal Harms Register of Risks and our draft Children's Register of Risks for detailed discussions of how online harms manifest and the wider impacts, including how certain kinds of harm such as intimate image abuse, controlling and coercive behaviour, and abusive and hateful content affect women and girls. [accessed 13 February 2025].

[9] This is the term we use to describe our approach to 'content and activity that disproportionately affects women and girls' as set out in section 54 of the Act. There are other commonly used terms to describe these harms, including online violence against women and girls (VAWG) and tech-facilitated gender-based violence. While these other terms refer to a similar subset of harms, we have selected online gender-based harms as it best reflects the scope of the Guidance.

1.14 Using this safety-by-design approach, we set out nine actions providers can take to address the challenge of online-gender based harms across each stage of operating and designing their service.[10] Figure 1 sets out the nine actions, which are split across Chapter 3, Chapter 4, and Chapter 5:

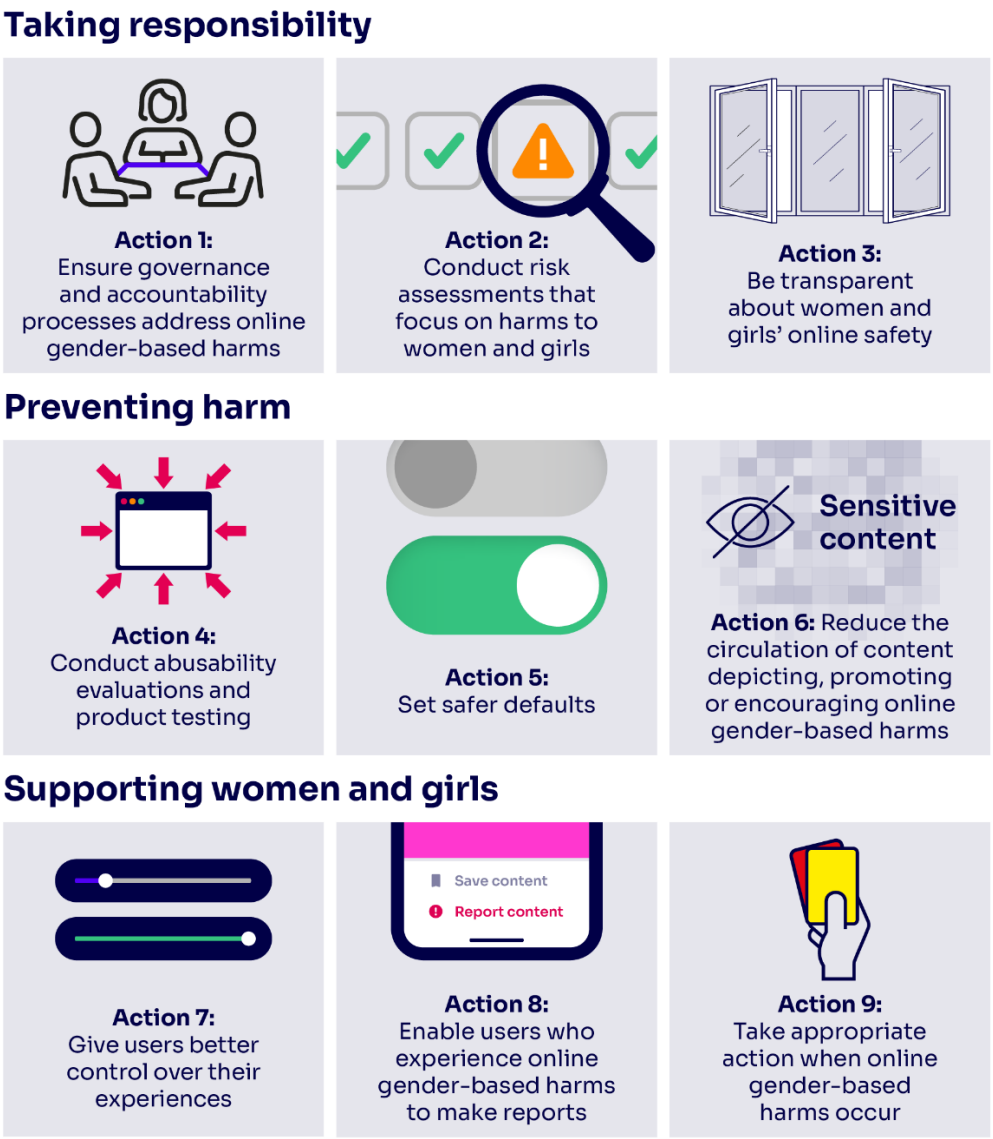

Figure 1: Nine actions areas in Chapter 3, Chapter 4, and Chapter 5

1.15 For each action, we include two types of information.

1.16 First, we explain where we have already set out the measures providers can implement to meet their duties under the Act related to illegal content and protection of children. This information is taken from the Codes and risk assessment guidance we have set out across our work on illegal harms, protection of children, and (only as applicable to a smaller number of providers) transparency. We refer to these collectively as '**foundational steps**.'

[10] Our approach draws on previous work which uses similar concepts to demonstrate how to address online gender-based harms and online safety more generally, including academic literature on safety-by-design and a report by the eSafety Commissioner on tackling 'gendered violence' through safety-by-design. [accessed 13 February 2025].

These steps are wide-ranging and cover areas including conducting risk assessments, [11] transparency reporting[12] and codes related to implementing improvements to user safety.[13] For some of these foundational steps, we include **case studies** to showcase what implementation could look like in a particular context. and how the foundational steps will address harms to women and girls.

1.17 Second, we include '**good practice steps'** which set out practical ways providers can go further to demonstrate a commitment to women and girls' safety. These good practice steps draw on a range of evidence sources including the workshops we ran with a wide range of stakeholders to refine our thinking collaboratively. To develop these proposals, we also drew on academic research, expert reports, industry practice and our work under our Media Literacy duties.[14] For some of these good practice steps, we also include case studies to showcase what implementation could look like in a particular context and how they can address harms to women and girls.

1.18 While the foundational steps are connected to the duties under the Act related to illegal content and protection of children which are enforceable, the good practice steps show practical and feasible ways to go further. We urge providers to implement the good practice steps we have set out to demonstrate a new industry standard prioritising women and girls' safety.

Additional considerations

1.19 The good practice steps are non-exhaustive. We recognise that technology evolves rapidly. There are likely to be ambitious, effective and innovative interventions to secure the safety of women and girls which we have not included or could emerge as safety technology develops.

1.20 We also recognise that harms evolve rapidly and expect providers to regularly look at what they may need to do to respond to changing threats and risks to women and girls. We

---

[11] Our risk assessment guidance is intended to assist services in fulfilling their legal obligations under the Act related to risk assessments. It does not represent a set of compulsory steps that services must take. The Illegal Content Risk Assessment Guidance was published in December 2024. The draft Children's Risk Assessment Guidance was published for consultation in April 2024. [accessed 13 February 2025].

[12] This applies to categorised services only. Section 78(1) Act requires Ofcom to produce guidance for such services about the transparency reporting framework, and we have published draft guidance in accordance with that duty. [accessed 13 February 2025].

[13] Codes describe measures recommended for the purpose of compliance with duties. If service providers implement measures recommended in Codes, services will be treated as complying with the relevant duties. However, the Act allows service providers to adopt alternative measures, provided they keep a record of what they have done and explain how they think the relevant safety duties have been met. The Illegal Content Codes were published in December 2024. At the time of writing, the codes have not completed their parliamentary process and are not yet in force. The draft Protection of Children Codes were published for consultation in May 2024. They are subject to change until we publish the final versions of the Codes in April 2025. [accessed 13 February 2025].

[14] Ofcom has a statutory duty to promote media literacy and to carry out research into media literacy under section 11 of the Communications Act 2003, including specific work related to content and activity that disproportionately affects women and girls. Where relevant, we draw on this duty and our media literacy work to illustrate good practice for online services. In addition, some of these examples of good practice draw on our Media Literacy duties. We define media literacy as "the ability to use, understand and create media and communications across multiple formats and services".

would also expect services with the highest risk and largest reach to need to do more to ensure they have achieved safer experiences for women and girls.

1.21 Finally, some steps may only be relevant or applicable to certain services (including because of their user base size, risk level, user demographics, the design of the service or their business model) and we note this throughout. For example, the way online gender-based harms manifest on search services would be different to how they may manifest on social media services. While search services are unlikely to enable a pile-on or harassment in the same way a social media service could, search services can still be vectors for some types of this content if they make online content such as online misogyny or image-based sexual abuse more accessible and available to wider audiences than they otherwise would be.

2. What are online gender-based harms?

Warning: this chapter contains content that may be upsetting or distressing.

2.1 As noted in the Introduction, we use the term online gender-based harms to refer to harmful content and activity that disproportionately affects women and girls online as set out under the Act. This includes, but is not limited to, relevant illegal content and content harmful to children.[15]

2.2 This chapter explains our understanding of online gender-based harms.

2.3 For further information on the causes and impacts of the specific kinds of harms referenced, see the Illegal Harms Register of Risks and draft Children's Register of Risks. For further information on how we understand what is in and out of scope of these harms, see our Illegal Content Judgements Guidance and our draft Guidance on Content Harmful to Children.

Background

2.4 **Online gender-based harms are systemic and intersectional.** They recreate and amplify existing gender discrimination, sometimes introducing new forms of harm,[16] and intersect with other factors including age, race, ethnicity, socio-economic status, sexual orientation, gender identity, disability, and religion. Evidence indicates that where one or more of these factors apply to women and girls, they experience heightened and specific risks and impacts of harm.[17] [18]

2.5 **Online and offline harms often co-occur and overlap.** This is especially true for forms of harm that include patterns of behaviour. The National Stalking Helpline recently reported that all cases presenting to the helpline include online elements.[19] In addition, different forms of online gender-based harms overlap with one another. For example, intimate image abuse can be a form of coercive control in cases of domestic abuse.[20]

2.6 **Online gender-based harms are exacerbated by societal norms such as victim-blaming.** Women and girls who experience abuse are often told they should have done more to keep themselves safe and can be blamed for being online. This results in an unnecessary burden

[15] For a full list of illegal content set out under the Act, see Overview of Illegal Harms which defines and categorises over 130 'priority offences'. For a full list of content set out in under the Act which is harmful to children, see Overview of Protecting Children from Harm Online. [accessed 13 February 2025].
[16] Global Partnership, 2023. Technology-Facilitated Gender-Based Violence: Preliminary Landscape Analysis. [accessed 18 January 2025]. [accessed 13 February 2025].
[17] Crenshaw, K., 2013. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, *University of Chicago Legal Forum*, 1989 (1). [accessed 24 October 2024].
[18] The Global Partnership, 2023. Technology-Facilitated Gender-Based Violence: Preliminary Landscape Analysis. [accessed 18 January 2025]; Department for Science, Innovation & Technology and PUBLIC, 2025. Platform Design and the Risk of Online Violence Against Women and Girls. [accessed 12 January 2025].
[19] Suzy Lamplugh Trust, 2021. Unmasking Stalking: A Changing Landscape. [accessed 6 February 2025].
[20] Refuge, 2020. The Naked Threat. [accessed 6 February 2025].

on women and girls to respond to, avoid, and cope with gender-based harms.[21] This is known as 'safety work'[22] and hinders their ability to participate freely in online life, if at all.

Categories of online gender-based harms

2.7 To structure the recommendations we make throughout this Guidance, we focus on four categories of online gender-based harms, which evidence suggests disproportionately affect women and girls:

a) Online misogyny,
b) Pile-ons and coordinated harassment,
c) Online domestic abuse, and
d) Image-based sexual abuse, including intimate image abuse and cyberflashing.

2.8 We recognise that these categories do not cover all types of harmful content and activity that may impact women and girls' experiences online,[23] such as child sexual exploitation and abuse, modern slavery and human trafficking, and eating disorder content. These categories of harm are either priority offences for illegal harms or primary priority content for protection of children under the Act. We set out our recommendations for how providers can assess and mitigate risks on these harms in our publications related to those duties. However, we have included information on these harm areas where there are overlaps with our four focus areas.

Online misogyny

2.9 Online misogyny describes a wide range of content and behaviour online which engages in, normalises or encourages misogynistic attitudes and ideas. We discuss illegal online misogyny (such as harassment, threats and abuse, or hate) in our Illegal Harms Register of Risks and discuss online misogyny that is harmful to children (such as abuse and hate, violent or pornographic content) in our draft Children's Register of Risks.

2.10 Online misogyny is perpetrated and witnessed in a variety of online spaces, across both larger services serving many audiences, and smaller services dedicated to proliferating misogynistic views and behaviours. On the former, misogynistic content can consist of hypermasculine narratives about how boys and men should behave and act towards women and girls, often in partnership with broader criticism of feminism, gender messages, or women's rights. Much of this content is produced by users with large followings. The

[21] Vera-Gray, F. and Kelly, L., 2020. Contested gendered space: public sexual harassment and women's safety work, *International Journal of Comparative and Applied Criminal Justice*, 44 (4). [accessed 25 October 2024].
[22] Gillett, R., 2021. "This is not a nice safe space": investigating women's safety work on Tinder, *Feminist Media Studies*, 23 (1). [accessed 25 October 2024].
[23] We recognise that some forms of child sexual exploitation and abuse (CSEA) disproportionately impact girls. For instance, girls are at a greater risk of experiencing grooming and being depicted in child sexual abuse material. We are not focusing on CSEA in this Guidance, but we do highlight CSEA measures set out in the Illegal Content Code for user-to-user and search services as they apply to girls. We also recognise that some of the ways that CSEA manifests online, such as self-generated intimate images and harmful sexual behaviour, overlap with online gender-based violence. For more information see Chapter 2 of our Illegal Harms Register of Risks. [accessed 13 February 2025].

content is framed as entertainment, aligning with interests such as self-improvement or gaming, and using formats such as memes and inspirational stories.[24]

2.11 Das NETTZ highlights the influence of 'misogynistic influencers' on the rise of misogyny in schools in the United Kingdom. Internet Matters found that boys were significantly more likely to have viewed content from such influencers, and are significantly more likely to have a positive view of the content they produce. Increased engagement with misogynistic content has been linked to unhealthy perceptions of relationships; children and young people who reported exposure in a survey were five times more likely to agree with the statement that "hurting someone physically is okay if you say sorry after hurting them."[25]

2.12 Research has also found that young people searching for friends, advice or shared groups are served content that is increasingly misogynistic through their recommender feeds.[26] Young people who are lonely, isolated, or who have mental health concerns can be drawn into more radical and misogynistic content, and find social structure in dedicated online communities.[27] Though they vary in size, ideology, privacy and organisation, such communities are alike in their promotion, imagining and organisation of highly misogynistic attitudes and behaviours, often alongside other discriminatory views.[28]

2.13 The Institute for Strategic Dialogue reiterates that online gender-based harms occur in a continuum, and so misogynistic behaviour that begins online can lead to the perpetration of offline violence, in both public and private spaces.

2.14 Girls also report negative online experiences including bullying, hateful comments, receiving sexual messages from men and other people they do not know online. These experiences are accompanied by a reported feeling of social pressure to be visible online by sharing and engaging with content despite having to navigate unwanted comments or male attention when they do so.[29]

Misogyny and sexually explicit content

2.15 Some misogynistic content overlaps with sexually explicit content. Research from Durham University found that 12%, or one in every eight titles, of analysable content on the landing

[24] Women's Aid, 2023. Influencers and Attitudes: How will the next generation understand domestic abuse?. [accessed 24 October 2024]; Regehr, K., Shaughnessy, C., Zhao, M. and Shaughnessy, N., 2024. Safer Scrolling: algorithms popularise and gamify online hate and misogyny for young people. [accessed 10 March 2025].
[25] Women's Aid, 2023. Influencers and Attitudes: How will the next generation understand domestic abuse?. [accessed 24 October 2024].
[26] Internet Matters, 2023. "It's really easy to go down that path": Young people's experiences of online misogyny and image-based abuse. [accessed 6 January 2025]; Vodafone, 2024. AI 'Aggro-rithms': young boys are served harmful content within 60 seconds of being online. [accessed 30 December 2024].
[27] Griffin, J., 2021. Incels: Inside a dark world of online hate, BBC News, 13 August. [accessed 11 February 2025]; Das NETTZ, Textgain, and Federal Association for Countering Online Hate, 2024. Tracing Online Misogyny: An analysis of misogynist ideologies and practices from a German-international perspective. [accessed 10 January 2025]; Vodafone, 2024. AI 'Aggro-rithms': young boys are served harmful content within 60 seconds of being online. [accessed 30 December 2024].
[28] Das NETTZ, Textgain, and Federal Association for Countering Online Hate, 2024. Tracing Online Misogyny: An analysis of misogynist ideologies and practices from a German-international perspective. [accessed 10 January 2025].
[29] Ofcom, 2024. Draft Children's Register of Risks. [accessed 10 January 2025].

pages of the top three user-to-user adult services in the UK described sexual activity that constituted sexual violence.[30] [31]

2.16 Research looking at depictions of sexual violence argues that the availability of this content can contribute to 'sexual scripts' about what behaviours and attitudes are acceptable or pleasurable.[32] This can lead to normalising harmful and coercive behaviour and users being less likely to understand what sexual acts are unlawful or harmful. It can contribute to sexual violence being more easily dismissed or going unreported.[33]

2.17 In one study, 16-21-year-olds expressed concern about the implications of pornography in distorting their understanding of the difference between sexual pleasure and harm.[34] This can have a range of impacts, including the normalisation of harmful sexual behaviour such as sexual aggression towards girls and women.[35] Evidence suggests that pornography also affects attitudes towards consent, as it is implied (rather than discussed) in pornographic content.[36]

Pile-ons and online harassment

2.18 We examine illegal harassment in depth in the Illegal Harms Register of Risks and Illegal Content Judgements Guidance published in our December 2024 Statement. Additionally, we will consider relevant content harmful to children in our Children's Register of Risks which will be published in April 2025.

2.19 Pile-ons and harassment cover behaviours that involve many users targeting an individual victim or group of victims with abusive, hateful or threatening content, often repetitively or

---

[30] For their study, the authors used the World Health Organisation definition of sexual violence, which is likely to include both extreme pornographic content and legal pornographic content. They focused on four broad categories of sexual violence: sexual activity between family members; aggression and assault; image-based sexual abuse and coercive and exploitative sexual activity. Source: Vera-Gray, F., McGlynn, C., Kureshi, I. and Butterby, K., 2021. Sexual violence as a sexual script in mainstream online pornography, *The British Journal of Criminology,* 61 (5). [accessed 24 October 2024].

[31] Some of this content may meet the threshold for illegal content, including extreme pornography. Extreme pornography describes a category of illegal sexual material which includes content that depicts non-consensual behaviours, physical violence and threats to life. For more information, see the Illegal Harms Register of Risks. [accessed 13 February 2025].

[32] Marshall, E., Miller, H. and Bouffard, J., 2018. Bridging the Theoretical Gap: Using Sexual Script Theory to Explain the Relationship Between Pornography Use and Sexual Coercion, *Journal of Interpersonal Violence,* 36 (9- 10). [accessed 29 October 2024].

[33] Researchers have argued that the availability of extreme pornographic content – including rape and non-consensual sexual penetration – sustains a culture in which sexual violence is not taken seriously and risks being normalised. Source: Vera-Gray, F., McGlynn, C., Kureshi, I., Butterby, K., 2021. Sexual violence as a sexual script in mainstream online pornography, *The British Journal of Criminology*, 61 (5). [accessed 29 October 2024].

[34] 'For example, a younger individual who is not fully developed could find pornography that reinforces abusing women, and they might begin to think that is what women find pleasurable'. Source: Children's Commissioner, 2023. 'A lot of it is actually just abuse'- Young people and pornography. [accessed 20 February 2025].

[35] Research from the UK Government notes that those perpetrating harmful sexual behaviours are influenced by multiple factors which could include viewing pornography, but it is never one factor alone that leads to this behaviour. Source: Government Equalities Office (Upton, J., Hazell, A., Abbott, R. and Pilling, K.), 2020. The relationship between pornography use and harmful sexual behaviours. [accessed 29 October 2024]. See also: BBFC and Revealing Reality, 2020. Young people, Pornography & Age-verification. [accessed 6 January 2025].

[36] BBFC and Revealing Reality, 2020. Young people, Pornography & Age-verification. [accessed 6 January 2025].

at scale.[37] In the context of gender-based harms, such behaviours are often misogynistic, involving sexualisation, threats, descriptions of rape, or intimate image abuse, including the creation and sharing of deepfakes.[38]

2.20 Women and girls from marginalised groups face heightened risks of pile-ons because of their race or ethnicity,[39] as well as their gender identity or sexuality.[40] Their experiences of seeking support are also often affected by overlapping forms of discrimination.[41]

2.21 Women and girls in public life such as celebrities, politicians, journalists, influencers and activists face heightened risks of pile-ons and harassment.[42] They are often targeted by gendered mis- and disinformation campaigns which weaponise false narratives to achieve social, political or economic aims, though not all mis- and disinformation is organised.[43]

2.22 Nearly three quarters (73%) of women journalists answering the ICFJ survey had experienced some form of online violence in the course of their work. Journalists can receive thousands of abusive posts, often in response to reporting on politics or elections, as well as gender. Posts discredit their work and demean them personally. Often, they are called "stupid" or "hysterical", their personal lives are investigated and criticised, and 1 in 4 report receiving threats of physical violence.[44]

2.23 However, women and girls can face harassment and pile-ons even if they are not in the public eye; being present or participating online in any way presents a risk of being targeted.[45] Plan International's 2020 report found that 58% of girls and young women surveyed had personally experienced some form of online harassment. In some instances,

---

[37] Demos (Judson, E.), 2021. Silence, Woman: An investigation into gendered attacks online. [accessed 13 February 2025].
[38] Demos (Judson, E.), 2021. Silence, Woman: An investigation into gendered attacks online. [accessed 13 February 2025].
[39] In a study by Amnesty International, Black, Asian and Minority Ethnic (BAME) women MPs received almost half (41%) of the abusive tweets, despite there being almost eight times as many white MPs in the study. Source: Amnesty International UK, 2017. Black and Asian Women MPs Abused More Online. [accessed 6 February 2025]. See also: Ofcom, 2024. Experiences of using online services. [accessed 6 February 2025].
[40] Ofcom's Online Experiences Tracker data found that reports of stalking, cyberstalking or harassing behaviour are higher among transgender women and non-binary people (16%) compared to cisgender respondents (4%). Source: Ofcom, 2023. Experiences of using online services. [accessed 6 February 2025]; According to 2016 US research, 33% of the LGBTQ+ individuals sampled had been sexually harassed online, compared to 6% of heterosexual people. Source: Data & Society Research Institute and CiPHR (Lenhart, A., Ybarra, M., Zickuhr, K. and Price-Feeney, M.), 2016. Online Harassment, digital abuse and cyberstalking in America. [accessed 6 February 2025].
[41] Demos (Judson, E., Atay, A., Krasodomski-Jones, A., Lasko-Skinner, R. and Smith, J.), 2020. Engendering Hate: The Contours of State-Aligned Gendered Disinformation Online. [accessed 25 October 2024]; Demos (Judson, E.), 2021. Silence, Woman: An investigation into gendered attacks online. [accessed 13 February 2025]; Di Meco, L. and Wilfore, K., 2021. Gendered disinformation is a national security problem, *Brookings,* 8 March. [accessed 25 October 2024]; Wilson Center (Jankowicz, N., Hunchak, J., Pavliuc, A., Davies, C., Pierson, S. and Kaufmann, Z.), 2021. Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online. [accessed 25 October 2024].
[42] HM Government Stabilisation Unit, 2020. Quick-read guide: gender and countering disinformation. [accessed 25 October 2024]; US Department of State, 2023. Gendered Disinformation: Tactics, Themes, and Trends by Foreign Malign Actors. [accessed 25 October 2024].
[43] Wilson Center (Jankowicz, N., Hunchak, J., Pavliuc, A., Davies, C., Pierson, S. and Kaufmann, Z.), 2021. Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online. [accessed 25 October 2024].
[44] International Center for Journalists (Posetti, J. and Shabbir, N.), 2022. The Chilling: A global study of online violence against women journalists. [accessed 22 October 2024].
[45] The Global Partnership, 2023. Technology-Facilitated Gender-Based Violence: Preliminary Landscape Analysis. [accessed 13 February 2025].

the harassment was co-ordinated on dedicated sites where perpetrators would randomly select women and girls to cyberstalk, finding and sharing any available information about them.[46]

2.24 Intimate image abuse often plays a role in online harassment of women in and out of the public eye. The Revenge Porn Helpline recently reported that 63% of survivors and victims who had experienced harassment online were women, and this often involved the sharing, or threatening to share, of intimate images.[47]

Online domestic abuse

2.25 Domestic abuse encompasses using technology for the offence of coercive and controlling behaviour in the context of an intimate relationship.[48] It can include stalking and harassment. These harms are covered under the Act and we examine their causes and impacts in the Illegal Harms Register of Risks.

2.26 Technology exacerbates existing dynamics of power and control. Abuse can now be immediate, constant and reach a broad social network with minimal effort, having a faster and greater impact on different spheres of the victim's life.[49] Online domestic abuse includes a range of controlling behaviours, such as:

a) Device and app control: unauthorised access to a person's online accounts, by guessing passwords, or hacking.[50]
b) Monitoring and surveillance: monitoring messages and posts and identifying location information through metadata, shared images, or applications which track location.[51]
c) Impersonation, including catfishing: assuming the identity of a person to access private information; exploit, embarrass, discredit or shame them; contact or mislead them; or create fraudulent documents. This can happen via account hacking and can also include the online theft of documents.[52]
d) Public disclosure of private information and doxing: nonconsensual publication of intimate images or private information online, such as a person's address or phone number.[53] This can be particularly harmful in the context of honour-based violence,

[46] Plan International, 2020. Free to be Online? Girls' and young women's experiences of online harassment. [accessed 6 January 2025].
[47] Revenge Porn Helpline (Papachristou, K.), 2023. Revenge Porn Helpline: 2023 Report. [accessed 7 January 2025].
[48] The term 'tech-driven domestic abuse' can also be used to refer this offence. Many of these controlling behaviours, such as stalking and harassment, can also occur outside of an intimate relationship.
[49] Fernet, M., Lapierre, A., Hébert, M. and Cousineau, M., 2019. A systematic review of literature on cyber intimate partner victimization in adolescent girls and women, *Computers in Human Behaviour,* 100. [accessed 13 February 2025].
[50] Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T. and Dell, N., 2018. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology, *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* [accessed 28 October 2024].
[51] Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T. and Dell, N., 2018. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology, *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* [accessed 28 October 2024].
[52] Henry, N., Vasil, S., Flynn, A., Kellard, K. and Mortreux, C., 2021. Technology-Facilitated Domestic Violence Against Immigrant and Refugee Women: A Qualitative Study. *Journal of Interpersonal Violence*, 37. [accessed 13 February 2025].
[53] The Global Partnership, 2023. Technology-Facilitated Gender-Based Violence: Preliminary Landscape Analysis. [accessed 13 February 2025].

where leaked information or the threat of leaked information can lead to additional shame, stigma and violence.[54]

e) Psychological or emotional abuse or threats: harassment and stalking, including sexual and/or dating harassment behaviours. This can include:

   i) Unwanted contact: sending the survivor and victim repeated messages.
   ii) Gaslighting: making the survivor and victim question their feelings and experiences.

2.27 Online domestic abuse also co-occurs with many other forms of harm. The majority of stalking, intimate image abuse and harassment reported by women is carried out by a current or former partner.[55] Many of the techniques used for coercive control, such as monitoring and harassment, may overlap with methods used for human trafficking and sexual exploitation.[56] Often different abuse types are co-occurring across multiple devices and platforms over many years. Considering these harms in isolation can underestimate the scale and impact of abuse.

2.28 Coercive and controlling behaviour is heavily under-reported. Half of survivors (49%) responding to a survey by Refuge said they told no one about the abuse and only a small proportion of women (13%) reported the abuse to the social media platform they experienced the abuse on. Only 1 in 10 survivors (10%) felt empowered to report to the police. Under-reporting happens due to shame about the abuse and lack of trust in platforms or the police to address the problem.[57]

2.29 Coercive and controlling behaviour can be particularly hard to identify compared to some other forms of abuse. For example, a picture of a front door may seem innocuous to an outside observer or content moderator. However, for a survivor and victim who is fleeing an abusive partner and moving to a safe, secret location, receiving such a picture could be a sign that the user knows where they are. This can be traumatic for survivors and victims, with women feeling physically unsafe because the perpetrator knows their location.[58] Furthermore, perpetrators can psychologically manipulate the victim to sow self-doubt and confusion in their mind to conceal the abuse (gaslighting) which makes it harder to identify.

2.30 Issues with reporting and identifying coercive and controlling behaviour point to the importance of qualitative work and engagement with support services to understand the nuances of harm in this area.

---

[54] End Violence Against Women Coalition and Faith and VAWG Coalition, 2021, Response to the Law Commission Consultation on Intimate Image Abuse. [accessed 17 December 2024].

[55] Refuge, 2021. Unsocial Spaces. [accessed 13 February 2025]; Flynn, A., Powell, A., Scott, A. and Cama, E., 2022. Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse. *The British Journal of Criminology*, 62 (6). [accessed 28 October 2024].

[56] Office of Family Violence Prevention and Services (Dabby, C.), 2019. Domestic Violence and Human Trafficking: Advocacy at the Intersections. [accessed 5 February 2025].

[57] Flynn, A., Powell, A., Scott, A. and Cama, E., 2022. Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse. *The British Journal of Criminology*, 62 (6). [accessed 28 October 2024]; Ofcom Stakeholder Workshop 1 on Women and Girls Online Safety, 16 September 2024.

[58] Refuge, 2021. Unsocial Spaces. [accessed 13 February 2025].

Image-based sexual abuse

2.31 Image-based sexual abuse refers to the taking, creating, sharing, or threatening to share intimate images without consent.[59] We use the term to refer to both intimate image abuse offences and the offence of cyberflashing. These harms are covered by the Act and we examine their causes and impacts in the Illegal Harms Register of Risks.

2.32 Image-based sexual abuse is not simply a product of online spaces, it is a manifestation of existing structures of sexual violence and misogyny. Image-based sexual abuse can form part of a pattern of harmful behaviour, motivated by a desire to exert power.

Intimate image abuse

2.33 Intimate image abuse refers to sharing or threatening to share intimate images without consent. Intimate images are often sexually explicit but may also include intimate scenarios based on specific cultural and religious contexts.[60]

2.34 Threats to share intimate images and non-consensual sharing of intimate images can be perpetrated as part of a pattern of coercive and controlling behaviour both online and offline.[61] Non-consensual sharing of intimate images can also be perpetrated as part of a 'collector culture' where users exchange and discuss intimate images, often in dedicated misogynistic spaces online.[62] Intimate images shared without consent may then be re-shared by other users. This re-sharing means the images continue to circulate, sometimes across multiple platforms, causing re-victimisation and re-traumatisation for survivors and victims.

2.35 Intimate image abuse can include both images that were taken or made consensually and images that were taken or made without consent. This includes images which have been artificially generated or manipulated, including deepfakes.[63]

2.36 Deepfake intimate images are an exponentially growing harm, with more deepfake intimate image abuse posted online in 2023 than in every previous year combined.[64] This growth is driven by the availability of new tools powered by Generative AI (GenAI), which makes it easier for users to create realistic and life-like deepfake content, and a wider 'deepfake economy'. This includes apps and sites that offer 'nudification' and websites that are dedicated to hosting deepfake intimate image abuse.[65] Their user base has dramatically increased – one site reportedly receives 17 million views monthly – and they are easily accessed through internet search engines.[66]

---

[59] McGlynn, C. and Rackley, E., 2017. Image-Based Sexual Abuse, *Oxford Journal of Legal Studies*, 37 (3). [accessed 28 October 2024].
[60] Muslim Women's Network UK, 2021. Written Evidence to the Law commission in Respect of its Consultation on Intimate Image Abuse. [accessed 28 October 2024].
[61] Refuge, 2020. The Naked Threat. [accessed 28 October 2024].
[62] Moore, A., 2022. 'I have moments of shame I can't control': the lives ruined by explicit 'collector culture', The Guardian, 6 January. [accessed 28 October 2024].
[63] Ofcom, 2024. Deepfake Defences: Mitigating the Harms of Deceptive Deepfakes. [accessed 28 October 2024].
[64] My Image My Choice, 2024. Deepfake Abuse: Landscape Analysis 2023-24. [accessed 28 October 2024].
[65] Ofcom, 2024. Deepfake Defences: Mitigating the Harms of Deceptive Deepfakes. [accessed 28 October 2024].
[66] Tenbarge, K., 2023, Found through Google, bought with Visa and Mastercard: Inside the deepfake porn economy, NBC, 27 March. [accessed 28 October 2024].

2.37 Intimate image abuse has a profound negative impact on survivors and victims, including psychological, emotional, professional, and relational harm.[67] For many survivors and victims of this kind of abuse, the impacts are devastating and all encompassing, representing a "social rupture" that changes their lives irrevocably from that point onwards.[68]

2.38 Intimate image abuse cuts across other harms: for example, it can be used to intimidate and discredit women in public life during pile-on harassment, or as a form of coercion and control in situations of domestic abuse. Collector culture forums for sharing intimate images are also a part of online misogyny communities. Women can be coerced and exploited through the creation and circulation of intimate or sexual content, including through livestreams.[69]

2.39 Intimate image abuse will also affect women in different situations differently: for example, sex workers' and adult content creators' experiences of intimate image abuse are often missed in safety interventions. Sex workers often experience blackmail, threats of exposure, recording without knowledge, and deepfakes.[70] Sex workers can experience victim blaming and stigma. In cases where intimate images are monetised, consensual image exchange still becomes abuse when images are shared further without consent.[71]

Cyberflashing

2.40 The cyberflashing offence set out in the Act refers to the sending or giving of a photograph or film of genitals to another person: (a) with the intention that person will see the genitals and be caused alarm, distress, or humiliation; or (b) to obtain sexual gratification (and the person is reckless as to whether the other person will be caused alarm, distress or humiliation).[72] The term cyberflashing is also used more generally to refer to the sending of sexual images without consent.

2.41 Cyberflashing breaches the privacy and sexual autonomy of the receiver of the image and can have negative psychological impacts on those targeted. Many survivors and victims describe the experience as aggressive or intimidating and feel frightened or vulnerable as a result.[73] Evidence suggests that women in minority ethnic groups and LGBTQ+ groups disproportionately experience cyberflashing.[74] Cyberflashing is part of a wider harm in which women, particularly women in minoritised groups, are sexualised without consent in ways that harm their sexual autonomy.

---

[67] Flynn, A., Powell, A., Scott, A. and Cama, E., 2022. Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse, *The British Journal of Criminology*, 62 (6). [accessed 28 October 2024].

[68] McGlynn, C., Johnson, K., Rackley, E., Henry, N., Gavey, N., Flynn, A. and Powell, A. 2021. 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse, *Social & Legal Studies*, 30 (4). [accessed 10 January 2025].

[69] United Nations Office on Drugs and Crime, 2020. Global Report on Trafficking in Persons. [accessed 7 January 2025]; Office of Family Violence Prevention and Services (Dabby, C.), 2019. Domestic Violence and Human Trafficking: Advocacy at the Intersections. [accessed 5 February 2025].

[70] Sanders, T., Trueman, G., Worthington, K. and Keighley, R., 2023. Non-consensual sharing of images: Commercial content creators, sexual content creation platforms and the lack of protection, *New Media & Society*, 27 (1). [accessed 28 October 2024].

[71] National Ugly Mugs, 2022. Visual Violence: Sex workers' experiences of image-based abuses. [accessed 28 October 2024].

[72] See section 187 of the Act which amends the Sexual Offences Act 2003 by inserting section 66A.

[73] Law Commission, 2022. Intimate image abuse: a final report. [accessed 28 October 2024].

[74] McGlynn, C., 2021. Written evidence submitted by Professor Clare McGlynn, Durham Law School, Durham University. [accessed 10 January 2025]; McGlynn, C. and Johnson, K., 2022. Cyberflashing: Recognising Harms, Reforming Laws. [accessed 13 February 2025].

3. Taking responsibility

Warning: this chapter contains content that may be upsetting or distressing.

Overview

Context

3.1 Taking responsibility for online gender-based harms means establishing overarching governance and accountability processes that ensure women and girls' online safety is at the heart of design choices. Good governance and risk assessment are central to all aspects of online safety, and to service providers' compliance with the risk assessment duties in the Act.[75]

3.2 This chapter looks at actions providers can take to address online gender-based harms in their governance and risk assessments. We specifically look at how providers can take the following actions:

- Action 1: Ensure governance and accountability processes address women and girls' online safety.
- Action 2: Conduct risk assessments that focus on harms to women and girls.
- Action 3: Be transparent about women and girls' online safety.

3.3 For each action, we set out a solid baseline of what safety looks like ('**foundational steps'**) to help service providers meet their new duties to protect UK users. We also highlight additional **good practice steps** to illustrate how providers can take an ambitious approach to women and girls' online safety.

3.4 We see the application of these actions as an ongoing exercise where providers can continually assess and improve the experiences of women and girls on their service.

Our target outcomes

3.5 When service providers adopt a safety-by-design approach to women and girls' online safety, all forms of online gender-based harms are considered throughout the product development lifecycle.[76] Improving safety practices for women and girls will also likely have knock-on effects on improving the experience of other marginalised and vulnerable users who experience disproportionate rates of abuse, by ensuring user safety is central to service design .[77]

3.6 Gender-sensitive governance and risk assessment processes can lead providers to better understand and anticipate risks to users, increasing the likelihood they are prioritised and

[75] For more information, see our Illegal Content Risk Assessment Guidance, and our draft Children's Risk Assessment Guidance. [accessed 12 February 2025].
[76] Strohmayer, A., Slupska, J., Bellini, R., Neff, G., Coventry, L., Hairston, A. and Dodge, A, 2021. Trust and Abusability Toolkit: Centering Safety in Human-Data Interactions. [accessed 22 October 2024]; Chatham House (Wilkinson, I., Hofstetter, J.S., Shires, J. and Yahaya, M.S.), 2024. The role of the private sector in combatting gendered cyber harms. [accessed 22 October 2024].
[77] Constanza-Chock, S., 2020. Design Justice. [accessed 22 October 2024].

mitigated appropriately, and factored into strategic decision making. Being transparent about these processes makes the outcomes of the decisions accessible to external actors to increase accountability.

3.7 These governance and organisational design processes should also help providers in preparing to deal with changes in the online landscape that may increase risks to users, including sudden spikes in illegal content and sensitive events. They could also assist in monitoring and reviewing the effectiveness of measures designed to reduce risk.

Action 1: Ensure governance and accountability processes address online gender-based harms

3.8 Effective governance and accountability processes provide the foundation for service providers to identify, manage, and review risks to their users. By embedding accountability, oversight, independence, transparency, and clarity of purpose into their operations, we expect providers to have well-functioning governance and organisational design processes. This could ultimately lead providers to respond more effectively to online gender-based harms, for example through senior leaders setting it as a priority.

3.9 Active and representative leadership is critical for a service provider to effectively respond to online gender-based harms that exist on or are facilitated by its service.[78] Senior leaders should be engaged with, and educated about online gender-based harms and build this culture with relevant staff and decision-makers.

3.10 Governance and accountability measures can also support organisational transformation to prioritise safety-by-design. This will be reflected in the policies service providers set, both internally (for example in learning & development goals) and externally (for example in Terms of Service). Ultimately, it is up to services to decide which policies will be most appropriate for their service. However, in this section we include recommendations and evidence about good practice.

Foundational steps: What are the expectations for service providers?

3.11 Our Codes set out the following steps for service providers:[79]

a) **Board review**: A provider's most senior governance body should carry out and record annual reviews of risk management activities in relation to online safety, and how developing risks are being monitored and managed.[80]

[78] Carnegie UK, The End Violence Against Women Coalition, Glitch, NSPCC, Refuge, 5Rights, Woods, L. and McGlynn, C., 2022. Violence Against Women and Girls (VAWG) Code of Practice. [accessed 22 October 2024]; Department for Science, Innovation & Technology and PUBLIC, 2025. Platform Design and the Risk of Online Violence Against Women and Girls. [accessed 12 January 2025].

[79] The 'foundational steps' refer to a range of expectations we have already set out for service providers, either in draft for the purpose of consultation or in final form at set out in a decision at the time of publication of this Guidance. Where the measures which are currently in draft are changed in our final statement following the consultation on those measures, we will revise this Guidance accordingly. The 'foundational steps' apply to different services based on functionality, risk, and size. Information about which foundational step applies to which service is set out in the Guidance at a Glance document.

[80] Illegal Content (ICU A1 / ICS A1), draft Protection of Children (PCU A1 / PCS A1).

b) **Accountable individual:** Name an individual accountable to their most senior governance bodies for compliance with online safety duties.[81]
c) **Written statements of responsibilities** for senior managers who make decisions related to the management of online safety risks.[82]
d) **Internal monitoring and assurance function** to provide independent assurance that measures taken to mitigate and manage the risks of harm to individuals identified in the risk assessments are effective on an ongoing basis. [83]
e) **Monitoring trends:** Service providers should track evidence of new kinds of illegal content and content that is harmful to children on their services, and unusual increases in particular kinds of illegal content and content that is harmful to children.[84]
f) **Codes of Conduct** that set standards and expectations for individuals working for the providers around protecting users from online safety risks.[85]
g) **Terms of Service and Publicly Available Statement**: Terms and statements have clear and accessible provisions on how users are protected from illegal content (including illegal harms that disproportionately affect women and girls, such as stalking and intimate image abuse) as well as content harmful to children.[86]
h) **Compliance training:** Staff involved in the design and operational management of a service are sufficiently trained in the service's approach to compliance with online safety duties.[87]

3.12 **Case study 1** includes examples of how a service provider can ensure its governance systems capture the risk of online gender-based harms.

Case study 1: Governance and accountability

Ensuring that considerations for online gender-based harms are embedded within the organisation and its systems and processes could include:

- A senior person within the service provider being accountable to the most senior governance body for ensuring the service considers and addresses online gender-based risks.
- Terms of service and/or community guidelines that are specific and accessible. This should include regular, systemic reviews to ensure that they remain effective, proportionate, and responsive to developing trends in online gender-based harms.
- Monitoring and assurance focused on tracking emerging threats, such as deepfakes and other developments in GenAI-enabled abuse, and evaluating whether safety measures adequately address them. Please see Chapter 4 for further details on how products can be tested for resilience to emerging threats.

[81] Illegal Content (ICU A2 / ICS A2), draft Protection of Children (PCU A2 / PCS A2).
[82] Illegal Content (ICU A3 / ICS A3), draft Protection of Children (PCU A3 / PCS A3).
[83] Illegal Content (ICU A4 / ICS A4), draft Protection of Children (PCU A4 / PCS A4).
[84] Illegal Content (ICU A5/ ICS A5), draft Protection of Children (PCU A5 / PCS A5).
[85] Illegal Content (ICU A6 / ICS A6), draft Protection of Children (PCU A6 / PCS A6).
[86] Illegal Content (ICU G3 / ICS G3), draft Protection of Children (PCU D3 / PCS D3). Terms of Service refer to user-to-user services, while Publicly Available Statements refer to search services.
[87] Illegal Content (ICU A7 / ICS A7), draft Protection of Children (PCU A7 / PCS A7). In this context, staff refers to individuals working for the company.

Good practice steps: How can service providers go further?

3.13 In addition to the foundational steps which provide a solid baseline for safety, providers can go further by demonstrating both internally and externally their commitment to women and girls' online safety and their willingness to improve in response to feedback. This could include:

a) **Setting policies** that are designed to tackle forms of online gender-based harms that are prevalent on the service (see **Case study 2**).[88] This could include defining and prohibiting:

   i) Forms of gendered harm such as stalking, harassment and intimate image abuse.
   ii) Gendered abuse affecting specific groups, such as misogynoir (hate directed at Black women and girls)[89] and deliberate misgendering (referring to someone, especially a transgender person, using a word, especially a pronoun or form of address, that does not reflect their gender identity).
   iii) Promotion of offsite abuse, such as promoting deepfake creation services like nudification apps, or forums sharing explanations of how to monitor your partner.

b) **Considering intersectionality:** Ensuring that governance and decision-making processes consider intersectionality of online harms, or how people with multiple characteristics, including gender, race or disability, experience disproportionate risks on services. This understanding can help inform governance and decision-making processes by avoiding different types of harm being addressed in isolation.
c) **Consulting with subject matter experts**, particularly those with experience of supporting survivors of gender-based harms, when setting policies and terms of service.
d) **Training staff** involved in setting policies or governance and decision-making processes on online gender-based harms and safety-by-design.
e) **Creating a media literacy-by-design policy** to promote critical and informed use of its service, as set out in Ofcom's Best Practice Principles for Media Literacy by Design. [90]
f) **Establishing an oversight mechanism** for trust and safety decisions (see **Case study 3**).[91]

Case study 2: Sexualised harassment policy

Women, particularly women from ethnic minority backgrounds, can be sexualised and fetishised online in ways which harm their sexual autonomy.[92] This can include unsolicited sexual messages, sexual deepfakes, and images reposted with sexual comments that fetishise or degrade women. Sometimes sexualisation can be implicit or highly contextual.

[88] Service providers should have regard to the needs of their UK user base in considering what languages are needed to ensure their policies are accessible.
[89] Glitch, 2023. The Digital Misogynoir Report: Online abuse against Black women allowed and enabled to thrive. [accessed 22 October 2024]; Bailey, M., 2021. Misogynoir Transformed: Black Women's Digital Resistance. [accessed 28 October 2024].
[90] Ofcom, 2024. Best Practice Design Principles for Media Literacy. [accessed 22 October 2024].
[91] LEAF (Khoo, C.), 2021. Deplatforming Misogyny: Report on Platform Liability for Technology-Facilitated Gender-Based Violence. [accessed 30 January 2025].
[92] Glitch, 2023. The Digital Misogynoir Report: Online abuse against Black women allowed and enabled to thrive. [accessed 22 October 2024].

Where a social media provider notices users resharing women's posts in ways which sexualise them without their consent, the company can set out a harassment policy which makes it clear that sexualising someone without their consent is a violation of the policy.

Having clear, specific policies on the nuances of online gender-based harms is an important step for creating safer spaces for women online. This is particularly important for nonconsensual sexualisation, which can often be normalised so much that it goes unnoticed.

Case study 3: External oversight

Service providers' community guidelines and content moderation decisions make them powerful arbiters of public speech. This can have serious implications for users, particularly in cases where biases around characteristics such as gender and race are embedded in service providers' policies, such as content moderation guidelines.

For example, content moderation decisions based on an over-zealous application of policies have led to removal of posts about mothers breastfeeding, LGBTQ+ couples kissing, or survivors sharing their experiences of sexual violence.[93] Likewise, algorithms used for content moderation may have biases embedded in their training data, leading to biased outcomes in which women from ethnic minority backgrounds are disproportionately policed for online speech.[94]

To introduce accountability and oversight for bias in decisions and policies, companies can engage with external experts or set up an external appeals ombudsman. Such an ombudsman could accept complaints from users appealing content moderation decisions and provide feedback that helps clarify a provider's policies.

This process enables service providers to have clearer and more consistent rules for content moderation, leading to safer experiences for women and girls online.

Action 2: Conduct risk assessments that focus on harms to women and girls

3.14 Existing evidence suggests that women and girls' experiences and gendered harms can get broadly overlooked and culturally diminished in organisations.[95] For example, gendered threats like intimate partner violence have often been missed in industry and research threat modelling processes which look at how a system's security might be compromised.[96] Instead, security evaluations can often assume that the main source of threat is an external or unknown stranger. As a result, there are insufficient safety features or processes to respond when someone is being threatened by an intimate partner.

---

[93] Peters, J., 2020. Sexual Content and Social Media Moderation, *Washburn Law Journal*, 59 (3). [accessed 22 October 2024]; Institute of Strategic Dialogue (Matlach, P-D. and Small, A.C.), 2024. Off-limits: Sexual Violence on TikTok. [accessed 07 February 2025].

[94] Hawkins, I., Roden. J., Attal, M. and Aqel, H., 2023. Race and gender intertwined: why intersecting identities. matter for perceptions of incivility and content moderation on social media, *Journal of Communication*, 73 (6). [accessed 22 October 2024].

[95] Criado-Perez, C., 2019. Invisible Women. [accessed 22 October 2024].

[96] Slupska, J. and Tanczer, L., 2021. Threat Modeling Intimate Partner Violence: Tech Abuse as a Cybersecurity Challenge in the Internet of Things in *The Emerald International Handbook of Technology-Facilitated Violence and Abuse.* [accessed 22 October 2024].

3.15 This is why there is a need for gender-sensitive risk assessments, which are achieved by making sure that existing risk assessment processes consider gender (as set out in our Illegal Content Risk Assessment Guidance). Services can further build on these existing risk assessments to capture the particular dynamics of gender-based harms.[97]

3.16 Service providers need to build an understanding of factors that enable and promote online gender-based harms in their online safety risk assessments in order to design safer systems and processes. As technologies evolve rapidly, providers will also need to keep track of emerging risks and trends in perpetration, particularly when rolling out new measures and features.

Foundational steps: What are the expectations for service providers?

3.17 Our Codes and risk assessment guidance set out the following steps for service providers: [98]

a) **Risk assessment**: Service providers have a duty to conduct a suitable and sufficient illegal content risk assessment,[99] and providers of services that are likely to be accessed by children are also required to carry out a suitable and sufficient children's risk assessment.[100] In cases where service providers have not previously completed suitable and sufficient risk assessments, they will now need to conduct these. Some providers will have additional assessment requirements under the forthcoming user empowerment duties. We suggest four steps for a suitable and sufficient risk assessment:[101]

i) **Understand the harms**: Service providers must assess the level of risk for each kind of priority illegal content and other illegal content, as well as for content harmful to children. This includes content and activity which disproportionately affects women and girls, such as intimate image abuse and children's exposure to abuse and hateful content against women and girls.[102]

ii) **Assess the risk of harm:** Results of product testing are an example of an enhanced input service providers can use to assess the risk of harm. Service providers can also consider data or information from their reporting and complaints procedures to assess the prevalence of particular kinds of harms, and the service's track record in handling such complaints. Ofcom has published information to help service providers identify what is illegal in the Illegal Content Judgements Guidance and

[97] Enhanced processes for product testing, such as abusability and red teaming, can be particularly relevant in highlighting risks that disproportionately affect women and girls. These will be explored further in Chapter 4, with Chapter 3 focussing on overarching risk assessment and threat modelling practices that should take place before new products are developed.

[98] The 'foundational steps' refer to a range of expectations we have already set out for service providers, either in draft for the purpose of consultation or in final form at set out in a decision at the time of publication of this Guidance. Where the measures which are currently in draft are changed in our final statement following the consultation on those measures, we will revise this Guidance accordingly. The 'foundational steps' apply to different services based on functionality, risk, and size. Information about which foundational step applies to which service is set out in the Guidance at a Glance document.

[99] Section 9 (user-to-user) and section 26 (search) of the Act

[100] Section 11 (user-to-user) and section 28 (search) of the Act.

[101] For more information, see our Illegal Content Risk Assessment Guidance, and our draft Children's Risk Assessment Guidance. [accessed 12 February 2025].

[102] This also includes the perpetration of online harassment, domestic abuse, stalking as well as the circulation of content that normalises or promotes misogyny and gender-based violence.

what is harmful to children in the draft Guidance on Content Harmful to Children, respectively, and how these harms manifest in the Illegal Harms and draft Children's Registers of Risk.

iii) **Decide measures, implement and record**: One way for service providers to meet their duties will be to apply the relevant safety measures set out in Ofcom's Codes of Practice.[103] Services can, however, decide on the appropriate online safety measures for their service to mitigate risk of harm to users. Chapters 3-5 of this Guidance lay out actions providers can consider as a part of this process.

iv) **Report, review and update risk assessments**: we recommend that service providers report their risk assessment outcomes and online safety measures to a relevant internal governance body.

b) **Internal content and search moderation policies**: These policies should be set having regard to the findings of the risk assessment and any evidence of emerging harms on the service.[104]

3.18 **Case study 4** includes two examples of how service providers could account for the risks of harm to different users, considering the overlap of harms and the influence of age on risk, in line with our risk assessment guidance.[105]

Case study 4: Gender-sensitive risk assessments

- When assessing their risk level for different kinds of harm, service providers need to consider risk factors, including functionalities, business model and user base.
- As a starting point, service providers may assess their user base demographics to understand which harms disproportionately impact women and girls, particularly those with intersecting identities.
- For example, young women (aged 18-24) are particularly at risk: one in five women in the UK have suffered online abuse or harassment, increasing to one in three for young women aged 18 to 24.[106] Young women are particularly at risk of image-based sexual abuse including cyberflashing and intimate image abuse.[107] In addition, while they are at an especially high risk of online gender-based harm, they fall into a protection gap as many safety measures are aimed at children under 18.
- Service providers often collect demographic data about users, for example for advertising purposes or to improve users' experiences.[108] Sometimes services can also make inferences about demographic data on the basis of user behaviour, for example inferring age or gender from the kinds of videos users watch.[109] Providers can also use

[103] The Draft Illegal Content Codes of Practice for user-to-user services and Draft Illegal Content Codes of Practice for search services were published in final form on 16 December 2024 and are undergoing the parliamentary process. The Draft Children's Online Safety Codes were published for consultation in May 2024 and will be finalised in April 2025. [accessed 13 February 2025].
[104] Illegal Content (ICU C3 / ICS C2), Draft Protection of Children (PCU B2 / PCS B3).
[105] Illegal Content Risk Assessment Guidance and our draft Children's Risk Assessment Guidance. [accessed 13 February 2025].
[106] Amnesty International UK, 2017. Black and Asian Women MPs Abused More Online. [accessed 6 February 2025].
[107] Illegal Harms Register of Risks 6M and 6S.
[108] Federal Trade Commission, 2024. A Look Behind the Screens: Examining the Data Practices of Social Media and Video Streaming Services. [accessed 9 December 2024].
[109] Federal Trade Commission, 2024. A Look Behind the Screens: Examining the Data Practices of Social Media and Video Streaming Services. [accessed 9 December 2024].

such user data at an aggregate level to consider risks to women and girls on their platform (such as the proportion of their user base that could be at risk). When considering the use of personal information, providers must also consider privacy rights and comply with duties under the UK General Data Protection Regulation ('UK GDPR').[110]

- In addition, services can assess risk with a specific focus on women and girls. Taking the example of online harassment, services can understand how:

> User base demographics can show that online harassment disproportionately affects women and girls, and in particular women in public life, as well as women and girls with multiple protected characteristics; [111]

> The functionalities of the service and business model – such as reposts or trending hashtags which amplify virality of hateful content – can contribute to harassment;[112]

> Online harassment is often sexualised and includes elements of body-shaming and fetishisation[113] and can manifest in ways that overlap with other harms, including offline stalking or threats.[114]

Good practice steps: How can service providers go further?

3.19 Providers can gain additional insights into how design choices create risk for women and girls by seeking expert advice and hearing directly from users about their experiences. This can also help them understand what further mitigations may be effective on their services. This could be done by:[115]

a) **Using external assessors** for monitoring the threat landscape, including local partners with regional and cultural knowledge, and international partners with expertise in highly contextual risk areas such as cyberstalking and controlling or coercive behaviour (see **Case study 3**).[116]

---

[110] We encourage providers to consult the Information Commissioner's Office's (ICO) guidance on UK GDPR requirements and the Age-Appropriate Design Code when processing the personal information of children. [accessed 13 February 2025].

[111] Amnesty International, 2017. Social media can be a dangerous place for UK women. [accessed 8 January 2025]; Glitch, 2023. The Digital Misogynoir Report: Online abuse against Black women allowed and enabled to thrive. [accessed 22 October 2024]; Illegal Harms Register of Risks 6E.

[112] Gosse, C., Veletsianos, G., Hodson, J., Houlden, S., Dousay, T.A., Lowenthal, P.R. and Hall, N., 2020. The hidden costs of connectivity: nature and effect is of scholars' online harassment, *Learning, Media and Technology*, 46 (3). [accessed 8 January 2025]; Thompson, J. D. and Cover, R., 2021. Digital hostility, internet pile-ons and shaming: A case study. *Convergence: The International Journal of Research into New Media Technologies*, 28 (6). [accessed 8 January 2025]; Marwick, A. and Boyd, D., 2010. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience, New Media Society 13 (1). [accessed 8 January 2025].

[113] Glitch, 2023. The Digital Misogynoir Report: Online abuse against Black women allowed and enabled to thrive. [accessed 22 October 2024].

[114] Illegal Harms Register of Risks. [accessed 22 October 2024].

[115] The following could all be included as "enhanced inputs" in the risk assessment process laid out in the Illegal Content Risk Assessment Guidance and draft Children's Risk Assessment Guidance.

[116] Carnegie UK, The End Violence Against Women Coalition, Glitch, NSPCC, Refuge, 5Rights, Woods, L. and McGlynn, C., 2022. Violence Against Women and Girls (VAWG) Code of Practice. [accessed 22 October 2024]; Glitch, 2023. The Digital Misogynoir Report: Online abuse against Black women allowed and enabled to thrive. [accessed 22 October 2024].

b) **Engaging with survivors and victims** to better understand their experiences. However, this type of engagement can put a burden on survivors and victims, as well as organisations that support them.[117] Engaging with organisations that represent people with lived experience can be valuable, particularly when these organisations include individuals with lived experience as well. As these organisations are often under-resourced, service providers could consider appropriate compensation for any work dedicated to improving provider's policies or risk assessments.[118]

c) **Conducting user research** such as surveys, to better understand users' preferences and experiences of risk (see **Case study 5**).[119]

d) **Conducting an impact assessment** alongside other risk assessments to assess impacts on self-expression, freedom from discrimination and privacy, especially for those with protected characteristics.[120]

Case study 5: Trauma-informed user surveys

The nature of online risk changes rapidly, as perpetrators of abuse identify new ways to co-opt or subvert new technologies to coerce and harass their targets. Conducting user research such as surveys can be valuable to identify these developments and respond to them.

For example, users of dating platforms are particularly at risk of abuse, such as stalking, harassment and coercive control in the context of dating. Chayn, an organisation that supports survivors and victims of domestic abuse, has partnered with various online dating platforms to help them better understand and support the risks their users experience. Chayn helps companies apply principles of trauma-informed design in their user surveys, with a focus on informed consent and privacy, sharing context on how data will be used, and working with a localisation team who understand trauma for multi-language surveys.[121] Such surveys can lead service providers to, for example, develop safety tools to prevent cyberflashing on their service (please see Chapter 4 for further details on preventative measures).

The survey also signposted to relevant resources, including Bloom, a remote trauma service offered by Chayn to online dating users who report harassment, assault or abuse. The service includes courses on healing from sexual trauma as well as access to one-to-one chat support and up to six sessions with a trauma-informed therapist.

Action 3: Be transparent about women and girls' online safety

3.20 Transparency reporting is an important source of online safety information for users, which will shine a light on service providers' safety performance and empower users to make

[117] Ofcom Stakeholder Workshop 2 on Women and Girls Online Safety, 19 November 2024.
[118] Women's Aid, 2024. Domestic abuse services struggling to fill critical gaps in a challenging landscape, exacerbated by the rising cost-of-living. [accessed 10 November 2024].
[119] Service providers should have regard to the needs of their UK user base in considering what languages are needed to ensure their user surveys are accessible.
[120] Equality and Human Rights Commission, 2019. Human Rights and Business. [accessed 16 December 2024]; United Nations Human Rights Office of the High Commissioner, 2011. UN Guiding Principles on Business and Human Rights. [accessed 16 December 2024].
[121] Winfield, J., 2024. How can we make quantitative research more trauma-informed?, *Medium,* 31 July. [accessed 10 October 2024].

informed choices about the services they use. This could be particularly relevant for women and girls who face disproportionate risk online and therefore have to curate their experiences to keep themselves safe online.

3.21 As laid out in our Online Safety Transparency consultation, Ofcom will issue transparency notices to categorised services once a year. In some instances, we may ask for information regularly to look at trends over time. In other instances, we may request information on a specific issue related to, for example, a user group or topical event. This could include shedding light on specific issues related to online gender-based harms.

3.22 We expect transparency reporting to help lead providers of categorised services to take measures to reduce harms stemming from their activities. Providers of categorised services may respond to what they perceive to be legal, commercial, or reputational risks generated by the public disclosure of information about their operations and impact.

3.23 We also expect to use transparency reporting to equip stakeholders, including Ofcom and civil society, with new knowledge about the safety practices of providers of categorised services and their effectiveness. This will enable stakeholders to understand best practice and encourage evidence-based safety improvements at services.

3.24 Publishing data on how providers of categorised services enforce their policies and the effectiveness of safety features or innovations allows for clear assessment of what works. Interventions improving user safety or deterring online gender-based harms should be shared and adopted more widely.[122]

Foundational steps: What are the expectations for service providers?

3.25 Under the Act, a proportion of in-scope services (which we call "categorised services") will be required to comply with additional duties focused on ensuring they are transparent and accountable.[123] The relevant service providers will be required to publish transparency reports based on requirements laid out in transparency notices issued by Ofcom.[124]

Good practice steps: How can service providers go further

3.26 Although transparency reporting applies to categorised services, non-categorised services could also engage in transparency reporting to improve accountability and better enable informed decisions. In addition, all services could further develop their approach to transparency to disclose how they address online gender-based harms. This can enable providers to demonstrate to users the actions they take and their impact on women and girls' online safety. The good practice steps service providers could take include:

a) **Sharing information** about the prevalence of different forms of online gender-based harms and the effectiveness of measures in place to address them, where that data is already available to providers and is collected and managed in line with data protection regulations. This information could include data on user reports and their outcomes, as

[122] eSafety Commissioner, 2024. Technology, gendered violence and Safety by Design. [accessed 22 October 2024].
[123] See Section 95 of the Act. Ofcom will publish a register of these categorised services and keep it up to date. Categorised services will need to comply with a series of additional duties, including transparency reporting.
[124] See Section 95 of the Act.

well as gender-disaggregated and race-disaggregated data on reports and outcomes. [125] It could also reveal additional information on which harms disproportionately affect women and girls, as well as indicate any biases in how reports are dealt with. [126]

b) **Providing more detail** about which posts are flagged by automated content moderation, active bystanders who are not targeted by abuse but report content to support others, and the targeted users themselves. [127]

c) **Exercise caution** in sharing information that perpetrators could exploit to circumvent safety measures, as well as details of specific incidents that could identify an individual or group, including location, sexual orientation, religion or other sensitive information that could put them at risk. [128] For example, this could mean only sharing identity-related data at a population level (such as what percentage of reports were made by men vs women), and not disclosing details of a specific report that could identify people involved. When sharing involves personal data, services will need to comply with the requirements of data protection law.

[125] Glitch, 2023. The Digital Misogynoir Report: Online abuse against Black women allowed and enabled to thrive. [accessed 22 October 2024]; International Center for Journalists (Posetti, J. and Shabbir, N), 2022. The Chilling: A global study of online violence against women journalists. [accessed 22 October 2024].

[126] Appelman, N., 2023. Disparate Content Moderation Mapping Social Justice Organisations Perspectives on Unequal Content Moderation Harms and the EU Platform Policy Debate. [accessed 22 October 2024].

[127] Glitch, 2023. The Digital Misogynoir Report: Online abuse against Black women allowed and enabled to thrive. [accessed 22 October 2024].

[128] eSafety Commissioner, 2024. Technology, gendered violence and Safety by Design. [accessed 22 October 2024].

4. Preventing harm

Warning: this chapter contains content that may be upsetting or distressing.

Overview

Context

4.1 Harm prevention means taking action to minimise the risk of harm before it occurs. To prevent harm, service providers should focus on deterring and interrupting perpetrator behaviour, including identifying potential routes for abuse and allowing for changes in features and functionalities. This draws on safety-by-design concepts that anticipate how features and functionalities can be co-opted to harm women and girls online and changing them to make it harder for perpetrators to misuse them.[129]

4.2 Online gender-based harms are often addressed retrospectively through interventions after the harm has occurred, which largely rely on women and girls to act, for example by reporting abuse to the service provider or the police. This requires time and effort from survivors who are already navigating the psychological, emotional, and reputational effects of online gender-based harms.[130]

4.3 This chapter looks instead at the actionable ways providers can prevent online gender-based harms before they happen on their services, including by discouraging perpetrators from doing harm in the first place. We specifically look at three actions related to service design and harm prevention:

- Action 4: Conduct abusability evaluations and product testing.
- Action 5: Set safer defaults.
- Action 6: Reduce the circulation of content depicting, promoting or encouraging online gender-based harm.

4.4 For each principle, we set out a solid baseline of what safety looks like (‘**foundational steps’**) to help service providers meet their new duties to protect UK users. We also highlight additional **good practice steps** to illustrate how providers can take an ambitious approach to service design and harm prevention.

4.5 These steps require a proactive approach which uses insights and evidence from risk assessments (discussed in the Chapter 3) to make changes to the features and functionalities of the service.

[129] NSPCC response to the November 2023 Illegal Harms Consultation, p.14; OSAN response to the November 2023 Illegal Harms Consultation, p.1; EVAW response to the November 2023 Illegal Harms Consultation, p.3; CARE response to the November 2023 Illegal Harms Consultation, p.9; Domestic Abuse Commissioner response to the November 2023 Illegal Harms Consultation, p.2; Victims Commissioner response to the November 2023 Illegal Harms Consultation, p.2.

[130] Ofcom Stakeholder Workshop 1 on Women and Girls Online Safety, 16 September 2024.

4.6 While it is important to adopt this preventative approach when designing and adopting new technologies, for example GenAI[131] chatbots or virtual/augmented reality, it is also important to extend this approach to relatively everyday[132] digital technologies (social media, messaging, dating apps). These are the sites where the vast majority of online gender-based harms take place.[133]

4.7 As part of the risk assessment processes outlined in the previous chapter, providers may identify features or functionalities which were not made with a safety-by-design approach and therefore may need to be redesigned to be safer. This could include retiring certain features or implementing design changes to mitigate risks.

Our target outcomes

4.8 Addressing the issue of online gender-based harms systematically means this Guidance is not focused on identifying individual instances of harm. Rather, online gender-based harms are structural issues which require active correction within various systems, from technology design to education. This preventative and systemic approach should reduce the burden on women and girls to keep themselves safe by transferring the responsibility to service providers.

4.9 Safety-by-design can make features and functionalities harder to abuse, leading to a reduction in harms on the service.

4.10 Media literacy is also a vital part of preventing online gender-based harms.[134] By equipping people with the skills to critically engage online, we can prevent harm and foster a more resilient digital community in the UK. We include relevant examples of how service providers can promote media literacy as a preventative tool in this chapter.

4.11 This approach would include engaging with men and boys to tackle misogynistic narratives and cultural norms that justify and glorify harmful behaviour.[135] Platforms can support such education through providing supportive or deterrence messaging, and making sure they do not erroneously remove content which challenges misogynistic abuse online (this can be referred to as 'counterspeech').

---

[131] The Act applies to certain types of GenAI content, chatbots and services. See Ofcom's open letter to online service providers which outlines how the UK's Online Safety Act will apply to Generative AI and chatbots. [accessed 13 February 2025].

[132] Mundane and everyday technologies refer to technologies which are used so commonly that they do not generate interest, excitement or attention unless they malfunction or are misused. For more information on mundane technologies, see: Dourish, P., Graham, C., Randall, D. and Rouncefield, M., 2010. Theme issue on social interaction and mundane technologies. *Personal and Ubiquitous Computing*, 14. [accessed 31 October 2024].

[133] Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T. and Dell. N., 2018. A Stalker's Paradise": How intimate partner abusers exploit technology, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. [accessed 31 October 2024]; Refuge, 2022. Marked as Unsafe. [accessed 30 January 2025].

[134] Ofcom Stakeholder Workshop 1 on Women and Girls Online Safety, 16 September 2024.

[135] Burrell, S. 2018. The contradictory possibilities of engaging men and boys in the prevention of men's violence against women in the UK, *Journal of Gender-Based Violence*, 2 (3). [accessed 16 December 2024]; National Education Union, 2023. Working with boys and young men to prevent sexism and sexual harassment. [accessed 16 December 2024].

4.12 Preventative approaches are particularly valuable because they are often non-punitive and focus on deterrence and behaviour change rather than punishment.[136]

4.13 It is also important to tailor these approaches to a variety of different kinds of users who may cause harm, from those that may do so unintentionally or unknowingly, to persistent and highly motivated perpetrators who will try to evade all safety measures.

Action 4: Conduct abusability evaluations and product testing

4.14 Abusability evaluations draw on the concept of 'usability' which is used to evaluate how easy it is for users to navigate a website or device to accomplish their goals. In contrast, abusability evaluations test how easy it is to abuse a tool or feature for harm, and therefore point to ways that abusability can be minimised in design.[137]

4.15 Perpetrators of online gender-based harms can be very innovative in co-opting technologies to facilitate pre-existing patterns of harassment, coercion and control. Many perpetrators – such as those spreading gendered and sexualised misinformation – will tactics to bypass detection and safety measures meant to prevent abuse, for example, through intentional misspellings of slurs and insults.[138]

4.16 Services can use techniques that seek to anticipate these evolving forms of abuse (as outlined in Chapter 3). One way to do so is adapting models from cybersecurity such as 'red team' exercises in which some testers take on the role of the perpetrator to try to test safety measures. This section outlines product testing methods called abusability evaluations or 'red teaming' which service providers can use to anticipate and prevent abuse.

4.17 Applying the concepts of abusability or red teaming during product testing can be useful across all harm categories. Those conducting the tests should be familiar with the specific nuances and dynamics of online gender-based harms. Therefore, it is particularly valuable to partner with subject matter experts who have experience of supporting survivors and victims when conducting such exercises.

Foundational steps: What are the expectations for service providers?

4.18 Our Codes and risk assessment guidance set out the following steps for service providers:[139]

---

[136] Ofcom Stakeholder Workshop 1 on Women and Girls Online Safety, 16 September 2024.

[137] Beers, A., Nguyên, S., Sioson, M., Mayanja, M., Ionescu, M., Spiro, E. S., and Starbird, K., 2021. The Firestarting Troll, and Designing for Abusability. [accessed 31 October 2024]; Ofcom Stakeholder Workshop 1 on Women and Girls Online Safety, 16 September 2024.

[138] Wilson Center (Jankowicz, N., Hunchak, J., Pavliuc, A., Davies, C., Pierson, S. and Kaufmann, Z.), 2021. Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online. [accessed 31 October 2024].

[139] The 'foundational steps' refer to a range of expectations we have already set out for service providers, either in draft for the purpose of consultation or in final form at set out in a decision at the time of publication of this Guidance. Where the measures which are currently in draft are changed in our final statement following the consultation on those measures, we will revise this Guidance accordingly. The 'foundational steps' apply to different services based on functionality, risk, and size. Information about which foundational step applies to which service is set out in the Guidance at a Glance document.

a) **Product testing**: As a part of suitable and sufficient risk assessments (as outlined in Chapter 3), we have set out product testing as one of the types of evidence service providers could use as an input to improve the accuracy of their judgments on risk.[140]
b) **Significant change risk assessment**: Service providers must carry out a new risk assessment before making a significant change to their service.[141]
c) **Recommender system testing:** Service providers should, when carrying out existing on-platform testing of content recommender systems, collect additional safety metrics when making design adjustments, to evaluate whether the adjustment is likely to increase user exposure to illegal content.[142]

4.19 **Case study 6** includes further details on how service providers can conduct one kind of product testing, called an abusability evaluation.

Case study 6: Abusability product testing

- Abusability testing is one kind of product testing that can inform online safety risk assessments[143] by internally testing a product to see if or how it can be abused before deploying it. This testing can also be applied to products which have already been deployed, as not all incidents of abuse will be reported by users and therefore may go unnoticed.
- In the context of online gender-based harms, abusability testing involves understanding and mapping out which features and functionalities are likely to be misused for harms such as domestic abuse, harassment, intimate image abuse, and stalking.[144]
- Sometimes, these features could also be low risk for other users, or have high levels of utility for some users, creating difficult trade-offs. For example, while some users may abuse location-tracking for stalking and surveillance, others may benefit from being able to monitor their children or relatives.[145]
- In some cases, removing such a feature could reduce the use of the application for adversaries without substantially inconveniencing legitimate users. In other cases, the service provider may judge that the feature offers benefits to the majority of its users, but to manage the risk to some users, it could provide better information or customisable defaults.[146]

---

[140] Illegal Content Risk Assessment Guidance and draft Children's Risk Assessment Guidance on "core evidence" and "enhanced evidence" for risk assessments. [accessed 13 February 2025].
[141] Illegal Content Risk Assessment Guidance and draft Children's Risk Assessment Guidance. [accessed 13 February 2025].
[142] Illegal Content Codes of Practice (ICU E1). This measure applies only to user-to-user services that identify as medium or high risk for at least two specified harms.
[143] This refers to both illegal content risk assessment and draft children's risk assessment.
[144] Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T. and Dell. N., 2018. A Stalker's Paradise": How intimate partner abusers exploit technology, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. [accessed 31 October 2024]; Strohmayer, A., Slupska, J., Bellini, R., Neff, G., Coventry, L., Hairson, A. and Dodge, A., 2021. Trust and Abusability Toolkit: Centering Safety in Human-Data Interactions. [accessed 30 October 2024].
[145] Levy, K. E., 2015. Intimate surveillance, *Idaho Law Review* 51 (3). [accessed 17 December 2024].
[146] Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T. and Dell, N., 2018. **"**A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology, *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* [accessed 28 October 2024]. Parkin, S., Patel, T., Lopez-Neira, I., and Tanczer, L., 2020. Usability analysis of shared device ecosystem security, NSPW'19: *Proceedings of the new security paradigms workshop*. [accessed 30 October 2024]; Slupska, J. and Tanczer, L., 2021. Threat Modeling Intimate Partner Violence: Tech Abuse as a Cybersecurity Challenge in the Internet of Things in *The Emerald International Handbook of Technology-Facilitated Violence and Abuse.* [accessed 22 October 2024].

- Some features are commonly co-opted for specific forms of online harassment. For example, a feature which allows creating and sharing lists of other users can be misused to share lists of targets with specific characteristics for pile-ons and coordinated harassment. Design changes, like notifying users if they have been added to a list, seeking their permission before they are added, or allowing users to remove themselves from these lists, can mitigate the risk of harm from such features.

Good practice steps: How can service providers go further?

4.20 Providers can implement additional good practice for abusability evaluations and product testing, both before and after product deployment, to gain further insights and data on the potential misuses of their products. This will not only create safer environments for women, girls and other users at heightened risks of harm, but in pre-empting and limiting misuse, providers can reduce the resource and reputational risks from abuse of their service by perpetrators. This can include:

a) **Using red teaming** for abusability testing, in which a product testing team takes on the role of a malicious actor and tries to attack a system in order to find vulnerabilities (as mentioned previously and in **Case Study 7**).[147] This could be repeated periodically, even if there are no major developments, as perpetrators will adapt quickly to evade safety measures.
b) **Working with experts** with direct or relevant experience engaging with and understanding perpetrator behaviours.
c) **Using personas** to explore how different users may experience a feature and including intersectional perspectives.[148]
d) **Media literacy:** Adhering to the principles on monitoring and evaluating features in the Best Practice Design Principles for Media Literacy.

4.21 These tests are only effective if service providers take action to reduce the risks identified as a result. **Case Study 7** shows how a red teaming exercise can lead to changes.

Case study 7: Red teaming for non-consensual intimate image abuse deepfakes

The proliferation of audio-visual GenAI tools has facilitated the rise of non-consensual intimate image (NCII) abuse deepfakes, leaving a devastating impact on the lives of survivors and victims, most of whom are women and girls.[149] While many services with GenAI functionalities employ safeguards to prevent the generation of deepfake intimate images (such as safety filters), research shows that users can successfully break guardrails leveraging

[147] Ofcom, 2024. Red Teaming for GenAI Harms - Revealing the Risks and Rewards for Online Safety.[accessed 30 October 2024].
[148] World Wide Web Foundation, 2021. Tech Policy Design Lab: Online Gender-Based Violence and Abuse. [accessed 30 October 2024].
[149] Security Hero, 2023. 2023 State Of Deepfakes: Realities, Threats, And Impact. [accessed 30 October 2024]. Malicious actors are using these tools to generate deepfake IIA content which is then posted/ shared on user-to-user services or surfaced in search results. Some of the deepfake IIA content created originate from the same models that user-to-user services and search services are adopting. Source: Ofcom, 2024. Deepfake Defences: Mitigating the Harms of Deceptive Deepfakes. [accessed 14 February 2024]; My Image My Choice, 2024. Deepfake Abuse: Landscape Analysis 2023-24. [accessed 14 February 2024].

basic prompting techniques.[150] For example, malicious users have circumvented online tools to generate deepfake intimate images of women in public life.[151]

Ongoing red teaming can help service providers make their GenAI tools more robust against such attacks.[152] Red teaming is a type of model evaluation that seeks to find and fix vulnerabilities in GenAI models. Service providers have used red teaming to understand whether their model can produce explicit or harmful material. In our discussion paper on red teaming, we considered good practices for a red team exercise, which could involve a service provider testing the effectiveness of safety measures intended to prevent AI-generated intimate images of women in public life.[153] It could involve a service provider testing the effectiveness of safety measures intended to prevent the generation of sexual content. The provider could then test the model's ability to generate images of public figures. If both tests are successful, it could in theory indicate the likelihood of the model being used to generate deepfake intimate images of public figures.[154] In instances where such vulnerabilities are identified, the service provider should take steps to strengthen its existing safety measures. This can include improving its input and output filters (such as content filters), updating blocklists for specific public figures, and removing nudity content from training datasets.[155]

Action 5: Set safer defaults

4.22 A common outcome of abusability evaluations will be to set safer defaults. Default settings can be a powerful tool to encourage safer behaviour online. Our research shows that defaults (such as a pre-selected choice) strongly influence user choice even when changing the setting takes one click.[156]

4.23 Safer defaults can also embed better consent practice in service design by allowing for more points at which users can determine whether they want to participate in an interaction, or allow their data to be shared.

4.24 Taking steps to make the service less susceptible to abuse by default makes it easier for women and girls to keep themselves safe online.

4.25 These safety measures often cut across different types of harm, although they are particularly important for harms like stalking.

---

[150] Safety filters of a text to image model can be bypassed to produce sexual content using a few prompts. Source: Yang, Y., Hui, B., Yuan, H., Gong, N. and Cao, Y., 2024. SneakyPrompt: Jailbreaking text-to-image generative models, *2024 IEEE symposium on security and privacy*. [accessed 30 October 2024].

[151] Ofcom, 2024. Deepfake Defences: Mitigating the Harms of Deceptive Deepfakes. [accessed 14 February 2024].

[152] We are already seeing red teaming being implemented to reduce the likelihood of GenAI services being misused to produce deepfake intimate images.

[153] The red team could for example, develop subtle variations of prompts to mimic how bad actors might be able to circumvent the filter (e.g. by misspelling words or using coded language known to perpetrators).

[154] There could also be methods to directly red team models for deepfake intimate images of public figures as well as private individuals. Companies should assess the merits and limitations of such approaches, including the legal and ethical risks involved. They should seek legal counsel where possible before conducting such exercises.

[155] For more information on what red teaming is and best practices when deploying the methodology see our discussion paper. [accessed 13 February 2025].

[156] Ofcom, 2024. Behavioural insights to empower social media users. [accessed 30 October 2024].

Foundational steps: What are the expectations for service providers?

4.26 Our Codes set out the following steps for service providers:[157]

a) **Safe settings**: Defaults for user-to-user services are set to protect child users.[158]

i) Automated location information of child users' accounts should not be visible to any other users by default. Any location sharing functionality should be 'opt in'.
ii) Child users should not be visible in connection lists of other users. The connection lists of child users should also not be visible to other users.
iii) Child users are not presented with prompts to expand their network of friends, or be included in network expansion prompts presented to other users.
iv) Non-connected accounts do not have the ability to send direct messages to children using a service.
v) For services with no formal connection features, providers should implement mechanisms to ensure children using a service can actively confirm whether to receive a direct message sent from another user account before it is visible to them.

b) **Group chats**: Provide children with an option to accept or decline an invite to a group chat.[159]

c) **Supportive information** is provided to children using a service in a timely and accessible manner.[160] This is to help child users make informed choices about risk by giving them information, access to safeguarding processes, and support on a service, when they are:

i) seeking to disable one of the safer defaults recommended previously which are set to reduce risk;
ii) responding to a request from another user to establish a formal connection; and
iii) receiving a direct message from another user for the first time.

d) **Safe search**: Providers of large general search services should filter out identified pornography, eating disorder, suicide and self-harm content for any users believed to be children.[161]

4.27 **Case study 8** includes further details on how services can set safer defaults.

Case study 8: Preventing grooming through safer defaults

- Online grooming involves establishing and developing communication with children for the purpose of conducting child abuse. Perpetrators often use bribery, blackmail or coercion during grooming.[162] The majority of children targeted by grooming are girls, in

[157] The 'foundational steps' refer to a range of expectations we have already set out for service providers, either in draft for the purpose of consultation or in final form at set out in a decision at the time of publication of this Guidance. Where the measures which are currently in draft are changed in our final statement following the consultation on those measures, we will revise this Guidance accordingly. The 'foundational steps' apply to different services based on functionality, risk, and size. Information about which foundational step applies to which service is set out in the Guidance at a Glance document.
[158] Illegal Content (ICU F1).
[159] Draft Protection of Children (PCU G4).
[160] Illegal Content (ICU F2).
[161] Draft Protection of Children (PCS B2). For this draft measure, we propose that where providers choose to filter out such content via a 'safe search default',' we specifically add that providers should take steps to ensure that this safe search setting cannot be switched off by users believed to be children.
[162] For more information on grooming, please see Ofcom Illegal Harms Register of Risks: Grooming.

part because girls are often seen by perpetrators as being more vulnerable to being targeted.[163]

- Perpetrators often target services that encourage new connections, such as social media or gaming services, particularly those with many child users. After establishing communication, perpetrators often attempt to move communication to another service, particularly private messaging services which allow perpetrators access to image-sharing.[164]
- Default safety options to the highest protection level can help prevent online grooming, such as disabling the option for unknown adults to contact children (messaging) and hiding information about them (geolocation).

Good practice steps: How can service providers go further?

4.28 Safer defaults can be a powerful safety measure for all users, not only children. By implementing stronger default settings, providers can demonstrate their commitment to safety by ensuring women and girls have a safe and accessible online experience from their first moment on the service.

4.29 As adult users are ultimately more capable of managing their online safety than children, these strong default settings should be paired with options for control and customisation (this means that in some cases the highest safety option should be pre-selected by default, but a user can change or customise it if they want. We discuss user control in further detail in Chapter 5). This could include:

a) **Interaction defaults:** Setting strong and customisable defaults around interaction, such as who can contact a user or asking a user's permission before being added to a group chat.
b) **Privacy defaults:** Setting strong and customisable defaults around privacy, such as what information users can see about another user and their content and who can and cannot redistribute their content or username/identity in real time. Other settings could include providing users with tools to express their gender identity and giving them control over the company's collection and inference of user information related to their sexual orientation and gender identity.[165]
c) **Bundles:** Combining relevant safety and privacy settings into 'bundles'. Users can be overwhelmed by too many options for settings, including when creating a profile, meaning they disengage from making decisions. Research suggests that choice bundles that include customisation options can help users make informed choices about their online settings.[166] Grouping relevant choices together into a 'bundle' can reduce the time and effort required from users. On the other hand, offering manual customisation options could allow users to make more granular manual choices outside of the bundles.

---

[163] Girls may be twice as likely to be groomed. Source: Whittle, H. C., Hamilton, Giachritsis, C., Beech, A. and Collings, G., 2013. A Review of young people's vulnerabilities to online grooming, *Aggression and Violent Behavior*, 18 (1). [accessed 29 October 2024].
[164] Ringenberg, T.R., Seigfried-Spellar, K.C., Rayz, J. M. and Rogers, M.K., 2022. A scoping review of child grooming strategies: Pre-and post-internet, *Child Abuse & Neglect*, 123, [accessed 29 October 2024].
[165] GLAAD, 2024. GLAAD Social Media Safety Index Platform Scorecard. [accessed 29 October 2024]; Department for Science, Innovation & Technology and PUBLIC, 2025. Platform Design and the Risk of Online Violence Against Women and Girls. [accessed 12 January 2025].
[166] Behavioural Insights Team, 2021. Active Online Choices: Designing to Empower Users. [accessed 29 October 2024].

d) **Strengthening account security** with Two Factor Authentication ('2FA') or multi-factor authentication feature, while ensuring accounts can be recovered after being hacked and/or locked out.

e) **Account access:** Providing information about account access by making it clear which users are currently connected to an account, device or platform, as well as what unique devices (via IP/MAC addresses) are connected to an account.[167] This minimises opportunities for non-consensual monitoring and surveillance.

f) **Reminders:** Identifying optimal frequency and timing and giving users regular reminders for reviewing or updating privacy and security settings. Research finds that timing of prompts matters for user engagement, and overly frequent reminders can be annoying and lead to disengagement. [168]

4.30 **Case study 9** describes how service providers can set safer defaults related to location data for all users.

Case Study 9: Removing geolocation information by default

Providers often collect and share information about users' locations. For example, smartphone cameras use information from mobile data, Wi-Fi, GPS networks and Bluetooth and embed this as location metadata in photo and video files. Many users are not aware that when they share photos and videos that include location metadata on social media and messaging services, they can inadvertently reveal users' locations. Likewise, many providers will collect and share users' locations to enhance social networking, which can lead to unintended consequences.[169] Such information leaking can cause serious harms, up to and including homicide, in cases of coercive and controlling behaviour and stalking.[170]

To prevent such harms, some providers limit opportunities for location sharing, ensure geolocation options are off by default, and provide obvious signs and warnings for users when location tracking is active.[171] For example, services can ensure that metadata is removed from all images upon upload. Policies which ensure privacy settings by default can also be effective in protecting users from online gender-based harms.

Action 6: Reduce the circulation of content depicting, promoting or encouraging online gender-based harms

4.31 Service providers can use persuasion (supportive or deterrence messaging), removal (preventing uploads or taking down content), or reduction (limiting circulation or reducing

[167] Service providers should have regard to the needs of their UK user base in considering what languages are needed to ensure the information about account access is accessible.

[168] Wohllebe, A., Hübner, D., Radtke, U. and Podruzsik, S. 2021. Mobile apps in retail: Effect of push notification frequency on app user behavior. *Innovative Marketing* 17 (2). [accessed 30 January 2025].

[169] Dhondt, K., Le Pochat, V., Voulimeneas, A., Joosen, W. and Volckaert, S., 2022. A Run a Day Won't Keep the Hacker Away: Inference Attacks on Endpoint Privacy Zones in Fitness Tracking Social Networks, *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*. [accessed 29 October 2024].

[170] Baddam, B., 2018. Technology and its Danger to Domestic Violence Victims: How Did He Find Me?. *Albany Law Journal of Science & Technology*, 28 (1). [accessed 29 October 2024].

[171] Ensuring geolocation options are off by default aligns with standard 10 'Geolocation' of the ICO's Age Appropriate Design Code, which helps providers of online services that are likely to be accessed by children take necessary steps to protect their personal data. [accessed 05 February 2025].

visibility) to reduce the circulation of harmful content or messages. Many of these methods will rely on automated detection techniques.

4.32 It is up to services to decide which methods will be most appropriate in each case. However, we expect that in most cases, a mix of approaches will be most effective. Each of these methods can be a powerful tool to reduce the spread of content which normalises abuse and misogyny and drives forms of online gender-based harms.[172] It can reduce the 'safety work' burden on those having to report many individual pieces of content. It can also limit the second-hand impacts of witnessing abuse against women in public life. [173]

4.33 User-to-user service providers should implement proportionate systems to prevent uploads of illegal content, as well as remove it swiftly when they become aware of it.[174] Likewise, search providers should put in place search moderation systems that allow them to moderate illegal content.[175] For legal content, in some cases providers may also seek to limit the circulation of such content through persuasion, removal and reduction. This could be because it is harmful to children, and therefore providers need to take steps to protect children on their service from encountering it.[176] It could also be because such content violates a provider's Terms of Service.[177]

4.34 These methods work differently and may be more appropriate for certain types of content and on certain services. Some may use proactive technology, which Ofcom is only able to recommend in Codes of Practice for content communicated publicly.

4.35 Methods based on persuasion are more respectful of users' autonomy and may be more beneficial to educate users about respectful behaviour. However, they are less likely to be effective against highly motivated perpetrators. Methods of removal and reduction are more effective at reducing the impact of harmful content, but they raise a variety of concerns related to a service provider's control over users' expression.

4.36 In particular, reduction is often used by service providers for content that is deemed by a service provider to be misleading, offensive, or otherwise risky, but not necessarily illegal. While reduction actions interfere with users' expression less than content removal, they are less visible to users and therefore they pose issues for accountability and oversight for these decisions.[178]

---

[172] UN Women, 2023. Technology-facilitated violence against women: Taking stock of evidence and data collection. [accessed 31 October 2024].
[173] The Global Partnership, 2022. Technology-Facilitated Gender-Based Violence: Preliminary Landscape Analysis. [accessed 31 October 2024].
[174] Section 10 of the Act
[175] Section 27 of the Act
[176] Duties relating to the protection of children are set out in sections 11-13 and 20-21 of the Act for regulated user-to-user services and sections 28-30 and 31-32 of the Act for regulated search services.
[177] In this chapter, we do not specify what service providers' Terms of Service should regulate, but rather review how they can enforce the policies they set out in their Terms of Service (see Governance and Accountability in Chapter 3).
[178] Gillespie, T., 2022. Do Not Recommend? Reduction as a Form of Content Moderation, *Social Media + Society*, 8 (3). [accessed 31 October 2024].

Foundational steps: What are the expectations for service providers?

4.37 Our Codes and risk assessment guidance set out the following steps for service providers:[179]

a) **Automated content moderation**: Use automated technique known as 'hash matching' to detect and remove image-based Child Sexual Abuse Material (CSAM).[180] **Case study 10** includes further details on how CSAM hash matching can address online gender-based violence and abuse.

b) **Recommender systems**: Ensure that content that is likely to be harmful to children, including abuse on the basis of sex or gender reassignment, as well as content promoting gendered violence, is given a low degree of prominence on children's recommender feeds.[181] Other kinds of harmful content – pornography, eating disorder content, self-harm and suicide content – should not appear on children's recommender feeds at all.[182] **Case study 11** includes further details on preventing children from encountering online misogyny.

c) **Search moderation:** Put in place search moderation systems or processes that are designed to enable service providers to take appropriate moderation in respect of illegal content and content harmful to children of which they are aware.[183] In addition, service providers should take action to ensure that users do not encounter, in or via search results, search content present at or sourced from URLs on a list of URLs previously identified as hosting CSAM.[184] Service providers should also have a clear appeals process for content that wrongly no longer appears in search results or has been given a lower priority in search results in compliance with the duties in the Act. **Case Study 12** includes further details on preventing children from encountering online misogyny on search services.

d) **CSAM warnings for search:** Include content warnings and support resources for users making search requests where the wording clearly indicates that the user may be seeking to encounter CSAM and which use terms that explicitly relate to CSAM.[185]

e) **Highly effective age assurance**: Some user-to-user service providers should use highly effective age assurance to prevent and/or protect children from encountering content harmful to children which they do not prohibit on the service, and/or to apply the relevant recommender system measures mentioned previously.[186]

---

[179] The 'foundational steps' refer to a range of expectations we have already set out for service providers, either in draft for the purpose of consultation or in final form at set out in a decision at the time of publication of this Guidance. Where the measures which are currently in draft are changed in our final statement following the consultation on those measures, we will revise this Guidance accordingly. The 'foundational steps' apply to different services based on functionality, risk, and size. Information about which foundational step applies to which service is set out in the Guidance at a Glance document.

[180] Illegal Harms (ICU C9)

[181] Protection of Children (PCU F2). For a detailed discussion of content harmful to children, see the Introduction chapter and our draft Children's Register of Risks. [accessed 13 February 2025].

[182] Protection of Children (PCU F1).

[183] Illegal Harms (ICS C1), Protection of Children (PCS B1).

[184] Illegal Harms (ICS C7).

[185] Illegal Harms (ICS F2).

[186] Protection of Children (PCU H2, PCU H3, PCU H4, PCU H5, PCU H6, PCU H7). In our draft Protection of Children Code of practice for user-to-user services, we outline which user-to-user services should use highly effective age assurance to (i) prevent children from accessing the entire service where their principal purpose is to host or disseminate primary priority content or priority content; (ii) to ensure children are prevented from

f) **Signposting children to support** when they search for harmful content, including suicide and eating disorder content.[187] Provide children with crisis prevention information in response to search requests regarding suicide, self-harm and eating disorders.[188]

Case study 10: Hash matching for CSAM

- The circulation of CSAM online is increasing rapidly. Child sexual abuse and the circulation of CSAM online causes significant harm, including to girls, and ongoing circulation of historical imagery can re-traumatise victims and survivors of abuse. The IWF found that 96% of the reports processed in 2022 depicted exclusively girls.[189]
- Hash matching and URL detection can be useful and effective tools for combatting the circulation of CSAM for user-to-user and search services, respectively.
- Hash matching involves analysing images and videos communicated publicly on the service and comparing a digital fingerprint of that content to digital fingerprints of previously identified CSAM. URL detection enables providers to ensure that users do not encounter, in or via search results, search content present at or sourced from URLs on a list of URLs previously identified as hosting CSAM.

Case study 11: Gender-sensitive recommender system algorithms

- A growing community of misogynistic influencers (sometimes referred to as 'misogyny influencers') can have considerable influence over the propagation of misogynistic content. Online misogyny can glorify, justify and create tolerance for sexual violence.[190]
- Evidence shows that recommender systems reward influencers creating misogynistic content with greater reach, particularly to boys and young men.[191] This happens because algorithms are optimised for high engagement, which over time can incentivise the production and exposure to polarising and harmful content.[192]

Furthermore, recommender systems can also disproportionately show other kinds of harmful content to girls and young women, such as content promoting eating disorders and self-harm.[193] Such recommendations can be a result of embedded data bias in data service providers collect about users.[194]

---

encountering primary priority content identified on the service; (iii) to ensure children are protected from encountering priority content identified on the service; and/or (iv) to apply relevant recommender system measures to children. We set out what is 'highly effective age assurance' in our Part 3 Guidance on highly effective age assurance, published in January 2025. We will finalise our age assurance measures for Part 3 services in April 2025 and will update the Part 3 Guidance on highly effective age assurance as appropriate. [accessed 13 February 2025].

[187] Draft Protection of Children (PCU E3).

[188] Illegal Harms (ICS F3) and Protection of Children (PCS E3).

[189] Internet Watch Foundation, 2023. #Behind the Screens: A deep dive into the digital and social emergency happening #BehindTheScreens, in children's bedrooms. [accessed 18 December 2024].

[190] Institute of Strategic Dialogue (Bundtzen, S.), 2023. Misogynistic Pathways to Radicalisation: Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence. [accessed 31 October 2024].

[191] Vodafone, 2024. AI 'Aggro-rithms': young boys are served harmful content within 60 seconds of being online. [accessed 17 December 2024].

[192] Ofcom, 2024, draft Children's Register of Risks. [accessed 13 February 2025].

[193] Ecorys, 2022. Qualitative research project to investigate the impact of online harms on children. [accessed 31 October 2024].

[194] Gerrard, Y. and Thornham H, 2020. Content moderation: Social media's sexist assemblages. *New Media and Society* 22 (7). [accessed 27 January 2025].

- Gendered bias and gendered disinformation can also be shared via GenAI chatbots and voice assistants which can replicate biased algorithms and training data.[195]
- Gender-sensitive approaches can reduce the spread of online misogyny, including abusive and hateful content. Training content recommendation algorithms to be gender-sensitive could include:

> Auditing and evaluating recommender algorithms and other AI systems to assess whether they promote online misogyny, as well as evaluating gender bias in recommendations.[196]

> Retraining algorithms either after revising existing datasets by, for example, applying pre-processing bias-mitigation strategies,[197] or training the algorithm (or 'classifier') on a new training dataset put together by a diverse group that includes humans with a high level of sensitivity and training on gender-based harms, including intersectional aspects. This may also include giving human annotators sufficient time to evaluate content carefully, especially when the evaluation needs to consider the context within which the content was posted.[198].

Case study 12: Gender-sensitive search services

- Some websites and forums are dedicated to allowing users to create non-consensual intimate content, including nudification apps sexualising deepfakes.[199] Providers of general search services can reduce access to these websites, forums, and applications to help protect individuals and society from illegal and harmful non-consensual material.[200] This can include:

> Delisting: action that results in the content no longer appearing in search results.

> Deprioritising: ensuring that a particular piece of content is deprioritised in the overall ranking of search results and is therefore less discoverable to users.

> Reporting: making it easier for people to request removal of non-consensual content from search results.

Good practice steps: How can service providers go further?

4.38 Service providers can reduce the impact of harmful content on women and girls in three ways: persuasion, removal and reduction. Many of these methods will rely on automated

[195] Feine, J., Gnewuch, U., Morana, S. and Maedche, A., 2020. Gender bias in chatbot design, *Chatbot Research and Design*. [accessed 31 October 2024].
[196] Institute of Strategic Dialogue (Bundtzen, S.), 2023. Misogynistic Pathways to Radicalisation: Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence. [accessed 31 October 2024].
[197] Feldman, T. and Peake, A, 2021. End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning. [accessed 20 December 2024]; Park, J. H., Shin, J. and Fung, P., 2018. Reducing Gender Bias in Abusive Language Detection. [accessed 20 December 2024]
[198] Gillespie, T., 2022. Do Not Recommend? Reduction as a Form of Content Moderation, *Social Media + Society*, 8 (3). [accessed 31 October 2024]; Feldman, T. and Peake, A, 2021. End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning. [accessed 20 December 2024]; Excell, E. and Al Moubayed, N., 2021. Towards Equal Gender Representation in the Annotations of Toxic Language Detection, *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. [accessed 20 December 2024].
[199] Henry, N. and Flynn, A., 2019. Image-Based Sexual Abuse: Online Distribution Channels and Illicit Communities of Support, *Violence Against Women*, 25 (16). [accessed 31 October 2024].
[200] Wilson Center (Jankowicz, N., Hunchak, J., Pavliuc, A., Davies, C., Pierson, S. and Kaufmann, Z.), 2021. Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online. [accessed 25 October 2024].

detection techniques. This section is therefore split into four subsections looking at persuasion, removal, reduction, and automated content moderation.

4.39 These **good practice steps** give service providers the opportunity to innovate and gather data on new approaches – some of which we are actively exploring for future Codes of Practice.[201] This could help to strengthen not only providers' response to online gender-based harms, but also set themselves apart as industry leaders in this highly complex and vital space.

4.40 **Persuasion** refers to using nudging or deterrence messaging to convince a user not to post or upload harmful content. This could include:

a) **Nudges:** Introducing deliberate friction through nudges[202] at the point of upload, for example prompting users to reconsider posting detected misogynistic content or other harmful or illegal gendered abuse or violence (see Case Study 6 and 7).[203] Nudging can also be used to deter users from non-consensually re-posting content, for example by using nudity detection to prompt users to reconsider re-sharing intimate images which were shared with them.

b) **Allowing users to verify their identity**. This may add accountability and reduce the online disinhibition effect (where individuals engage in harmful behaviours which they would typically refrain from engaging in real-life due to online affordances such as anonymity or asynchronous communication).[204] Identity verification may be useful in some circumstances (see **Case Study 7**), but it also comes with significant privacy issues, particularly for survivors and victims.[205] [206]

4.41 **Removal** refers to using technical tools to block uploads of harmful content, or to remove it after it has been uploaded.[207] This could include:

a) **Using hash matching** to prevent uploads of known intimate image abuse.[208] [209]

---

[201] See Page 5 of Ofcom's progress update implementing the Online Safety Act where we describe our plans to consult on additional measures in Spring 2025. See footnote 213 for more information. [accessed 20 February 2025].

[202] Nudges are design measures within an online environment to promote some behaviours and/or discourage others.

[203] Cox, A. L., Gould, S. J., Cecchinato, M. E., Iacovides, I. and Renfree, I., 2016. Design frictions for mindful interactions: The case for microboundaries, *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. [accessed 31 October 2024].

[204] Cheung, C.M., Wong, R.Y.M. and Chan, T.K., 2021 Online disinhibition: conceptualization, measurement, and implications for online deviant behavior , *Industrial Management & Data Systems*, 121 (1). [accessed 17 December 2024]; Ofcom Stakeholder Workshop 2 on Women and Girls Online Safety, 19 November 2024.

[205] We will be considering the details of identity verification further in a future consultation so providers who implement this may wish to review this in due course: Phase 3 consultation: Duties on categorised services, incl. transparency, see Ofcom's progress update implementing the Online Safety Act. [accessed 29 January 2025].

[206] We expect to update this reference to reflect our position at the time we publish the final Guidance.

[207] In this section, we focus on methods for doing this automatically rather than in response to user reports. Content moderation following user reports is covered in Chapter 5.

[208] Domestic Abuse Commissioner response to the November 2023 Illegal Harms Consultation, p.5; Refuge response to the November 2023 Illegal Harms Consultation, p.13; The Cyber Helpline response to the November 2023 Illegal Harms Consultation, p.18.; Victims Commissioner response to the November 2023 Illegal Harms Consultation, p.6.

[209] We are planning an additional consultation in Spring 2025 which will include measures on intimate image abuse hash matching to prevent the sharing of non-consensual imagery. We expect to update this reference to reflect our position at the time we publish the final Guidance.

b) **Implementing time out features** to users who repeatedly attempt to use service feature or functionality to perpetrate online gender-based harms. This means the affected user would not be able to send a message or use other platform features for a set amount of time while in time out. If a user continues to misuse the service repeated strikes could lead to an account ban (see Action 9: Take appropriate action when online gender-based harms occur in Chapter 5).
c) **Requiring consent** from those depicted in intimate content prior to uploading where adult content is allowed on a service to prevent intimate image abuse, including deepfakes (see **Case Study 13**).
d) **Implementing prompt and output filters** for GenAI models (see **Case Study 7**).

4.42 **Reduction** refers to limiting the circulation and visibility of misleading or harmful kinds of content rather than removing them entirely.[210] It is increasingly commonplace among services which allow user-to-user sharing of content mediated by recommender algorithms, as well as search services which mediate the results users find in response to queries but cannot remove content from those sites. This step could include:

a) **Deprioritising harmful content** in recommender algorithms to reduce its visibility and reach.[211]
b) **Removing links** to sites dedicated to hosting non-consensual images, or to services such as nudification apps used to generate non-consensual intimate content.
c) **De-monetising** user-generated content which promotes online gender-based harms but is not clearly illegal to prevent it from earning advertising income.[212]
d) **Blurring nudity and harmful content** using automated detection, giving users the option to unblur it if they want to see it.
e) **Scanning for duplicates**: When someone successfully requests the removal of explicit non-consensual fake content featuring them from search, making sure systems scan for – and delist from search – any duplicates of that image that are found.

4.43 **Automated detection**: the methods described in this section often rely on automated content moderation techniques to scan, identify, and filter content depicting online gender-based harms. .[213]

a) It is important that these automated systems are accurate, effective, contextually nuanced and minimise bias in terms of race, gender, and other characteristics.[214] For example, automated systems should seek to avoid erroneously removing content by survivors and victims that calls attention to online gender-based harms (see **Case Study 14**).[215]

---

[210] Gillespie, T., 2022. Do Not Recommend? Reduction as a Form of Content Moderation, *Social Media + Society*, 8 (3). [accessed 31 October 2024].
[211] Appelman, N., 2023. Disparate Content Moderation Mapping Social Justice Organisations Perspectives on Unequal Content Moderation Harms and the EU Platform Policy Debate. [accessed 22 October 2024].
[212] Jankowicz, N., Gomez-O'Keefe, I., Hoffman, L. and Vidal Becker A. 2024. It's Everyone's Problem: Mainstreaming Responses to Technology-Facilitated Gender-Based Violence. [accessed 13 February 2025.]
[213] We are planning an additional consultation in Spring 2025 on how automated detection tools can be used to mitigate a number of types of priority illegal content and content that is harmful to children. We expect to update this reference to reflect our position at the time we publish the final Guidance.
[214] Appelman, N., 2023. Disparate Content Moderation Mapping Social Justice Organisations Perspectives on Unequal Content Moderation Harms and the EU Platform Policy Debate. [accessed 22 October 2024].
[215] Kwarteng, J., 2022. Misogynoir: Challenges in Detecting Intersectional Hate, *Social Network Analysis and Mining,* 12 (1). [accessed 31 October 2024].

b) Continuously improving automated content moderation is important to identify content that could be harmful and avoid recommending content, forums or groups that are likely to encourage misogynistic content. While automated tools are unlikely to ever perfectly differentiate between categories of speech like irony, humour, counterspeech, and online-gender based harms, they can be continually evaluated and improved to reduce the risk of both overblocking (for example, blocking content recounting an experience of online-gender violence) and underblocking (for example, not adapting to new trends in content moderation circumvention) by applying more sophisticated methods that account for context such as fine-tuning content moderation models to be more culturally-sensitive, or developing multi-modal systems capable of understanding greater nuance.[216]

c) This process of improving automated content moderation could also include ensuring that moderation is effective for different kinds of formats like text, image, and voice. In-game voice chat moderation can also reduce exposure. For example, machine-learning driven voice moderation can detect toxic speech.[217]

Case study 13: Nudging to deter uploading of harmful content

Introducing deliberate friction[218] using nudges aimed at potential perpetrators can discourage uploading harmful behaviour or content without blocking it.

A popular example implemented by a range of services is 'preliminary flagging'. When a user attempts to post harmful content, a moderation algorithm classifies the content as harmful or violating a service's community guidelines. Deterrence messaging can also be deployed where potentially harmful behaviour (rather than content) – such as repeatedly messaging a user they haven't engaged with before without a response – is detected.

For example, a dating service could use deterrence messaging to reduce harmful interactions. Within a conversation, if harmful content is detected in a message, the sender could be prompted with a warning to reconsider the language used and be given the option to edit the message before sending. Deterrence messaging can be used in tandem with supportive messaging (for more information on supportive messaging, see Chapter 5). For instance, if a user who is prompted with a warning message opts to send the content anyway, then the user who receives the message could be prompted with a supportive message with information about how to report and block other users on the service. Feedback loops between user reporting and deterrence messaging can improve both the efficacy of content moderation and deterrence messaging.

---

[216] Peterson-Salahuddin, C., 2024. Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation, *Big Data & Society*, 11 (2). [accessed 23 December 2024]; Chan, A. J., Redondo García, J. L., O'Donnell, C., Silvestri, F. and Palla, K., 2024. Enhancing Content Moderation with Culturally-Aware Models. [accessed 23 December 2024]; Arya, P., Pandey, A. K., Patro, S. G. K., Tiwari, K., Panigrahi, N., Naveed, Q. N. and Khan, W. A., 2024. MSCMGTB: A Novel Approach for Multimodal Social Media Content Moderation Using Hybrid Graph Theory and Bio-Inspired Optimization, *IEEE Access* (12). [accessed 23 December 2024]; Wang, W., Huang, J., Huang, J., Chen, C., Gu, J., He, P. and Lyu, M. R., 2023. An Image is Worth a Thousand Toxic Words: A Metamorphic Testing Framework for Content Moderation Software. [accessed 23 December 2024].

[217] Pappas, M., 2023. Social Safety in Games: Moderating Voice Chat in the Metaverse, *ACM Games: Research and Practice*, 1 (3). [accessed 31 October 2024].

[218] Cox, A. L., Gould, S. J., Cecchinato, M. E., Iacovides, I. and Renfree, I., 2016. Design frictions for mindful interactions: The case for microboundaries, *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. [accessed 31 October 2024].

While such nudges show promising effects in improving online pro-social behaviour, existing studies emphasise the importance of getting the messaging right.[219] Particular care needs to be taken with deterrence messages shown to children so that the messages do not lead children to be ashamed and avoid discussing negative online experiences with their parents or guardians.[220]

Deterrence messaging will always be limited in deterring highly motivated offenders, and in identifying subtle forms of abuse (such as coercive control or misgendering).[221] There is also the risk of the impact of such nudges reducing over time as users become accustomed to the prompts and automatically click-through them. An intervention with a learning component or rotating several nudge messages could help to maintain effectiveness.[222]

Case study 14: Preventing image-based sexual abuse on adult services

Providers of adult content services face higher risks of hosting non-consensual intimate content because their sites allow sexual content.[223] Where a service provider identifies it is at risk of hosting nonconsensual intimate content, there is a variety of measures it can implement to mitigate this risk, such as:

Persuasion:

- **Consent nudging:** Implementing a prompt asking the user if they have asked for consent from all parties depicted within the content (if the algorithm detects more than one person). A study by the Cyber Civil Rights Initiative found that 66% of perpetrators listed "if I had taken more time to think about what I was doing" as a reason that would have stopped them from posting nonconsensual images.
- **Uploader verification**. Users must verify their identity to upload content, providing a full legal name, date of birth, a piece of matching government-issued photo ID, and a live face scan check. This can also include removing historic videos from unverified accounts.
- **Deterrence messaging**: Warning messages about the illegality and consequences of intimate image abuse.

Removal:

- **Hash matching**. An automated system cross-references uploaded content against a database of hashes for previously reported non-consensual intimate images, with matches removed and prevented from being shared. This can involve cross-industry initiatives such as StopNCII.org, which allow survivors and victims to generate hashes

[219] Warner, M., Strohmayer, A., Higgs, M., Rafiq, H., Yang, L. and Coventry, L., 2024. Key to kindness: reducing toxicity in online discourse through proactive content moderation in a mobile keyboard. [accessed 30 October 2024]; Katsaros, M., Yang, K. and Fratamico, L., 2022. Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content, *Proceedings of the International AAAI Conference on Web and Social Media*, 16 (1). [accessed 30 October 2024]; OpenWeb (Simon, G.), 2020. Nudge Theory Examples In Online Discussions [accessed 30 October 2024].

[220] Ofcom Stakeholder Workshop 2 on Women and Girls' Online Safety, 19 November 2024.

[221] Ofcom Stakeholder Workshop 2 on Women and Girls' Online Safety, 19 November 2024.

[222] Jahn, L., Rendsvig, R. K., Flammini, A., Menczer, F. and Hendricks, V. F., 2023. Friction Interventions to Curb the Spread of Misinformation on Social Media. [accessed 24 October 2024]; Kovacs, G., Wu, Z. and Bernstein, M.S., 2018. Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition, *Proceedings of the ACM on Human-Computer Interaction*, 2. [accessed 24 October 2024].

[223] Illegal Harms Register of Risks.

from their intimate images. These hashes are shared across participating service providers to detect and prevent the circulation of these images.[224]

- **Consent verification**. Users must certify that all individuals depicted in uploaded content have consented to appear and must provide identity verification for those depicted. Service provides can use facial recognition and nudity detection to block uploads of content if the uploader cannot provide proof of consent. This approach could also allow depicted users to withdraw consent, especially in content involving nudity.

Providers can layer different techniques to prevent intimate image abuse. Service providers may also refer to the Image-Based Sexual Abuse Principles on preventative approaches for the development of industry best practices.

Case Study 15: Automated detection of misogynoir content and results

Digital misogynoir refers to online hate and dehumanising language experienced by Black women online, particularly on social media services.[225] Although some of this content is likely to be illegal hate speech,[226] existing automated hate speech detection tools are ineffective at detecting it due to a lack of sensitivity to context.[227] This issue is particularly likely when such algorithms are trained on datasets which are tagged as just racist speech or misogynistic speech, therefore missing the intersections between the two. The effectiveness of these tools can be strongly influenced by the identity of annotators labelling hate speech, as well as other decision-makers within service providers.[228]

Researchers at Glitch, an online safety charity, and the Open University have been developing methods to better detect misogynoir, including through training on datasets of self-reported misogynoir.[229] For such techniques to be successful, they need to recognise that safety measures which treat different kinds of abuse (such as racism and misogyny) in isolation will fail to account for intersectional hate.

Digital misogynoir can also occur in search services. For example, researcher Safiya Noble found that searches for the term "Black girls" were more likely to result in pornographic results and sexually explicit terms than searches for "white girls".[230] Likewise, searches for variations of "Black women" led to racist and sexist suggestions in autocomplete. To address these harms, search services could monitor their systems for embedded bias.

---

[224] Ofcom has not assessed this particular tool for accuracy, effectiveness and freedom from bias.

[225] Bailey, M., 2021. Misogynoir Transformed: Black Women's Digital Resistance. [accessed 31 October 2024]. Glitch, 2023. The Digital Misogynoir Report: Online abuse against Black women allowed and enabled to thrive. [accessed 22 October 2024].

[226] See Illegal Content Judgements Guidance Section 3 Threats, abuse and harassment (including hate). [accessed 13 February 2025].

[227] Kwarteng, J., 2022. Misogynoir: Challenges in Detecting Intersectional Hate, *Social Network Analysis and Mining,* 12 (1). [accessed 31 October 2024].

[228] Kwarteng, J., 2022. Misogynoir: Challenges in Detecting Intersectional Hate, *Social Network Analysis and Mining,* 12 (1). [accessed 31 October 2024].

[229] Glitch, 2023. The Digital Misogynoir Report: Online abuse against Black women allowed and enabled to thrive. [accessed 22 October 2024]. Kwarteng, J., 2022. Misogynoir: Challenges in Detecting Intersectional Hate, *Social Network Analysis and Mining,* 12 (1). [accessed 31 October 2024].

[230] Noble, S. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. [accessed 19 December 2024].

5. Supporting women and girls

Warning: this chapter contains content that may be upsetting or distressing.

Overview

Context

5.1 Harm prevention addresses how online gender-based harms can be prevented or minimised before they arise, including during the development of a product. This chapter considers how service providers can embed safety-by-design to effectively support users and address harm, including once a product is in operation. This includes user control tools that give users more agency over their experiences, user support tools that provide helpful and accessible information, and harm response systems that support women and girls when harm occurs. These mechanisms consider the experiences of women and girls as they interact with the product and other users.

5.2 Some services may have no user controls, user support, or harm response mechanisms. Other services may have some mechanisms, but these may not sufficiently account for online gender-based harms. For example, evidence shows many services do not offer users easily navigable tools or clear information about how to use them.[231] A study from Refuge found that 95% of survivors who reported online abuse to a service were not satisfied with the response they received – and over half never received a response.

5.3 This chapter looks at how providers can improve women and girls' online safety through user controls, user support, and harm response. We specifically look at how providers can take the following actions:

- Action 7: Give users better control over their experiences.
- Action 8: Enable users who experience online gender-based harms to make reports.
- Action 9: Take appropriate action when online gender-based harms occur.

5.4 For each action, we set out a solid baseline of what safety looks like ('**foundational steps**') to help providers meet their new duties to protect UK users. We also highlight additional '**good practice steps**' to illustrate how providers can build on the foundational steps to create safer experiences for women and girls, give their users more autonomy, and provide assurance that users can seek appropriate redress for any harm that does occur.

5.5 Harm response systems are especially important as some perpetrators of online gender-based harms are highly motivated to cause harm and will attempt to evade safety measures – even if the measures have been designed and deployed to prevent abusability.

5.6 Actions 7 and 8 are focused on how providers can empower users and give them control over their online experiences.[232] These two actions are important because they acknowledge that online safety may look different for different women and girls and can

[231] W3C, 2024. Web Content Accessibility Guidelines (WCAG) 2.2. [accessed 10 January 2025].
[232] The examples of good practice that this chapter outlines do not prejudge anything we might recommend for the purposes of complying with the user empowerment duties that will apply to Category 1 services, which we will be consulting on no later than early 2026.

change over time. However, this layer of personalisation should be supported by strong groundwork from the provider to ensure that responsibility for user safety does not sit solely with users of the service.

Our target outcomes

5.7 Women and girls should have more agency to shape their online lives, and those that experience harm should be offered appropriate support. Providers should offer user control, user support, and harm response tools that account for the dynamics and complexities of online gender-based harms.

5.8 If providers develop and deploy user controls and user support tools, women and girls will gain more control over the content and users they encounter online and will have accessible information available to help them make informed choices about risk. This will allow them to make personalised decisions about what safety looks like for them – including the ability to make changes if their circumstances change.

5.9 Harm response includes providers designing and operating reporting systems which are easy to find, easy to use, and fit for purpose, as well as supporting women and girls when harm occurs. This includes specialised support for survivors and victims of online gender-based harms such as domestic abuse and stalking.

5.10 Media literacy also plays an important role in enabling women and girls to respond to online gender-based harms by equipping users with the skills and critical thinking needed to engage effectively with user control, user support, and harm response tools such as reporting systems. Tools to give users more control and facilitate reporting should be designed in line with the Best Practice Design Principles for Media Literacy.

Action 7: Give users better control over their experiences

5.11 Users often feel that they lack control over their online experiences.[233] Women and girls frequently experience unwanted behaviour and encounter unwanted content on online services.[234] Too often, providers do not deploy easily accessible[235] and navigable tools or provide users with clear information about how to change their content and safety settings. This makes it harder for users to curate their experiences online and increases the risk of harm.[236]

5.12 Providers can empower users by providing them with greater and more granular control over their own experiences. Increased options over who contacts them, what they see and what information about them is visible can allow users to reduce the risk of experiencing or encountering gender-based harms and minimise their impact when they occur.[237]

---

[233] Centre for Data Ethics and Innovation, 2020. Online targeting: Final report and recommendations. [accessed 30 December 2024].
[234] Ofcom, 2024. Experiences of using online services. [accessed 7 January 2025].
[235] W3C, 2024. Web Content Accessibility Guidelines (WCAG) 2.2. [accessed 10 January 2025].
[236] Centre for Data Ethics and Innovation, 2020. Online targeting: Final report and recommendations. [accessed 30 December 2024]; The Behavioural Insights Team, 2020. Active Online Choices: Designing to Empower Users. [accessed 30 December 2024].
[237] Grimani, A., Gavine, A. and Moncur, W., 2022. An evidence synthesis of covert online strategies regarding intimate partner violence, *Trauma, Violence, & Abuse*, 23 (2). [accessed 30 December 2024].

Increased control enables users to provide feedback on the content they see and make informed choices about what safety looks like for them.

5.13 Providers can use tools such as choice bundles (see Chapter 4) to make it easier for users to engage with increased options over their online experiences.[238]

Foundational steps: What are the expectations for service providers?

5.14 Our Codes set out the following steps for service providers:[239]

a) **Block and mute:** Users can block[240] and mute[241] other user accounts.[242]
b) **Disable comments**: Users can disable comments on their own posts[243] (see **case study 15**).
c) **Negative feedback:** Children can give negative feedback on content that is recommended to them by a content recommender system.[244]
d) **Group chats:** Children are given the option to accept or decline an invite to a group chat.[245]
e) **Supportive information:** Children using a service are provided with supportive information to help them make informed choices about risk.[246] Children are provided with information when they restrict interactions with other user accounts or content.[247] See **case study 16** for further details on how this can be put into practice.
f) **Support materials**: Children are provided with age-appropriate user support materials explaining the user safety tools available to them on a service.[248] Children are signposted to support when they report content and when they post, re-post, or search for harmful content.[249]

---

[238] The Behavioural Insights Team, 2021. Active Online Choices: Designing to Empower Users. [accessed 30 December 2024].
[239] The 'foundational steps' refer to a range of expectations we have already set out for service providers, either in draft for the purpose of consultation or in final form at set out in a decision at the time of publication of this Guidance. Where the measures which are currently in draft are changed in our final statement following the consultation on those measures, we will revise this Guidance accordingly. The 'foundational steps' apply to different services based on functionality, risk, and size. Information about which foundational step applies to which service is set out in the Guidance at a Glance document.
[240] A user tool that enables: the blocked user(s) cannot send direct messages to the blocking user; the blocking user will not encounter any content posted by blocked user(s) on the service and vice versa; the blocking user; and blocked user(s) if they were connected, will no longer be connected.
[241] A user tool that enables: the muting user will not encounter any content posted by muted user(s) on the service; and the muted user(s) is not aware that they have been muted and continues to encounter content posted by the muting user.
[242] Illegal Content (ICU J1), Draft Protection of Children (PCU G1).
[243] Illegal Content (ICU J2), Draft Protection of Children (PCU G2).
[244] Draft Protection of Children (PCU F3).
[245] Draft Protection of Children (PCU G4). This measure also appears as a foundational step for 'Action 5: Set safer defaults' as the measure is relevant both for preventing harm and supporting women and girls.
[246] Illegal Content (ICU F2). This measure also appears as a foundational step for 'Action 5: Set safer defaults' as the measure is relevant both for preventing harm and supporting women and girls.
[247] Draft Protection of Children (PCU E2).
[248] Draft Protection of Children (PCU E1, PCS E1).
[249] Draft Protection of Children (PCU E3). This measure also appears as a foundational step for 'Action 6: Action 6: Reduce the circulation of content depicting, promoting or encouraging online gender-based harms' as the measure is relevant both for preventing harm and supporting women and girls.

Case study 15: Disabling comments

- Pile-ons are a common type of coordinated harassment where a large group of users target either an individual, or a much smaller group of users. Perpetrators of pile-ons intend to shame and silence the individuals they target, particularly women and girls in public life such as politicians, journalists, activists, and athletes.
- Comment features can be used by perpetrators in pile-ons to respond directly and publicly to other users' posts with threatening or abusive messages.[250]
- Enabling users to disable comments on their own posts (before or after posting) means that women and girls can respond to this harm when it occurs, or when they judge that it is at risk of occurring.

> A user could disable comments on a post during a pile-on to hide any existing comments and prevent any further comments.

> A user who is concerned about experiencing a pile-on could disable comments before posting to prevent this harm.

- Women and girls can disable comments on their posts without having to restrict the visibility of their posts.

Case study 16: Supportive information

- Online misogyny circulates through a wide range of content online, from posts on dedicated forums trivialising sexual assault to viral videos on social media sites that glorify domestic abuse. This type of content can cause harm and evidence shows it can lead to women and girls withdrawing from online participation.[251]
- While a small number of users deliberately search for this content, many encounter it unintentionally through content recommender systems, including boys.[252]
- Signposting to supportive information that is clear and accessible can increase users' awareness of the user control tools available to them and encourage users to consider their safety online.

> A user who encounters a misogynistic video online and reports the content could receive supportive information. This information could set out steps they can take to further restrict content, accompanied by an explanation of the kind of content they may be restricting.

- Informed choices about content restriction can prevent users being exposed to further harm.

Good practice steps: How can service providers go further?

5.15 Service providers can build on the foundational steps to offer users greater and more granular control over their experiences. This can encourage users who may be at heightened risk of harm – for example public figures or those experiencing stalking or

[250] Wilson Center (Jankowicz, N., Hunchak, J., Pavliuc, A., Davies, C., Pierson, S. and Kaufmann, Z.), 2021. Malign Creativity: How gender, sex, and lies are weaponized against women online. [accessed 30 December 2024]; PEN America (Vilk, V. and Lo, K.), 2023. Shouting into the Void. [accessed 30 December 2024].
[251] Glitch and EVAW, 2020. The Ripple Effect: COVID-19 and the epidemic of online abuse. [accessed 30 December 2024].
[252] Vodafone, 2024. AI 'Aggro-rithms': young boys are served harmful content within 60 seconds of being online. [accessed 30 December 2024].

domestic abuse – to curate what safety looks like for them. These tools, in tandem with the tools aimed at preventing and reducing abuse discussed in Chapter 4, could help tackle the 'silencing' of women and girls on online services where abuse constrains their ability or willingness to participate in discourse for fear of abusive responses. The good practice steps service providers could take include:

a) **Visibility settings:** Allowing users to delete or change the visibility settings of content they upload, including content uploaded in the past.
b) **Block and mute:** Providing users with tools to block and mute multiple accounts simultaneously (see **case study 17**).
c) **Identity verification:** Allowing users to filter out content from all users who have not completed identity verification.
d) **User controls:** Providing adult users with greater control over what content is recommended to them by content recommender systems.
e) **User preferences:** Allowing users to signal what kind of content they do not want to see, and what kind of content they want to see more of (see **case study 18**).
f) **Supportive information:** Signposting users to supportive information which addresses specific harms such as domestic abuse or image-based sexual abuse. This could include supportive resources, specialist services, or information about where and how to report a crime.[253]

Case study 17: Mass blocking

When online gender-based harms occur, survivors and victims should be able to limit their interactions with and exposure to perpetrators. The ability to block other users provides women and girls with greater control over who can follow their accounts, who can interact with their posts, and who can directly message them.

A social media platform could provide users with more extensive blocking options. For example, the option to block not only another user's account but also any other accounts the user might have, as well as any new accounts the user may create in future. The platform could also allow users to block any current or future accounts connected to a particular phone number or email address. A social media platform could also offer users more automated blocking options. For instance, if a user sees a post that is offensive or disturbing to them, they could be given the option to block not only the post's author but all users who have reposted.

Providing users with additional blocking options reduces the burden of safety work on survivors and victims, and creates friction for perpetrators attempting to evade blocks by creating new accounts. Providing women and girls with greater control over which users they can restrict interactions with enables them to reduce their risk of experiencing online gender-based harm.

[253] Department for Science, Innovation & Technology and PUBLIC, 2025. Platform Design and the Risk of Online Violence Against Women and Girls. [accessed 12 January 2025].

Case study 18: Content filtering

Filters give users greater control over their experience online and prevent them from encountering unwanted content, including the use of words and phrases which amount to online gender-based harm.

Content filters can provide users with control in a variety of ways, such as by allowing them to reduce the amount of violent or sensitive content they are shown. A social networking platform could provide content filtering tools that allow users to identify topics they do not want to engage with by flagging specific tags, keywords, and phrases they do not want to see. Content filters are not usually case sensitive, and keyword filters should also work on any terms which include the keyword.

It takes time for users to set up filters, but it allows them to personalise the content they see. The tool empowers women and girls to shape their online experiences and avoid content which contains words and topics likely to be offensive, disturbing, or upsetting to them.

Action 8: Enable users who experience online gender-based harms to make reports

5.16 Users who experience or encounter online gender-based harms may face challenges in reporting this to providers.[254] Where reporting systems do not effectively account for harms such as online domestic abuse and image-based sexual abuse, survivors and victims can experience invalidation and re-traumatisation from the failure of providers to recognise the harm they have experienced, as well as frustration caused by lack of action to tackle the harm.[255] Poor experiences with reporting systems erode trust with users and many survivors and victims stop reporting online gender-based harms to providers.[256]

5.17 Improving reporting systems is crucial, as this will allow users to inform providers when harm occurs so that providers can take further action where appropriate. However, it is important providers recognise that user reporting relies on survivors and victims of online gender-based harms, and that reporting processes are time-intensive and risk re-traumatising survivors and victims.

5.18 Providers can encourage and enable users to make reports by designing reporting systems which are accessible, transparent, easy-to-use, and account for the specific dynamics of online gender-based harms. A safety-by-design approach incentivises providers to encourage and enable users to report online gender-based harms. For instance, they can design trauma-informed reporting systems which consider the specific needs and requirements of survivors and victims.[257]

[254] Department for Science, Innovation & Technology and PUBLIC, 2025. Platform Design and the Risk of Online Violence Against Women and Girls. [accessed 12 January 2025].
[255] Refuge, 2021. Unsocial Spaces. [accessed 30 December 2024]; The Open University (Jurasz, O.), 2024. Online violence against women: a Four Nations study. [accessed 30 December 2024].
[256] Victims Commissioner (Storry, M. and Poppleton, S.), 2022. The Impact of Online Abuse: Hearing the Victims' Voice. [accessed 30 December 2024]; PEN America (Vilk, V. and Lo, K.), 2023. Shouting into the Void. [accessed 30 December 2024].
[257] Chayn (Hussain, H.), 2021. Chayn's trauma-informed design principles. [accessed 30 December 2024].

Foundational steps: What are the expectations for service providers?

5.19 Our Codes set out the following steps for service providers:[258]

a) **Complaints processes**: Complaints processes enable users, affected persons, and interested persons to make relevant complaints[259] (see **case studies 19** and **20**).
b) **Complaints systems:** Complaints systems are easy to find, easy to access, and easy to use and allow users to add supporting information to their complaints.[260]
c) **Complaints communications**: Complaints systems acknowledge receipt of complaints, provide indicative timeframes,[261] and set out information about how the complaint will be handled[262] (including giving users the option to opt-out of communications from a service).[263]
d) **Predictive search**: Users are given means to easily report predictive search suggestions, and those which the provider determines present a clear and material risk of users encountering harmful content should no longer be presented to users.[264]

Case study 19: Affected persons

- Intimate image abuse can include the non-consensual sharing of both images created consensually and images created non-consensually, such as deepfakes.
- User reporting is an important mitigation against intimate image abuse, especially for services on which users can encounter nudity and sexually explicit content.
- Allowing affected persons[265] to make complaints makes it easier for women and girls who experience intimate image abuse to report it.

> A survivor and victim can report intimate image abuse depicting them as an affected person, even if they are not a user of the service it has been uploaded to. This means that intimate image abuse can be reported without the survivor and victim having to create a user account to access the service's reporting system.

> A survivor and victim of intimate image abuse can ask another individual to report images on their behalf. This reduces the risk of re-traumatisation.

- Complaints processes reduce friction for survivors and victims in reporting intimate image abuse.

[258] The 'foundational steps' refer to a range of expectations we have already set out for service providers, either in draft for the purpose of consultation or in final form at set out in a decision at the time of publication of this Guidance. Where the measures which are currently in draft are changed in our final statement following the consultation on those measures, we will revise this Guidance accordingly. The 'foundational steps' apply to different services based on functionality, risk, and size. Information about which foundational step applies to which service is set out in the Guidance at a Glance document.
[259] Illegal Content (ICU D1, ICS D1), Draft Protection of Children (PCU C1, PCS C1).
[260] Illegal Content (ICU D2, ICS D2), Draft Protection of Children (PCU C2, PCS C2).
[261] Illegal Content (ICU D4, ICS D3).
[262] Illegal Content (ICU D5, ICS D4), Draft Protection of Children (PCU C4, PCS C4).
[263] Illegal Content (ICU D6, ICS D5).
[264] Illegal Content (ICS F1), Draft Protection of Children (PCS E2).
[265] Affected persons include a person – other than a user – who is the subject of the content, a member of a group with a certain characteristic targeted by the content, a parent or adult with responsibility for a child who is a user of the service, or an adult providing assistance in using the service to another adult.

Case study 20: Reporting options

- Harassment is an offence involving a course of conduct, which includes causing another person alarm or distress.[266] The perpetration of harassment is often highly personal and can involve patterns of behaviour aimed at isolating the survivor and victim.
- This behaviour does not always occur through written and visual communication such as images, comments, and direct messages which can be easily recorded and reported.
- Offering accessible reporting for all types of content and interaction supported on a service ensures that women and girls are always able to report harassment to providers.

> A user on a gaming service can report harassment when it happens during in-game voice chat. This prevents perpetrators subverting the service's enforcement systems by using voice chat.

> A user on a virtual reality service can report a perpetrator sexually harassing them and invading their physical space.

- Reporting systems must be updated to cover any changes providers make to their services, including new possibilities for user interaction.

Good practice steps: How can service providers go further?

5.20 To further strengthen their reporting systems, service providers can follow additional good practice that creates a more trauma-informed and tailored process for reporting online gender-based harms. This could encourage more users to report content and help users to provide the information a service needs to make an appropriate decision. This would lead to better experiences for women and girls, as well as giving services a better picture of how harms are manifesting on their service. The good practice steps providers could take include:

a) **Exit buttons:** Providing a 'quick exit button' throughout the reporting process which immediately takes the user out of the reporting system.[267]
b) **Report tracking:** Allowing users to track and manage their reports and tailor their experience throughout the complaints process (see **case study 21**).
c) **User feedback:** Allowing users to give feedback to the service provider on their reporting process.[268]
d) **Trusted flaggers:** Establishing a trusted flagger programme in partnership with organisations that have expertise in online gender-based harms.[269] (see **case study 22**).
e) **Incident reporting:** Allowing users to report incidents of abuse, including abuse that happened on another service or offline. Service providers may take various approaches in responding to these reports (see **case study 23**).[270]

---

[266] The offence is only committed if the perpetrator knows or ought to know that their conduct amounts to harassment.
[267] Chayn (Hussain, H.), 2021. Chayn's trauma-informed design principles. [accessed 30 December 2024].
[268] Service providers should have regard to the needs of their UK user base in considering what languages are needed when designing their reporting process.
[269] Refuge response to the 2023 Illegal Harms Consultation, p.12; Suzy Lamplugh response to the 2023 Illegal Harms Consultation, p.21; SWGfL response to the 2023 Illegal Harms Consultation, p.15; UCL Gender and Tech response to the 2023 Illegal Harms Consultation, p.10; VAWG Sector Experts response to the 2024 Protection of Children Consultation, p.13.
[270] Service providers should have regard to the needs of their UK user base in considering what languages are needed when designing their off-service incident reporting process.

f) **Media literacy:** Adopting the principles on user-centric design and timely interventions in Ofcom's Best-Practice Design Principles for Media Literacy.

Case study 21: Track and manage reports

Experiences of online gender-based harms are often complex and highly contextual, and frequently involve multiple interactions or pieces of content. Reporting is time-consuming and can be re-traumatising for survivors and victims.[271] Reporting systems that allow users to track and manage their reports can provide survivors and victims with greater agency and transparency over the process.[272]

The Web Foundation's Tech Policy Design Lab developed a prototype for a reporting dashboard that enables users to track their reports and see when reports are resolved. The dashboard could allow users to add additional context to their reports and collect and archive evidence of harmful content. Providers could also give users the option to invite a trusted contact to support with the reporting process.

Providing users with greater choice over the reporting process allows women and girls to tailor the reporting process to their experiences and preferences. This can help overcome challenges in the reporting of online gender-based harms and build trust between providers and survivors and victims.

Case study 22: Trusted flaggers

Reporting online gender-based harms and engaging with providers can be a challenging process for individual survivors and victims. Trusted flagger programmes can assist with this process by building relationships between providers and organisations with expertise in harms such as online domestic abuse and intimate image abuse.[273]

A 'Violence Against Women and Girls Code of Practice' developed for industry by a civil society coalition recommends that providers set out clear criteria for what content trusted flagger organisations can report and provide a specific route for escalation if providers do not respond to trusted flagger reports. The coalition also emphasise that trusted flagger organisations should be provided with the necessary resource and support when carrying out additional work to make a service safer.

Developing partnerships between providers and organisations with expertise in gender-based harms gives survivors and victims additional support and advocacy. These partnerships can also be used to alert providers to emerging forms of harm.

Case study 23: Reporting off-service behaviour

Online gender-based harms are not always restricted to a single interaction or piece of content. Harms such as stalking are part of a wider pattern of behaviour, both online and offline. Reporting systems which allow survivors and victims to flag gender-based harm that

[271] Chayn (Hussain, H.), 2021. Chayn's trauma-informed design principles. [accessed 30 December 2024].
[272] PEN America (Vilk, V. and Lo, K.), 2023. Shouting into the Void. [accessed 30 December 2024].
[273] Ofcom Stakeholder Workshop 2 on Women and Girls Online Safety, 19 November 2024.

has happened offline or on another service enable providers to reduce the risk of harm occurring on the service.

A livestreaming service could introduce a policy that enables users to report harmful off-service behaviour. This would allow users of the service who have experienced stalking offline to inform the provider. The provider could investigate and, if satisfied that sufficient verifiable evidence has been given, the provider may take enforcement action. Enforcement action could include blocking the perpetrator from interacting with the survivor and victim on the service to prevent any future abusive behaviour.

Accounting for off-service instances of gender-based harms such as stalking helps tackle harmful patterns of behaviour and enables providers to take proactive action to prevent online gender-based harms occurring on their service.

Action 9: Take appropriate action when online gender-based harms occur

5.21 Providers often do not allocate sufficient resource and expertise to ensure appropriate action is taken when users experience or encounter online gender-based harms. Where this is the case, providers may fail to respond to user reports, and their internal content and search moderation systems may not identify harmful content. In other cases, providers do act but without specific consideration of the nuances of gender-based harms and the importance of survivor and victim agency.[274]

5.22 Online gender-based harms are often highly contextual. Harms such as intimate image abuse and harassment may not immediately be identified by content or search moderation. Existing evidence shows particularly poor response rates for reports of intimate image abuse.[275]

5.23 Providers can reduce the impact of online gender-based harms by taking appropriate action when it occurs on their service. Content and search moderation that accounts for harms such as intimate image abuse and harassment enables providers to identify patterns of harmful behaviour. Improving responses to user reports allows providers to better support women and girls who have experienced online gender-based harms.

Foundational steps: What are the expectations for service providers?

5.24 Our Codes set out the following steps for service providers:[276]

[274] Refuge, 2021. Unsocial Spaces. [accessed 30 December 2024]; Refuge response to the 2023 Illegal Harms Consultation, p.4; Glitch response to the 2023 Illegal Harms Consultation, p.6.
[275] Qiwei, L., Zhang, S., Kasper, A., Ashkinaze, J., Eaton, A., Schoenebeck, S. and Gilbert, E., 2024. Reporting Non-Consensual Intimate Media: An Audit Study of Deepfakes. [accessed 30 December 2024].
[276] The 'foundational steps' refer to a range of expectations we have already set out for service providers, either in draft for the purpose of consultation or in final form at set out in a decision at the time of publication of this Guidance. Where the measures which are currently in draft are changed in our final statement following the consultation on those measures, we will revise this Guidance accordingly. The 'foundational steps' apply to different services based on functionality, risk, and size. Information about which foundational step applies to which service is set out in the Guidance at a Glance document.

a) **Take down**: Content moderation function that allows the swift take down of illegal content and content harmful to children. [277]
b) **Performance targets**: Set targets for content and search moderation functions. [278]
c) **Prioritisation**: Prepare and apply policies on the prioritisation of content for review. [279]
d) **Moderation teams**: Content and search moderation teams are resourced [280] and receive training [281] (see **case study 24**).
e) **Complaints**: Complaints about suspected illegal content and content harmful to children should be handled in accordance with content prioritisation processes. [282]
f) **Appeals**: Appeals should be determined according to performance targets, [283] depending on the provider. Complainants should, as far as appropriate and possible, be restored to the position they would have been in had the decision not been made, following upheld appeals. [284]

Case study 24: Domestic abuse training

- Domestic abuse is perpetrated in complex and highly personal ways, and online domestic abuse often replicates and extends the same dynamics as offline domestic abuse. [285]
- User reports of online domestic abuse are contextual and may focus on a pattern of behaviour rather than a single interaction or piece of content. [286]
- Training content and search moderation teams on domestic abuse enables moderators to better identify instances of this harm and respond to user reports.
  - Trained content and search moderation teams could take into account considerations such as how to respond to user reports appropriately without escalating offline risks to survivors and victims.
  - Providers, specifically content and search moderation teams, could develop their understanding of domestic abuse and how it occurs on their service through partnerships with organisations who have frontline experience and expertise.
- Content and search moderators should receive adequate support and safeguarding from providers to undergo training and carry out this work.

Good practice steps: How can service providers go further?

5.25 By building on the foundational steps, providers can further embed understanding of online gender-based harms into their systems and processes. This will enable them to respond to harm in a way that supports survivors and victims, while also minimising the risk of future harmful behaviour and content circulating on their services. Appropriate action can include:

[277] Illegal Content (ICU C2), Draft Protection of Children (PCU B1).
[278] Illegal Content (ICU C4, ICS C3), Draft Protection of Children (PCU B3, PCS B4).
[279] Illegal Content (ICU C5, ICS C4), Draft Protection of Children (PCU B4, PCS B5).
[280] Illegal Content (ICU C6, ICS C5), Draft Protection of Children (PCU B5, PCS B6).
[281] Illegal Content (ICU C7, ICS C6), Draft Protection of Children (PCU B6, PCS B7).
[282] Illegal Content (ICU D7, ICS D6), Draft Protection of Children (PCU C5, PCS C5).
[283] Illegal Content (ICU D8, ICS D7), Draft Protection of Children (PCU C6, PCS C6).
[284] Illegal Content (ICU D10, ICS D9), Draft Protection of Children (PCU C8, PCS C8).
[285] Woodlock, D., McKenzie, M., Western, D. and Harris, B., 2020. Technology as a Weapon in Domestic Violence: Responding to Digital Coercive Control, *Australian Social Work,* 73 (3). [accessed 14 March 2025].
[286] Refuge, 2021. Unsocial Spaces. [accessed 30 December 2024].

a) **Enforcement action:** Taking enforcement action against users who continually violate a service's terms of service (see **case study 25**).
b) **Fact-checking and labelling:** Adding fact-checking and labelling to content can be a useful tool to address gendered disinformation, which can be used in coordinated harassment campaigns against women and girls in public life.[287]
c) **Watermarks and metadata:** Adding watermarks and metadata to content can indicate where content is synthetic and how it has been edited or manipulated.[288] This can also be a useful tool for sex workers and adult content creators whose intimate images have been re-shared without their consent.[289]
d) **Upholding bans:** Identifying and preventing the creation of new accounts by banned users.
e) **Moderator review:** Sending high risk and highly contextual user reports of gender-based harms for review by specifically trained moderators.[290]
f) **Hiding content:** Hiding potentially harmful content while it is assessed in content moderation, such as potential harassment or intimate image abuse, can minimise harm to survivors and victims while the assessment is completed.
g) **Reporting channels:** Creating dedicated reporting and review channels for online gender-based harms.[291]

Case study 25: Action on serial perpetrators

A small number of users are often responsible for a large amount of online gender-based harm, particularly in cases of co-ordinated harassment.[292] Evidence shows these users engage in repetitive and abusive behaviour which targets women, such as repeatedly posting the same sexually explicit content.[293]

A social networking service with a Generative AI feature could implement a strike-based enforcement system, where a user receives a 'strike' against their account for misuse of the service. For instance, if a user attempts to generate harmful content through the Generative AI feature, they could receive a strike against their account. If a user receives multiple strikes, their access to the Generative AI feature could be removed for a given period of time. If a perpetrator continues to misuse the service, repeated strikes could lead to an account ban.

---

[287] Internet Governance Forum, 2021. Best Practice Forum on Gender and Digital Rights: Exploring the concept of gendered disinformation. [accessed 30 December 2024]; National Democratic Institute, 2021. Addressing Online Misogyny and Gender Disinformation: A How-To Guide. [accessed 30 December 2024].
[288] Ofcom, 2024. Deepfake Defences: Mitigating the Harms of Deceptive Deepfakes. [accessed 30 December 2024].
[289] Sanders, T., Trueman, G., Worthington, K. and Keighley, R., 2023. Non-consensual sharing of images: Commercial content creators, sexual content creation platforms and the lack of protection, *New Media & Society*, 27 (1). [accessed 30 December 2024].
[290] Automated content moderation tools can be effective. However, it is important that service providers recognise that certain user reports of online gender-based violence are nuanced and highly contextual and so human moderators with specific training are likely to be highly effective in addressing these.
[291] Service providers should have regard to the needs of their UK user base in considering what languages are needed when designing their dedicated reporting and review channels for online gender-based violence
[292] Ofcom Stakeholder Workshop 2 on Women and Girls Online Safety, 19 November 2024.
[293] Wilson Center (Jankowicz, N., Hunchak, J., Pavliuc, A., Davies, C., Pierson, S. and Kaufmann, Z.), 2021. Malign Creativity: How gender, sex, and lies are weaponized against women online. [accessed 30 December 2024].

Users should be informed when they receive a strike and what the consequences of a strike are. Providers should also include the ability for users to appeal strikes and related enforcement action. Applied effectively, strike-based enforcement systems can act as a form of deterrence to prevent a single act of abusive behaviour from becoming a pattern.

It is worth noting that content moderation and tools assessing user behaviour are likely to involve processing of personal data. This includes moderation actions applied to a user's account (such as a strike, service restriction or ban). We encourage services to consult guidance from the Information Commissioner's Office. [294]

[294] See ICO guidance Content moderation and data protection and planned forthcoming ICO guidance Behaviour ID Tools for Online Safety (Due for publication: Spring 2025). [accessed 13 February 2025].

A1. Glossary

Warning: this chapter contains content that may be upsetting or distressing.

A1.1 This glossary sets out definitions of terms used throughout the Guidance.

Term	Definition
Abusability testing/evaluations	Abusability testing or abusability evaluations draw on the concept of "usability" which is used to evaluate how easy it is for users to navigate a website or device to accomplish their goals. In contrast, abusability evaluations test how easy it is to abuse a tool or feature for harm, and therefore point to ways that abusability can be minimised in design.
Action	Actions are intended to assist user-user and search service providers in designing, testing, deploying and operating their services in a manner that takes responsibility for the safety of women and girls on their services.
Blocking	To take action that will result in the blocking user and blocked user being unable to send direct messages to each other or encounter each other's content, and to become unconnected if they were connected.[295]
Coercive and controlling behaviour	The repeated or continuous engagement in behaviour by a perpetrator towards a victim, with whom they are personally connected, that is controlling or coercive, and this behaviour has a serious effect on the victim, putting them in fear of violence or causing serious alarm or distress which has a substantial adverse effect on their usual day-to-day activities.[296] This can include assault, threats, humiliation and intimidation or other abuse that is used to harm, punish, or frighten the victim.
Connection list	A feature that allows users to find other users that they may not already be connected with. This list can sometimes be suggested by the platform based on user interests and existing connections.
Consent	Permission for something to happen.
Co-opt	In the context of this document, we use this term to refer to using a good or service in a way that is different from the usual or intended purpose.
Content recommender systems	An algorithmic system which determines the relative ranking of an identified pool of content (that includes regulated user generated content) from multiple users on content feeds. Content is recommended based on factors that it is programmed to account for, such as popularity of content, characteristics of a user, or predicted engagement. References to content recommender systems do not include a content recommender system employed exclusively in the operation of a search functionality which suggests content to users in direct response to a search query, product recommender systems or network recommender systems.

[295] A more precise definition is contained in paragraph ICU J1.3 of Recommendation ICU J1 (user blocking and muting) in the Draft Illegal content Codes of Practice for user-to-user services

[296] An offence under section 76 of the Serious Crime Act 2015.

Term	Definition
Counterspeech	The practice of responding to speech that seems harmful or offensive. It can take many forms such as challenging, debunking, or critiquing harmful speech, amplifying alternative viewpoints, providing accurate information, and fostering empathy and understanding.
Cyberflashing	The sending of a photograph or film of genitals, intending the recipient will be caused alarm, distress or humiliation, or sending a photograph or film of genitals to obtain sexual gratification and being reckless as to whether the recipient will be caused alarm, distress or humiliation.
Deepfake	Forms of audio-visual content that have been generated or manipulated using AI, which misrepresent someone or something. Deepfake content also include Intimate Image Abuse (IIA).
Domestic abuse	Incident or pattern of behaviour that is used by someone to control or obtain power over their partner or ex-partner: this is also known as coercive and controlling behaviour, which is a crime. It is never the fault of the person who is experiencing it.
Doxing	The intentional online exposure of an individual's identity, private information or personal details without their consent. Doxing can take place in the context of pile-on harassment, honour-based violence, and coercive control.
Duties	The relevant legal duties we are required to focus on for the purposes of this guidance under s.54 of the Act are set out in Parts 3 and 4 of the Act and are applicable to providers of Part 3 services (user-to-user and search services). The Act imposes duties which require providers to identify, mitigate and manage the risks of harm from illegal content and activity and content and activity that is harmful to children. Certain additional duties are set out in Part 4 of the Act, including some which apply only to categorised services.
Foundational step(s)	Expectations we have already set out for service providers – either in final or draft form – relevant to achieving an action. This includes actions service providers can take to help them comply with their duties related to risk assessments and transparency, and when looking to implement measures set out under Codes of Practice, in the context of protecting women and girls.
Gaslighting	Psychological manipulation of a person usually over an extended period of time that causes the victim to question the validity of their own thoughts, perception of reality, or memories and typically leads to confusion, loss of confidence and self-esteem, uncertainty of one's emotional or mental stability, and a dependency on the perpetrator.
Gendered disinformation	Gendered abuse online that uses misleading or false gender-based narratives against women and their participation in public life. It is a combination of online disinformation through falsity, coded language to evade moderation and detection, and coordination.
Generative AI or GenAI	Refers broadly to machine learning models that can create new content. Models create a wide variety of outputs including text, images, video, and audio.

Term	Definition
Good practice steps	Further information on how services can tackle online gender-based harms that build on the expectations we have set out in the foundational steps.
Harm prevention	Refers to systems which attempt to anticipate and mitigate risk before harm happens. In the context of product development, it can include testing to identify potential routes for abuse, and to allow for changes in features and functionalities to prevent harm.
Harm response	Refers to systems which address harm and minimise its effects, leading to support or restitution for the person who experienced harm. In the context of online gender-based harms users experience or encounter on a service, this includes reporting systems which are easy to find, easy to use, and fit for purpose, and taking appropriate action to address the impact of harm.
Hash-matching	Hashing is an umbrella term for techniques that create a 'fingerprint' of a given piece of content. In practice this means using an algorithm to analyse content and create a 'hash' that can represent it. Hashes are then stored in a database that can be accessed by multiple parties as required. In the context of online safety, online platforms can use hashing to notify other platforms of illegal or harmful content they have identified, and vice versa. Hashing databases exist for CSAM, terror content, and non-consensual intimate images.
Image based sexual abuse	Taking, creating, sharing, or threatening to share intimate images without consent, including intimate image abuse and cyberflashing.
Intimate image abuse	An offence of sharing or threatening to share intimate images or film. Other terms for this include nonconsensual intimate image abuse, commonly referred to as "NCII", and "revenge porn".
Intersectionality	The interconnected nature of social categorisations such as race, class, and gender, regarded as creating overlapping and interdependent systems of discrimination or disadvantage; a theoretical approach based on such a premise. The term was coined in 1989 by Kimberlé Crenshaw describing how traditional feminist theory and antiracist policies exclude Black women who face overlapping discrimination unique to them.[297]
Malign creativity	The use of coded language; iterative, context-based visual and textual memes; and other tactics to avoid detection on social media platforms.
Media literacy	The ability to use, understand and create media and communications across multiple formats and services. Ofcom has specific media literacy duties as set out in Chapter 2 of this document.
Misogyny influencers	Influencers on social media who promote misogynistic views via videos and podcasts, often in combination with offering advice about relationships, mental health and wellbeing, and achieving material success and status.

[297] Crenshaw, K., 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, *University of Chicago Legal Forum*, 1989(1). [Accessed 12 December 2024].

Term	Definition
Monitoring and assurance	Function to provide independent assurance that measures taken to mitigate and manage the risks of harm to individuals identified in the risk assessment are effective on an ongoing basis. This function should report to, and its findings should be considered by, either: a) the body that is responsible for overall governance and strategic direction of a service; or b) an audit committee.
Muting	To take action that will result in the muting user not encountering the content of the muted user unless the muting user visits the user profile of the muted user.[298]
Non-designated content	A category of content harmful to children defined in the Act, broadly: content, which is not primary priority content or priority content, of a kind which presents a material risk of significant harm to an appreciable number of children in the United Kingdom.[299]
Nudges	Design features which lead or encourage users to follow the designer's preferred paths in the user's decision making.
Online domestic abuse	Using technology for coercive and controlling behaviour in the context of an intimate relationship.
Online gender-based harms	Harmful and abusive content and activity directed towards women and girls. For the purposes of the Guidance, we focus on four overlapping forms of harm which seek to enable, enact or normalise forms of misogyny, sexism and gender-based violence: online misogyny, piles on and online harassment, domestic abuse, and image-based sexual abuse.
Online misogyny	Content that normalises, actively encourages, or cements misogynistic ideas, attitudes, or behaviours. This includes content that incites hatred towards, abuses or threatens women and girls, as well as sexual or explicit content that normalises or encourages harmful sexual behaviour.
Perpetrator	A user, individual or group who conducts and participates in online gender-based harms.
Pile-ons	Refers to when a user is criticised or targeted by a large number of other users, often as part of bullying campaigns.
Product	An all-encompassing term that includes any functionality, feature, tool, or policy that a service provides to enable users to interact with or use the service
Service provider	We refer to a platform as 'service', and the legal entity that provides the service as 'provider' or 'service provider'. In other words, a website might be the 'service', and the company that owns and runs it would be the 'provider'.

[298] A more precise definition is contained in paragraph ICU J1.3 of Recommendation ICU J1 (user blocking and muting) in the Draft Illegal content Codes of Practice for user-to-user services.
[299] Section 60(2)(c) of the Act.

Term	Definition
Primary priority content	A category of content that is harmful to children, as defined in section 61 of the Act.[300] This includes pornography, suicide, self-harm, and eating disorder content.
Priority content	A category of content that is harmful to children, as defined in section 62 of the Act.[301] This includes hate and abusive speech.
Red teaming	A type of evaluation method that seeks to find vulnerabilities in GenAI models.
Reduction	Limiting the circulation and visibility of misleading or harmful kinds of content rather than removing them entirely.
Re-traumatisation	The re-experiencing of thoughts, feelings or sensations experienced at the time of a traumatic event or circumstance in a person's past. This can be triggered by, for example, hearing a smell or sound associated with the traumatic event, or by being put into a similar situation or experience.
Safety-by-design	A proactive approach to integrating safety considerations into the design and development of products, systems, or processes.
Safety work	Online and offline strategies women employ to respond to, avoid, and cope with gender-based violence. This can include avoiding certain actions (like posting online or walking alone at night), spending time thinking or planning about safety risks, or moderating how you dress or present yourself online to avoid experiencing violence. Safety work is an unfair burden which limits women's space for action and self-expression, and makes the responsible for preventing violence.
Search service	An internet service that is, or includes, a search engine. For the purpose of this Guidance, we primarily focus on "general search services" which are services that enable users to search the internet by inputting search requests. It derives search results from an underlying search index (developed by either the provider of the service or a third party). Search results are presented using algorithms that rank based on relevance to a search request (among other factors). This is because we don't have substantial evidence on the risks of online gender-based harms on vertical search services.[302]
Serial perpetrators	A user, individual or group who repeatedly conducts and participates in online gender-based harms. In many cases, the majority of online gender-based harms comes from a relatively small proportion of a services' user base, who are highly motivated to continue this pattern of behaviour.

[300] We have typically grouped the different kinds of primary priority content as follows: pornographic content, suicide and self-harm content, eating disorder content.
[301] We have typically grouped the different kinds of priority content as follows: abuse and hate content, bullying content, violent content, harmful substances content, dangerous stunts and challenges content.
[302] For more information on vertical search services, please see: Protecting people from illegal harms online - Annex 3: Glossary [accessed 24 January 2025].

Term	Definition
Supportive information	Also referred to as Crisis Prevention Information. Refers to information provided by a search service in search results that typically contains the contact details of helplines and/or hotlines and links to trustworthy and supportive information provided freely by a reputable and reliable organisation.
Trauma-informed approach	An approach which acknowledges (i) the prevalence of trauma; (ii) how trauma affects all individuals involved with the programme, organisation, or system, including its own workforce; (iii) and responds by putting this knowledge into practice. This often includes an emphasis on informed consent as well as content warnings. Examples include trauma informed design and trauma-informed reporting systems.
Usability	The quality or state of being convenient and practicable for use. In the context of product design, it is a measure of how well a specific user in a specific context can use a product/design to achieve a defined goal effectively, efficiently and satisfactorily.
User centric design	Refers to a design process which is inherently iterative and puts user needs at the centre of every stage of this process, to create highly usable and inclusive products.
User-to-user service	An internet service by means of which content that is generated directly on the service by a user of the service, or uploaded to or shared on the service by a user of the service, may be encountered by another user, or other users, of the service.
User controls and user support	Refers to tools which give users more agency over the content, users, and experiences they encounter online and accessible information to make informed choices about risk.
Survivor and victim	A person who has experienced online gender-based harms.
Victim-blaming	Explicitly stating or implying that the victim is to blame for the abuse they have experienced. It often focuses on actions that a victim could have taken (or not taken) to avoid experiencing abuse.
Viral	A piece of content – such as a post, image, or video – that achieves a high level of popularity by being quickly and widely shared online, particularly on social media.
Virality	The degree to which online content spreads easily and/or quickly across many online users, alongside how much engagement and/or views a piece of content received (such as ‘shares’, ‘likes’, and ‘view’, etc.).