July 2026

Exploring how internet users navigate sensitive and harmful content

YouGov®

Report prepared by YouGov Qualitative.

CONTENTS

YouGov®

CONTENT WARNING

The research findings include references to users' experiences of online content that in some cases may be distressing, including reference to hate and discrimination, and graphic violence.

YouGov®

1 Key Insights

Key Insights – Experiences of sensitive & harmful content

In Autumn 2025, YouGov conducted qualitative research - via an online community and text-based focus groups - with 75 adult internet users from diverse backgrounds across the UK. Key insights included the following:

- Sensitive and harmful content was **ubiquitous in participants' online experiences, often encountered daily** while scrolling, or through personalised feeds, making it hard to avoid, according to participants.

- **Women and participants from minority backgrounds** were **more likely** to share experiences of encountering content they characterised as **sensitive or harmful**, such as hate, and violent content. They spoke about the negative potential impact these types of content could have on their wellbeing and sense of safety. In contrast, other participants shared fewer of these experiences.

- **Sensitive and harmful content** was often described as **intrusive**, appearing unexpectedly even when participants had taken steps to avoid it. There were perceptions that inflammatory or shocking content was prioritised and amplified by platform algorithms.

- It was felt that the amount of sensitive and harmful content circulating online had **increased** over time. Some **platforms and features** (such as comment sections) were more **frequently associated with sensitive and harmful content,** though few platforms were described positively in relation to online safety.

Key Insights – Attitudes towards user control and content filtering tools

- While participants **enjoyed the unpredictability of discovering new content**, this sat alongside a broader feeling **that users lacked control** over the risk of seeing something in their feed that could be harmful or upsetting to them. There were concerns around the effectiveness of online tools for navigating sensitive and harmful content, as well as issues with awareness, undermining interest in these.

- Participants had **mixed feelings** about the idea of **filtering content** out of their online experience. The idea of such tools being more available was welcomed and seen as having the potential to create a safer, healthier online experience for individuals. Participants saw content filter tools as particularly **relevant for supporting their wellbeing, and some envisaged more 'temporary' use – for example, if they felt seeing certain kinds of content was contributing to poor wellbeing –** and for other users they thought might be 'vulnerable.'

- However, there were **a range of concerns about the limitations of content filter tools and whether some content might be important to see even if it was potentially upsetting.** Many concerns related to **effectiveness and accuracy:** whether content a user didn't want to see could be mistakenly shown to them or prevent them from encountering content they wanted to engage with. Related to this, participants were worried about **potential biases** in platforms' decisions about what content to filter, and that the filtering tools would encourage users to live in **echo chambers.**

- When asked about whether they wanted to filter out **content that was deemed hateful or offensive** to an aspect of the participant's own identity or community – such as racist, sexist, or homophobic content – users often wanted **to know what hateful things were being said**, rather than filter this out, with some participants noting they might want to take action on it (such as supporting users being targeted with hate).

YouGov

Key Insights – Attitudes to other features for navigating sensitive content

- Participants expressed a preference for **content filtering and related tools** (e.g., setting preferred topics) to involve more **granular choices based on personal preferences**, and for them to be clearly visible, with explanations as to what difference applying such tools would mean to their online experience.

- Participants generally **valued content overlays** that required a **conscious decision to proceed**, as these were seen as allowing users to prepare themselves and retain agency. This meant users often **preferred the idea of having overlays compared to excluding content**, though some admitted that this approach might tempt them into seeing content they might not want to.

- There was widespread existing use of tools that let **users take action against individuals** (e.g., blocking/muting) and were seen as **valuable** in managing what they saw online and **allowing granular control of what to hide**. Users were familiar with **reporting** accounts and content and had some negative experiences of feeling their reports were not acted on.

- **Actions against individual pieces of content** (e.g., “see less of this”) were in some respects seen as a **helpful and relatively familiar** way in which participants **already navigate online content**, but in some scenarios **involved high effort with little reward**, for handling content that people were repeatedly bothered by.

Key Insights – Attitudes to verified status

- Participants associated **verified status* with public figures** and professional contexts, which meant they carried **some legitimacy and authenticity.** However, these positive associations were tempered by **distrust of monetised verification models and lack of information about how verified status is obtained** across platforms.

- The majority of participants **did not see strong value in filtering out users without verified status in their daily online experiences**. They **did not easily see** how this would **help avoid sensitive or harmful content** and wanted to continue seeing content they enjoyed and hear from a wide range of voices (who they did not think would have verified status) including friends and family.

- Participants recognised **contexts where filtering users without verified status might be relevant**— for example**, on dating apps, where user identity was linked to personal safety**, or in circumstances of **intense harassment or coordinated abuse.** Filtering out users without verification status was rarely seen as a sufficient safety solution on its own.

- Participants were asked to consider that verified status could be granted based on **platforms verifying that the internet user's identity was authentic. This had little impact on the perceived value of verified status or filtering**. Participants generally expressed **little motivation to verify their own identities**, viewing it as more relevant for public figures or specific professional settings. Many were wary of sharing official identification, though more comfortable with platforms using information they already held. Participants also highlighted how they didn't necessarily share any of their identity with other users on platforms where they mainly consumed content without posting.

** 'Verified status' refers to when a user who has been 'verified' by the platform. We explored perceptions of this feature as it is used on a range of services, including (but not limited to) when it is accompanied by a badge or 'tick' on the user's profile, and/or when verification is based on verifying someone's identity.*

2 Background

Background to this report

As the UK regulator for online safety and media literacy, Ofcom's mission is to make communications work for everyone, which includes to make online services safer for the people who use them. To do this, it relies on evidence to inform its work and fulfil its duties. In-depth research on the experiences and views of people across the UK supports Ofcom to better understand citizens and consumers, which is central to its policymaking.

Ofcom commissioned YouGov to carry out qualitative research among adults looking into content controls and tools that can be used to navigate "sensitive" and potentially harmful types of content, as well as the verification status of other users. Research was considered relevant to support the following areas of Ofcom's work:

1. **Additional Online Safety Duties** for categorised services, particularly those related to user empowerment and identity verification. The research was intended to form part of the evidence base of the perceptions of adult internet users towards a range of online tools and features that are relevant to these duties, particularly those for which there was less existing evidence, such as tools for filtering out specific categories of content, and tools for filtering out content from users who do not have 'verified' status.
2. Contributing to Ofcom's evidence base on **Women and Girls' Online Safety** to support its understanding of the harmful ways in which regulated services may be used, especially harmful content and activity disproportionately affecting particular groups, including women and girls.
3. Researching and promoting **Media Literacy**, including via Ofcom's Making Sense of Media programme (MSOM), in which it is important to understand how users can benefit from online tools that support them with navigating online content.

The research was published at the same time as the Additional Duties consultation and is one of a wider range of evidence sources that assisted the development of the consultation proposals.

How to read this report and consider the research findings

The following factors should be considered when interpreting the findings of this research:

- This study is qualitative and exploratory in nature, meaning that the findings are **not statistically representative of UK adults** more generally. The sample was also skewed towards participants with protected characteristics more commonly targeted in online hate.
- The findings rely on participants' **experiences, attitudes, and speculations about how they might behave in certain scenarios**, rather than being based on observed behaviour.
- Participants' views on online safety tools and features referenced **should not be taken as evidence of their technical feasibility, proportionality, or effectiveness**. The report does not seek to verify whether participants' perceptions accurately reflect the functionality or safety measures implemented by the platforms referenced.
- **References to specific online platforms** are included to illustrate participants' experiences and perspectives. These references should not be understood as indicating the prevalence or origin of particular types of content on those platforms, but rather as reflecting the platforms used by participants and their individual experiences.
- Any findings or observations presented in this report are based on the perceptions of the adults who participated in the research and **do not represent the views of Ofcom or YouGov**.
- While this report references online tools and features that are also discussed in Ofcom consultation documents and regulatory outputs, its findings **should not be interpreted as any decision of Ofcom or as reflecting any policy position that Ofcom may adopt** in its role as the online safety regulator.

Research objectives

Research objectives:

- Understanding, experiences and attitudes towards sensitive and potentially harmful content.
- Understanding and experiences of user verification status.
- Attitudes towards using content controls and user empowerment features to navigate sensitive and potentially harmful content, including filtering out certain types of users and content.
- How user attitudes vary, in particular among people whose identity is more likely to be targeted in hateful content.

Key research questions, stage 1 – Online community

- How do users interact with social media and video-sharing platforms, what are their general experiences, and what types of content do they come across?
- What does ‘sensitive content’ mean to users, and what are their attitudes and experiences with sensitive content online and on different social media and video-sharing platforms?
- What are user attitudes towards managing sensitive content and their reactions towards relevant content controls and features?
- What are user attitudes towards filtering out content that is hateful towards different characteristics?

Key research questions, stage 2 – Focus groups

- What do users understand by user verification status and how does it impact experiences online?
- What are user attitudes towards verification status, and how do these vary across different platforms and content types?
- What are user attitudes towards managing content from users with and without verified status?
- What are user attitudes towards sharing information about their identity to obtain a verified status?

Methodology: Online community

YouGov conducted a **2-stage research project** with **75 participants. Stage 1** ran between **20th October to 24th October,** and **Stage 2** took place between **10th to 12th November 2025**, with a 2-week break in between to identify emerging findings and inform the focus group discussion guide.

Stage 1: 5-Day Online Community

Participants logged in daily to the research platform to complete structured activities, discussions, and diary tasks focused on experiences with sensitive content and platform controls. The two discussion tasks were the only ones in which participants could see responses from others.

Day 1 – Online experiences

- *Explored participants' typical online routines, platforms used, and recent positive/negative experiences.*

Day 2 – Attitudes to managing sensitive content

- *Perceptions of "sensitive" content, concerns, and views on hiding or avoiding such content.*

Day 3 – Discussions (filtering sensitive content)

- *Live discussion boards debating the advantages and drawbacks of filtering sensitive content.*

Day 4 – Tools for managing content

- *Assessed awareness, usage, and perceived effectiveness of existing platform tools.*

Day 5 – Discussions (avoiding hateful content)

- *Interactive discussions on filtering hateful content, participants considered scenarios for using filters.*
- *On Day 3 and 5, each participant was added to a discussion board made up of people with whom they shared specific personal characteristics in order to tailor the discussion about online hate (e.g. racism, misogyny, ableism) in a way that was relevant to individual circumstances.*

Daily diary task

- *Across all days, participants uploaded examples of content they found on social media platforms, with commentary on whether they disliked, found concerning or considered it inappropriate.*

Sample frame: Online community

A total of 75 participants took part in the online community (Stage 1).

- Across the sample, participants had a mix of social media and online usage (high, medium, low).
- All were using at least one major social media/video sharing platform, and a range of such online platforms were represented across the sample.

The sample was purposefully skewed to participants with minority identities that are more likely to be targeted in hateful content.

- According to the Online Experiences Tracker and other evidence in Ofcom's Illegal Harms and Children's Registers of Risk, hate and other forms of content that target abuse towards identities and communities, most often target people from minority ethnic groups, people from minority religious groups, women, people with minority gender identities and sexual orientations, people born outside the UK, and people with disabilities.
- Most participants (n=62) belong to one of the groups/identities above.
- We took this approach in order to capture a diverse range of user experiences and perspectives, and ensure the research captured experiences of types of content that are relevant to the Online Safety Act (e.g. content relevant to additional duties for some online services).

Demographics of online community sample:

Gender
- 28 Male and 39 Female
- 8 Minority gender identity (non-binary, transgender, and gender fluid)

Sexual orientation
- 25 identified as being gay, lesbian, bisexual, or queer.

Ethnicity
- 18 Asian and British Asian (Indian, Pakistani, Bangladeshi and Chinese), 33 White (English, Welsh, Scottish, Northern Irish, British), 13 Black and Black British (African and Caribbean), 10 from other minority ethnic groups, 1 'Prefer not to say' ethnicity)

Religion
- 8 Islam, 4 Buddhism, 2 Hinduism, 1 Sikhism, 6 Church of England/Anglican/Episcopal, 1 Evangelical – independent/non-denominational, 1 Pentecostal, 1 Presbyterian/Church of Scotland, 7 Roman Catholic, 6 'Other, 4 'Prefer not to say', 34 No religion

Other
- 20 were born outside the UK*
- 24 had a disability*

Age
- 18-24: 17 participants
- 25-34: 22 participants
- 35-44: 23 participants
- 45-54: 8 participants
- 55+: 5 participants

*Excluding those who preferred not to share this information

Methodology and sample frame: Text-based focus groups

Stage 2: Three text-based Focus Groups (90 mins – 7-9 participants per group – total 26 participants)

A subset of participants from the online community further took part in text-based focus groups. The text-based focus groups explored perceptions of verification status across different platforms, as well as exploring related tools and identity verification.

Participants were allocated to groups according to specific behaviours and attitudes that might influence views on this topic:

- **Group 1** (Heavy and medium digital users who post regularly online)
- **Group 2** (Participants who agreed to either both or either of the following statements: *"I think that a lot of "sensitive" content that's available online should not be there" and "I would like to have more control over how much "sensitive" content I am shown online"*.
- **Group 3** (Light digital users who do not regularly post online)

Discussion guide

The discussions focused around examining how people understand and manage their online identity, what verified status means to them, and how it affects trust, safety, and engagement with content on different platforms.

The groups also explored attitudes toward making verified status more widely available, including potential filtering tools that limit interactions with users without verified status.

Finally, participants discussed what types of personal information they would feel comfortable sharing for verification and reflected on principles for an ideal, trustworthy, and accessible verification process.

Demographics of text-based focus group sample:

Gender: 9 Male, 13 Female, 4 Minority gender identity (non-binary/Prefer to use another term)

Sexual orientation: 9 identified as being either gay, lesbian, or bisexual.

Ethnicity: 5 Asian and British Asian (Indian, Pakistani, Chinese); 7 Black and Black British (African, Caribbean); 13 White (English / Welsh / Scottish / Northern Irish / British, Irish, Any other white background); 1 from other minority ethnic background.

Religion: 1 Buddhism, 2 Church of England/Anglican/Episcopal, 1 Evangelical – independent/non-denominational, 1 Hinduism, 3 Islam, 1 Presbyterian/Church of Scotland, 2 Roman Catholic, 3 from other religions not listed.

Terminology Note

Sensitive and potentially harmful content – This research explores how participants feel about and navigate sensitive and potentially harmful content online. We decided the best way to explore this with participants was to use the phrase “sensitive content.” This was purposely used to allow for subjective interpretation of the term and not to prime participants’ attention on types of content that could fall under its definition, e.g. hateful content. This was also to encourage discussion about how users engage with content that has not been removed by a platform they use. Sensitive content was defined in the community as:

- Content that could be offensive, distressing or upsetting to some people
- Content that’s not “family-friendly” or “safe for work”
- Other content that participants might think shouldn’t be allowed on the platforms, but which is currently allowed

Participants therefore included examples of content that is considered harmful under the law, as well as other content types, when discussing this topic. Within the report, we have used the phrase “sensitive and potentially harmful content” to refer to how participants understood the phrase “sensitive content”, unless reporting on insights in which participants provided alternative interpretations or are focusing on a specific form of content.

Platform – is used to refer to video sharing and social media services (e.g., YouTube, LinkedIn, X formerly Twitter, Facebook, Instagram, TikTok and Pinterest).

Default settings – settings relating to whether a feature is automatically applied (e.g. ‘on’ or ‘off’) in a user’s experience.

Online usage – was used to define the usage of social media and video sharing platforms on a weekly basis.
- High usage: Over 22 hours (per week)
- Medium user: 11 to 22 hours (per week)
- Low user: Less than 11 hours (per week)

Terminology Note – Continued

'Navigating' online content – describes internet users' experiences of online content, including the actions they take to respond to content that they see, as well as other reactions they have and choices, they make when managing their online experience. This includes using the range of online platform features explored in this research report, as well as other actions like scrolling past content or closing an app.

Public online spaces – spaces that can be accessed by most members of the public e.g. a general content feed of a social media/video sharing platform.

Private online spaces – spaces that are more controlled or restricted in access, e.g. a closed group chat where moderators approve access for individuals.

Verified status – refers to when a user has been 'verified' by the platform. We explored perceptions of this feature as it is used on a range of services, including (but not limited to) when it is accompanied by a badge or 'tick' on the user's profile, and/or when verification is based on verifying someone's identity.

Minority groups – refers to participants from groups including participants with a disability, participants from minority ethnic groups, participants from minority religions, participants with a minority gender identity and participants from a minority sexual orientation.

Participants from non-minority groups – refers to participants who do not belong to any of the above-mentioned groups.

Minority sexual orientation – refers to participants who identify as e.g. gay, lesbian, bisexual, or queer.

Minority gender identity – refers to participants who identify as e.g. trans-gender and/or non-binary.

Cisgender – refers to participants whose gender identities correspond with the sex they were assigned at birth.

YouGov

Terminology Note – Continued

We have used some overarching terms to categorise relevant content controls and tools:

- Content filtering tools – This refers to tools that enable users to change their online experience to avoid seeing certain categories of content, such as content types considered more ‘sensitive’ or potentially harmful.

- Content preference tools – This refers to tools that allow users to say change what content they are shown in other ways, such as what topics they prefer to see, or if they wish to ‘reset’ their recommendations.

- Global blocking tools – This refers to tools that block a large number of accounts from the user experience, for example, only seeing content from users you are ‘connected’ or ‘friends with, or only seeing content from ‘trusted’ or ‘verified’ users.

- Content overlays – This refers to tools that mask a user from seeing a specific piece of content (e.g. via a blur, interstitial, or warning label) and requires them to ‘click through’ or take a similar action in order to see it.

- Actions against individual users – This refers to actions like blocking and muting individual users.

- Actions against individual pieces of content and content types – This refers to reporting, ‘See less of this’ features, ‘Community notes’ features, and muting individual words and hashtags.

3 Experiences of online platforms, sensitive & harmful content

3.1 How participants used online platforms

Social media and video-sharing platforms are ingrained in participants' everyday lives

Social media and video-sharing platforms were described as deeply embedded in their daily routines. Many checked their phones immediately after waking, with heavier users spending up to 6 hours per day on weekends. Usage tended to drop during school or work hours, due to time pressures or deliberate efforts to set boundaries. Engagement increased afterwards, when platforms such as TikTok, Instagram, and YouTube were used for both entertainment and social connection.

Many participants expressed frustration at how easily screen time accumulated. They described getting drawn in despite intentions to cut back on social media usage, through habitual scrolling, 'doom-scrolling', or completely losing track of time. A recurring theme was difficulty avoiding phones during bedtime, with participants mentioning scrolling to delay sleep or even buying a blue light filter so they could continue using their devices.

Type of content was a major driver of both positive and negative experiences online.
Participants mainly consumed visual and short-form content, spanning entertainment, lifestyle, personal interests, educational resources, and news.

- Users going on platforms mainly for entertainment curated their feeds for relaxation, actively avoiding heavier topics or topics they found more sensitive, e.g. by scrolling past content they didn't want to see, and using 'dislike' / 'not for me' buttons.
- News-focused users considered social media an essential source of information, despite describing regularly encountering misinformation, sensationalism, and divisive perspectives.

Some participants still wanted to have the choice to see challenging or difficult content to stay informed. Even when content was upsetting or uncomfortable (e.g. political updates, conflicts, social issues), several participants felt it was important to remain aware of events, access multiple viewpoints and challenge their own assumptions. This contributed to them tolerating certain posts they otherwise disliked.

Platform experiences, including exposure to potential harms, were influenced by online behaviours

Content consumption was the primary use for online platforms:

- Participants used platforms to **view and engage with others' content**, while several **posted** short video content about travels or special activities. A few occasionally posted advice on Reddit.
- Passive users referenced trying to avoid seeing **political debates, repetitive content, and excessive ads**, while active users saw social media as a creative outlet but faced risks like unwanted comments and bullying during livestreams.
- Overall, many participants were frustrated by content they viewed as toxic and unpleasant, even if it was closely linked to some of their interests (e.g. staying up to date with the news).

"Mostly browse content... very little of my "online" presence outside of WhatsApp requires interactions with others." (Male, 39, Bangladeshi, No disability, Islam, Cisgender, Straight, Online usage – light)

"On Instagram, I usually just scroll the feed and share videos with friends. I can post something once or twice a month, and it's usually my pictures of travelling and countries that I visited." (Female, 27, White, No disability, Roman Catholic, Cisgender, Straight, Online usage – light)

"Instagram: I use it to post on my Stories, pretty much on a daily basis, sometimes 2-3 times a day." (Female, 31, White, No disability, Religion – Other, Cisgender, Straight, Online usage – medium)

Some engaged in specialised online communities tied to hobbies or interests:

- This included Facebook groups for neighbourhood news, gaming communities on Discord or Steam, or forums dedicated to topics like drones, African history, and religious history.
- **Those who used online communities tended to describe less frustration about online safety and sensitive content,** compared to other participants.
- Related to this, participants in this research also talked about going on social media platforms to communicate with family and friends.

"Recently been chatting in an online forum for the grey arrows drone club, I learned a lot from others about what drone to get and what to expect when taking exams" (Male, 42, White, No disability, No religion, Cisgender, Straight, Online usage – light)

Personalised feeds allowed participants to discover new content, which led to mixed reactions

Participants described much of their browsing as driven by **recommended or suggested short form content**, rather than deliberate searching. This often exposed them to posts they would not otherwise have encountered, creating **positive experiences of discovery but also surfacing content that felt irrelevant, intrusive, or distressing.**

While many valued the serendipity and variety of algorithm-driven content, participants also highlighted **clear distinctions between content they enjoyed, content they did not want, and content that was distressing but felt important.**

Some appreciated discovering new ideas, creators, or perspectives. Others found that the same algorithmic processes surfaced content that **felt irrelevant, intrusive, or misaligned with their interests**. Participants also frequently encountered distressing political or crisis-related content, which they sometimes wanted to avoid emotionally but still felt compelled to stay informed about.

These distinctions shaped participants' mixed feelings: enjoying discovery while wanting more control over content they did not choose to see.

Participants described wanting to:

- Stay aware of political and social issues
- Understand what is happening in the world
- Hear perspectives different from their own
- Maintain control over exposure to upsetting or intrusive content

> *"I prefer to see this type of content because people need to know what's going on in our society and to younger generations."* (Female, 39, Asian background, No disability, Buddhism, Cisgender, Straight, Online usage – high)

> *"Some people may not want to be reminded of racism or confronted with it however, it is important we have open conversations and discussions in society, it is history."* (Non-binary, 31, White, Disabled, No religion, Lesbian, Online usage – medium)

> *"Even though I find this issue very distasteful, it is important to be aware of the facts of this matter...to be aware of the nonsense that [political and public figures] are attempting to spread, that is why I access content like this."* (Male, 58, Indian, Disability, Islam, Cisgender, Straight, Online usage – light)

> *"It was distressing to hear, but I would not have skipped it. These children need to be heard and saved [in reference to the Gaza crisis]"* (Female, 60, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)

3.2 Attitudes to sensitive and harmful content

Participants' perceptions of "sensitive content" covered a range of content types, including online harms

Most felt "sensitive content" includes **upsetting or disturbing material, as well as offensive or discriminatory language which can negatively impact someone's mental health**. Some felt this content was directly linked to **that which could be deemed inappropriate for those under 18** (such as extreme violence/political opinions, or sexual content).

Primary areas of concern regarding sensitive content included:

- **Graphic violence and death**, such as beheadings, shootings, and war-related imagery (e.g., images / videos of Gaza and videos of the Charlie Kirk shooting were commonly mentioned). These were considered highly distressing yet easily accessible. Lesser mentioned violent content was animal abuse.
- **Hate speech and discrimination**. Most noted, there was a perceived increase in political views seen as extreme and Islamophobia, which they felt could harm minority communities. Participants also mentioned homophobic comments, though less frequently.
- **Triggering topics**. Sexual assault, mental health struggles, self-harm, and eating disorder content were seen as harmful to mental well-being. Lesser mentioned was body image-related content (e.g. pro anorexia content).
- **Sexually explicit content**. Participants were concerned about inappropriate exposure of younger audiences to this content.
- **Other mentioned concerns included** recordings of others who have not consented to be filmed/photographed/featured. For example, deepfakes or AI-generated content of people (including famous/deceased people) without prior permission.

Overall, concerns were concentrated around content that could provoke emotional distress, perpetuate social divisions, or expose vulnerable groups to harm. Participants had different thresholds for what kind of sensitive content they are comfortable seeing.

"'Sensitive content' means material that could be upsetting, offensive, or inappropriate for some people. Things like violence, graphic images, strong language, or topics that deal with trauma or personal issues. It's basically a way of saying, 'hey, just a heads up, this might not be comfortable for everyone" *(Male, 31, Caribbean, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – high)*

"Sensitive content means not suitable for all viewers, especially for younger audiences, those who are vulnerable or affected by certain strong themes. Anything that is labelled sensitive should be treated with caution." *(Female, 41, Mixed ethnicity, No disability, No religion, Cisgender, Straight, Online usage – light)*

Participants fear that sensitive and harmful content can deepen societal divisions and undermine safety for minority groups

Younger users and **marginalised** groups, including ethnic minorities and women, were more concerned about **hate speech, harassment, and sexualised content. Many ethnic minorities and some Muslim participants** were specifically most concerned about **racist and Islamophobic content and misinformation on social media.**

This type of content made them feel unsafe and reinforced harmful narratives, e.g., linking Muslims to terrorism. Participants noted how **real-world events escalate harmful online discourse**; for example, Islamophobic abuse after the Southport murders. These concerns extended beyond minority groups, with many perceiving an increase in extreme political rhetoric, which was viewed as fuelling racism and polarisation in the UK.

There were female participants who were concerned about **sexually explicit content being accessible to audiences under 18,** feeling it to be inappropriate.

Female participants also reported **mental health impacts from misogynistic comments and body image-related content.** Viewing this was seen to potentially trigger harmful behaviours such as 'body checking', dieting pressures, and unrealistic beauty standards.

There were participants, including those from religious minorities, who were concerned about **sexually explicit content or nudity**, describing it as something they didn't want to see and as potentially '**vulgar**'. There were also concerns about content that was seen as sexualising teenage children **or children suffering,** such as children starving or children in war zones.

Those who had **reported hateful content or comments**, such as racism, mentioned they received **inadequate responses from platforms**. They claimed they either received little or no acknowledgement/communication or were informed that the post/comment they reported was not deemed "sensitive" or "offensive" to warrant further action.

"[Racist and Islamophobic] content can incite violence, as we have seen in the recent past when people were sharing content and news without verifying…I came across religiously upsetting content… I was told, 'Why do I watch such content, if I don't like it?'" *(Male, 39, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)*

"Whenever crimes happen, we hold our breath to what colour or religion the perpetrator is – to prepare for the backlash." *(Male, 60, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)*

"I know a woman who runs social media for museums and gets a lot of sexualised messages directed at them through those museum accounts... It makes me sad." *(Female, 34, White, Disabled, No religion, Cisgender, Straight, Online usage – medium)*

"There is a lot of nasty behaviour on gaming platforms, i.e. misogynistic and racist" *(Non-binary, 31, White, Disabled, No religion, Gay or lesbian, Online usage – medium)*

3.3 What kinds of sensitive and harmful content participants encountered

Within these online experiences, it was common for participants to come across sensitive or harmful content, sometimes daily

In addition to the concerns, they voiced about harmful content in discussion, a ‘daily diary’ exercise in the online community provided concrete examples of what participants were actually seeing in their feeds.

They captured up to 10 pieces of content (posts, ads, videos, comments) they came across online that they: would have preferred not to see, believed they shouldn’t be shown, did not want to see again, thought others might not want to see, or had mixed feelings about.

What participants were seeing

Content ranged from **violent videos, hateful rhetoric, political misinformation, scams, and sexualised posts,** to algorithmically irrelevant content.

How the content they highlighted made them feel

The content participants showed to researchers had often **elicited strong emotional reactions** (such as discomfort, anger, shock, or mixed feelings). In other cases, participants highlighted instances of content that seemed irrelevant or was overly intrusive in their feeds.

What they did in response

The majority of participants **simply scrolled past or ignored content,** especially when it was irrelevant or mildly annoying.

When posts **felt more harmful, such as sexualised short-form clips or AI-generated posts presented as real**, they used light controls such as ‘not interested’ or unfollowing to reduce similar content.

The most **severe uploads, such as gruesome bodily injury videos and content inciting racism/violence,** prompted **reporting**, although many doubted its effectiveness. Occasionally, participants shared positive or informative posts or discussed disturbing ones with friends or family.

Exposure to hateful content, and at times distressing conflict or news contributed to discomfort during participants' browsing

General news and negative current affairs:

- Unexpected crime and accident reports, sensational headlines, and trending stories that participants found unpleasant surfaced, making feeds feel dominated by **unsolicited 'doom' updates** that spoiled the mood and felt intrusive when participants were not seeking news.
- Most did not encounter these stories intentionally and would often minimise exposure and click 'see less', citing frustration with clickbait framing.

Conflict, war and geopolitics:

- **Participants described seeing a wide spectrum of conflict-related content.** Many referenced standard news reporting of conflicts such as Israel/Palestine and Ukraine, which some found difficult or emotionally heavy even when the content was less graphic. Others encountered more distressing forms of content, including emotionally charged commentary, graphic or violent footage. This left them feeling **overwhelmed, anxious, angry, or helpless** during what they intended to be light browsing. In many cases, they **scrolled past to protect their mood**, but a few shared for awareness, discussed with friends, or reported content that looked misleading.

Hate, prejudice, and discriminatory content:

- Users reported encountering anti-immigration rhetoric, racist and antisemitic slurs, Islamophobic comments, homophobic or transphobic rhetoric, and posts supporting political figures or ideologies seen as extreme and hateful. Participants described such experiences on X, Instagram and Facebook.
- Content was perceived as **normalising hate or risking radicalisation**, prompting reports, blocks/unfollows and deliberate non-engagement to avoid algorithmic amplification.

"Seeing a stark mugshot attached to a story about double murder was instantly unsettling and injected a heavy, negative feeling into my feed."
(Male, 31, Caribbean, No disability, Church of England, Cisgender, Straight, Online usage – high)

"The footage was of bereaved family members trying to identify the bodies/remains of their loved ones who have been killed in Gaza. I feel uncomfortable watching people experience such a harrowing experience in a video that is sandwiched in between trivial reels designed for light-hearted entertaining consumption."
(Female, 41, White and Black Caribbean, Straight, Online usage – light)

"I don't enjoy seeing it [a video of a man shouting about foreigners], but it's also important to see, as it does give people an idea about some of the views that are still shared in this country and the struggle that people have day to day with racism, violence, etc." *(Male, 39, White, Cisgender, Straight, Online usage – medium)*

Violent content was described as particularly distressing, and sexualised content caused embarrassment

Violence, abuse and disturbing incidents:

- Participants encountered a range of violent or disturbing content, including auto-playing clips of fights and assaults, animal abuse, car crashes and content related to domestic abuse. One participant reported seeing a video depicting “a woman being beaten and abused”, which they found deeply distressing. Some users also came across videos featuring abusive, aggressive, or threatening language.
- These experiences were described as shocking, upsetting or inappropriate, particularly because the content often auto-played without warning. Parents highlighted additional concern when browsing near children who might briefly see disturbing material, even if not using the device themselves.
- Typical responses involved attempts to reduce exposure, such as selecting “not interested/see less”, reporting the content, or exiting the app altogether.

“This was the first post I saw when I opened X today, first one. A man physically abusing a donkey. I wish I didn't have to see videos like this everyday. It just straight away started playing when I opened X with no warnings at all beforehand.”
(Female, 44, White, Disabled, Church of England, Cisgender, Lesbian, Online usage – medium)

Sexualised or nudity content:

- Clips of a sexual nature, creator self-promotion of sexual content and unexpected nudity often appeared in short-form videos on feeds across most platforms. This content was described as causing embarrassment and violating, especially when family or colleagues were nearby.
- Several questioned the algorithm and felt concerned that this type of content was being pushed by low-quality moderation and monetised self-promotion.
- In more explicit cases, they did report the content, with the exception of a few that believed platforms rarely act on sexual-content reports.

“Sexually suggestive nature of content shown to me with a porn star... I'm not into this kind of stuff and not sure why these things keep popping up.”
(Male, 38, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)

Distrust in content, such as scams and misinformation, created concerns about others being influenced and misled

Scams, misleading content and misinformation:

- Participants uploaded examples of misleading ads, fake event promotions using celebrities, and clickbait headlines, describing these posts as deceptive and manipulative. Some worried that others, particularly less-confident and older users, might be taken in by content that appeared legitimate. Due to **past fraud experiences**, some participants started to **distrust platforms** because the content often looked legitimate at first glance.

Unpleasant ads and promotions:

- Participants flagged intrusive and repetitive adverts that they found suspicious, or were on topics they disliked/found irrelevant. The majority unfollowed or clicked 'not interested' on these adverts, as they felt overwhelming, irritating, or out of sync with their actual interests. There were also examples of 'inappropriate' adverts being shown on a child account being used by one research participant.

AI-generated and deepfake content:

- AI-generated videos, fake public-figure audio, manipulated political clips and deepfake videos/memes, made them feel uneasy, **distrustful** and confused **about what was real** because of a lack of labels.
- They worried such content could distort political debate or spread hate and described feeling mentally "on guard" while scrolling.

Health and body image content:

- Participants saw frequent appearance and weight-focused posts, alongside some eating-disorder-related material such as "body-checking" videos (which highlight extreme thinness unrelated to fitness videos) or "recovery" content, especially on short-form feeds. These posts **felt over-targeted**, repetitive and **triggering for some**, leading to participants ignoring or clicking 'see less'.

"I have been a victim of online fraud and honestly I do not wish for my friends or family members to come across such misleading posts that can cause them financial harm and emotional distress"
(Female, 29, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – high)

"A couple days ago I saw a fake AI generated movie poster and while I'm able to tell it's not real it can be very misleading for some people."
(Female, 26, White, Disabled, No religion, Cisgender, Straight, Online usage – light)

"This advert on Instagram about abs plastic surgery in Turkey. Looks suspicious and may promote body dysmorphia in other people watching the advert"
(Male, 21, Asian, No disability, Buddhist, Cisgender, Straight, Online usage – medium)

For around half of participants, encountering sensitive and harmful content is a daily occurrence and felt to be unavoidable

Participants were split **fairly evenly** between those who encountered sensitive content **daily,** and those who encountered it less often, with the former group made up more of **women and those in minority groups**.

Participants **wanted to stay informed and sometimes** accepted they might **inevitably see certain kinds of sensitive content**, and may even need to do so **in order to keep up with news**.

There is little tolerance from participants for overtly hateful content or comments towards certain groups, which are deemed unacceptable, and this is especially felt by women and minority groups.

"I come across a lot of sensitive content, it's the political content, the violent imagery, that is most distressing to me because that is what is so harmful to us as a society. But I do feel the need to keep aware of what's happening in the world, however painful that is."
(Male, 35, White, Disabled, No religion, Cisgender, Bisexual, Online usage – high)

"I do come across sensitive content fairly often even when I'm not looking for it, it just shows up in my feed…I understand that being online means I might still run into some sensitive content now and then. The internet is huge, and it's impossible to completely avoid it." *(Male, 31, Caribbean, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – high)*

"I need to hear more about these stories to understand the world around me. So while I might not like listening to them I am quite grateful I can hear their stories." *(Female, 19, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – medium)*

Participants felt that some platforms they used featured more sensitive and harmful content (1/2)

There were platforms sometimes perceived as doing less to moderate sensitive and harmful content

There were perceptions that some platforms acted in ways that enabled **more 'toxic' behaviour** from their users.

Participants sometimes felt these services **promoted harmful narratives** and did not moderate enough content.

These perceptions were **more common** for platforms with **comment/reply sections**, which were seen as spaces where some kinds of harmful content were common (e.g. hate).

Platforms that featured video content came with risks of being overwhelming or intrusive

While there were examples of video sharing platforms that participants felt safer on, it was **common** for participants to talk about **the risk of** being **overwhelmed by short-form video content**, and the risk of seeing unexpected, **intrusive content** they didn't want to see without a warning (e.g. graphic violence).

Platforms were generally seen as less risky when they were used for a narrower purpose

Participants expressed **minimal concern** about sensitive content on platforms used for creative hobbies and recipes, and that they used for professional reasons.

Community spaces were also viewed more positively than other more general online spaces.

While participants acknowledge the challenges platforms face in moderating vast volumes of content, there is **scepticism** about platforms' **ability and willingness** to adequately and appropriately manage sensitive material. Regardless, this sentiment does not imply a desire to ban all sensitive and harmful content.

Participants felt that some platforms they used featured more sensitive and harmful content (2/2)

Quotes from participants on seeing sensitive content…

There were platforms sometimes perceived as doing less to moderate sensitive and harmful content

"I think apps like X (formerly Twitter), TikTok, and sometimes Reddit tend to have more sensitive content floating around. Because people can post almost anything so quickly, it's easy for graphic or upsetting stuff to spread before it's taken down." (Male, 31, Caribbean, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – high)

"I see the most offensive content on Twitter (racist / anti immigration), and probably the most potentially sensitive on there as well as the content feels less controlled than on Facebook / Insta/ LinkedIn." (Female, 53, White, No disability, No religion, Cisgender, Straight, Online usage – light)

Platforms that featured video content came with risks of being overwhelming or intrusive

"TikTok is still fresh, therefore its sensitivity controls are most likely being developed as the days go along." (Female, 21, African, Disabled, Cisgender, Straight, Online usage – light)

I've noticed that platforms like Instagram and YouTube try to filter things more, but even there, you can still stumble across disturbing videos or comments." (Male, 31, Caribbean, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – high)

Platforms were generally seen as less risky when they were used for a narrower purpose

"I don't see sensitive content so much on Instagram or LinkedIn." (Female, 37, Mixed ethnicity, No disability, No religion, Cisgender, Straight, Online usage – light)

"I don't think there is anything to be concerned about on LinkedIn. I don't think it has any content that will cause distress to anybody." (Male, 38, Pakistani, No disability, Muslim, Cisgender, Straight, Online usage – light)

"I use [Reddit] occasionally, it serves a niche sometimes that nothing else can, the subcultures are great, the community vibe is great." (Female, 37, Mixed ethnicity, No disability, No religion, Cisgender, Straight, Online usage – light)

4 Attitudes towards features for controlling online experiences

4.1 Attitudes toward controlling online experiences

There is a sense that control over what participants see online has decreased, though exposure to a wide variety of content was sometimes beneficial

A recurring theme was a **perceived loss of control over what appeared in feeds on platforms.**

- Participants shared concerns about **seeing a lot of sensitive and harmful content they didn't want to see,** including content they aren't interested in, content they might find upsetting, and content that could have a more serious negative impact on them (e.g. content promoting eating disorders).

- **However,** there was also evidence that **participants enjoyed the unpredictable nature of encountering online content:**
  - People described enjoying seeing entertaining posts from users they didn't know, discovering new information and interests, and hearing from people with different perspectives on their social media feed.
  - Being exposed to a range of opinions and perspectives – this was both a cause of frustration when they felt other users were disrespectful/toxic, but nevertheless important to engage with diverse opinions/perspectives.

- **There was acknowledgement that it was important for there to be tools to help with managing and navigating content, but this didn't always translate into a desire to use these extensively.**

"It has become harder and harder to control the type of content I see, the algorithm constantly pushes stuff in my feed whether I want to watch it or not… now it's anything from anyone" (Female, 29, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – high)

"I am thinking of Instagram when I'm saying this, but I like that random people show up when I'm browsing on the Reels tab – that's how I discover (and share) new content with my friends!" (Female, 28, Asian, No disability, Hinduism, Cisgender, Straight, Online usage – medium)

Participants had mixed levels of experience with using tools to navigate online content, with limited awareness for some tools

Participants had **varying approaches to control and act** on what they saw online, in some cases taking this to consciously 'train' their algorithm.

Reactive and lighter-touch steps

Some participants, including a few heavier or more active users, appeared more familiar with basic, day-to-day tools for navigating content. However, this familiarity was inconsistent and highlights a broader awareness and effort gap. Actions included:

- **Blocking, muting, unfollowing, 'not interested', and reporting** – tools that often involved taking action against individual users or pieces of content.
- **Ignoring unwanted posts**, using likes/dislikes to influence algorithms, but generally avoiding advanced controls due to low perceived value or awareness.

"I saw the upsetting headline, but I didn't interact with it in any way… no blocking, no reporting, no lingering. I just kept scrolling to remove it from my view and continue browsing." *(Male, 31, Caribbean, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – high)*

Proactive steps including granular control

More advanced tools were generally unfamiliar to most. Even when people knew they existed, they often felt 'too much effort' to locate or set up, which limited regular use. Examples included:

- **Changing settings to see less 'sensitive' content and/or select topics of interest** – participants had mixed views about the success of these (see rest of section).
- **Algorithm resets, topic bans, and keyword muting** - used by some intermittently to update interests, or when topics became overwhelming, during stressful periods.
- **Safe Search (Reddit) and adult-content controls (Instagram)**, mainly to prevent children seeing inappropriate content on shared devices/accounts, rather than for their own viewing needs.
- **Subscriptions to block ads** and **share less personal data with advertisers** (e.g. YouTube Premium).

"Recommendation reset – This is essential when the algorithm starts reinforcing emotionally triggering or irrelevant patterns. It's a way to reclaim control and start fresh." *(Female, 41, Other ethnic group, Disabled, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – light)*

There were concerns that many tools were ineffective, which sometimes undermined interest in using them

While participants were supportive of having tools to control and act on the content they saw, there were many comments about poor experiences with these:

The majority felt that tools offered **short-term** help but failed to achieve lasting feed changes.

"It always ends up cycling the same content through regardless of how many times you click 'not interested'" (Male, 22, White, No disability, No religion, Online usage – medium)

Reporting was perceived as **ineffective**, with platforms rarely acting on flagged content.

"I report scam friend request… that rarely gets accepted though by Facebook" (Female, 51, White, Disabled, Roman Catholic, Cisgender, Straight, Online usage – high)

There were frustrations when previous efforts to change what content participants were seeing had to be **repeated**.

"Sometimes I feel everything reverts back to previous settings on its own." (Male, 38, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)

Content management (e.g. uninterested button) disturbed their casual scrolling, becoming **time-consuming**.

"It has become a time-consuming job of pressing the 'uninterested' button." (Female, 29, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – high)

Even after using tools to try and change their online experience, users could feel there was **little change in their experience**.

"There's no action from the platform." (Female, 26, White, Disability, No religion, Cisgender, Straight, Online usage – unknown)

Some users had workaround strategies for avoiding sensitive content (e.g., reducing time spent online, being selective about who they followed).

"I just expect to see stuff I don't want to, so I take that as a sign to switch off for the day." (Female, 35, White, Disabled, Presbyterian, Cisgender, Straight, Online usage - light)

4.2 Attitudes towards filtering out sensitive types of content

Participants said that filtering out sensitive content was useful for online safety and supporting people's wellbeing online

- The majority of participants could see **benefits to filtering content, by helping to create a safer online experience for an individual user,** if applied to more sensitive or harmful types of content.
- They could see in principle that the tool could **reduce exposure to hateful content, misinformation, and distressing content, acting as a layer of protection for users' emotional wellbeing.** Participants sometimes said they would be more likely **to use filters temporarily**, e.g. if they were having a difficult day and didn't want to come across upsetting content. This was **recognised especially** by participants from minority ethnic or religious groups, younger people, and those with mental health problems.
- It was also suggested though that filters may need to **differentiate between news content** and general internet users **posting their own harmful or sensitive content.**
- Among participants who showed interest in using filters, some said they would **use** them on **platforms they thought** were more **'toxic',** and others on platforms where they didn't communicate with other people (e.g. video-sharing platforms).

"I would use the filter because I don't have any tolerance for hate speech/content. If it were a video from a reputable news organisation that was reporting a news story containing homophobic abuse, it's important that the story is shown to viewers with a trigger warning. I think it's important to consider how homophobia, along with other forms of hate, can present in subtle, insidious ways, which may or may not be picked up by filters, meaning that hate comments or hateful user content may easily slip by."
(Male, 21, Asian, No disability, Buddhism, Cisgender, Gay, Online usage – light)"

*"I think I would for social media platforms that are less about interacting with individuals, so, like YouTube, Tik-Tok, Instagram. But on platforms like BlueSky I tend to trust that I'm able to curate my experience enough that any content like that is being shared as a call to action *against* it, or generally raising awareness. If I'm interacting with people I trust I'm less worried, but if I'm being exposed to people I don't know then I would have my guard up more."*
(Male, 35, White British, Disabled, No religion, Cisgender, Bisexual, Online usage – high)

There were concerns that platforms could not quickly or effectively find and hide content for a user wanting to use filters

Participants were concerned about *under-filtering,* meaning content that should be filtered out was not, making the tool ineffective.

- **Trust in platforms to enable effective filtering was also low**, given that participants felt that some content types that are supposed to be taken down are not. There were also some who had **mixed experiences with existing sensitive content filters**.
- **There were concerns about some platforms being biased and having political agendas,** which were seen to encourage and enable hateful viewpoints on the platform.
- There were also concerns about **users' ability to bypass filtering by using coded language and subtlety** in the content they post.

Many participants were concerned about whether comments sections would be tackled by content filters. They said current comment-filtering tools fall short, with platforms being slow to act and automated filters missing harmful comments.

- They noted that the most distressing content often appears in comment sections, seeing sexist, abusive, or hateful language, and felt using reporting tools did not help.
- Participants felt that stronger platform-level moderation, including human review, is needed for this issue, and felt current systems are too unreliable to protect users effectively.

"You can report or dislike comments on most platforms I guess - but those are often the places I see the most outrageous sexism and misogynist etc."
(Female, 49, African, No disability, No religion, Cisgender, Straight, Online usage – high)

"A lot of the actual posts on my feed are fine, but the comments can sometimes be upsetting. A way to filter certain words or types of comments would be helpful."
(Female, 27, White, No disability, Roman Catholic, Cisgender, Straight, Online usage – light)

There were even more concerns about the potential impact of not having the choice to access content participants wanted to see

As well as under-filtering, there were concerns about *over*-filtering, meaning filtering out more than needed, removing access to valuable informational content such as news, politics, and a diverse range of opinions.

- Some participants **did not want** to **alter their access to content** via filters, and said they would continue with other features instead.
- It was widely agreed that people **had different thresholds for what they considered sensitive content** and content that they would prefer not to see. Participants expressed concerns that a blanket option could remove content that they would otherwise want to see due to the presence of specific keywords.
- Similarly, there were also concerns about **the subjective nature of classifying some content types.** Participants acknowledged the **difficulty of recognising when debate was healthy and legitimate** (e.g. discussing societal issues respectfully vs. aggressive conversations with overt attacks on others). **There was limited trust in platforms' ability to interpret tone and context accurately.**
- Participants also feared that excessive use of filtering could impact their ability to stay informed with current affairs and be exposed to a diverse range of opinions. They were concerned that this could only reinforce existing beliefs, potentially leading to **'echo chambers' and polarisation**.

"The idea of filtering out and avoiding sensitive content depends on audience maturity and setting. I have a young son. I would absolutely apply filters to online content if he had access to online social media sites. Outside of work, at home or with friends, I would not apply such a filter. I am a strong believer in the power of education and knowledge... the amount of knowledge you can gain is limited by filters." *(Male, 39, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)*

"What is regarded as 'sensitive' material is subjective, and I would rather decide for myself and make my own mind up. I don't filter out content; I tend to begin reading it, and if it's extreme or, for the most part, just plain wrong, I (for the most part) just scroll on by and ignore it." *(Male, 58, White, Disabled, No religion, Cisgender, Gay, Online usage – light)*

"[in reference to the risks of filtering tools] You stay within an echo chamber and don't get exposed to other views different from your own, it can cut people off from convincing each other otherwise/ this could make people more extreme in their views and dehumanise 'the other' more" *(Non-binary, 31, White, Disabled, No religion, Lesbian, Online usage – medium)*

Discussion of filtering tools also prompted concerns about why some content was available on the platform at all

Participants had **concerns about whether content subject to filtering tools should be on platforms at all**. There were frustrations at perceived lack of action against some types of content such as racist content and hate towards religious groups, as well as major concerns about the wider presence of hateful discourse in UK society.

Related to this, some participants had experience of already seeing content they perceived as neutral being taken down and were concerned filters could enable this further, for example if the platform was seen as having a political agenda against a certain community.

This also prompted a worry that filtering out sensitive or potentially harmful content was **a way for platforms to ignore content that shouldn't be there,** and that if widely used, filters could mean this content went **unchallenged and unreported,** reducing accountability.

It was suggested that if content was permitted, then it could be important for it to be accessible. For example, if there was a news story about a public figure saying something offensive, then it could feel important for what they said, and other internet users' responses to this, to remain available and not filtered out.

"Filtering is being used as an excuse to continue harmful content being available online. It's a complicated time in society at the moment. There's a growing far right, and racism is much more widespread and acceptable. Whilst I would prefer not to see racist content (particularly where it's aimed at people like me!), to filter it out would mean I'm not aware of what's being shared and what I'm up against. My logic for not filtering is sort of a "know your enemy" logic." *(Female, 32, Indian, No disability, No religion, Cisgender, Straight, Online usage – light)*

"Visibility is important; without seeing hateful content, you're unable to react or report it. If we filter out hateful content towards disabled people, that may improve our overall social media experience, but the hate may still end up in the feeds of those most affected by it. Ideally, any filtering of hateful content should come from the social media platforms who have a legal responsibility to safeguard vulnerable people. Hateful content towards any marginalised/vulnerable population is unacceptable, and allowing those perpetuating such content to hide behind filters doesn't seem like the right answer." *(Female, 26, White, No disability, No religion, Cisgender, Straight)*

"Simply filtering it out would mean it goes unchallenged, you can't report it or engage."
(Female, 23, White, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight)

There were concerns about filters preventing people's ability to challenge hateful content or support people being targeted with it

Reactions were sometimes affected by *how* content filters were presented to participants:

Removal of content from experience

Participants sometimes wondered if there were **alternatives to removing content from their experience.** They sometimes preferred the idea of being warned about content instead, rather than losing access to it (see slide 53), or suggested that someone using the filter could still access some content (e.g. news).

"Sensitive" content vs. specific types

The idea **of filtering out 'sensitive' content** prompted **very different reactions compared** to the idea of filtering out more specific *types* of 'sensitive' content.

Participants **had more objections and concerns when less granular terminology was used.**

Personal investment in content type

Participants had **stronger views** when the type of content that could have been filtered out was a **topic area that was more relevant to them** – for example, the significance of racist content to someone from a minority ethnic group.

When asked about filtering *hateful* content, people from groups targeted by hate sometimes wanted to filter content out, given the potential for it to negatively affect them.

But more often, people said they wanted **to continue accessing the content** – sometimes *because* of their objections to it, with users expressing interest in:

- **Seeing what hateful things are being said about their community**
- **Challenging users saying hateful things**
- **Supporting users they shared an identity/community with, if they were being personally attacked**

When thinking about what the default settings should be for content filtering, even those who were more supportive of the tools were torn about what was a fair and effective approach

- Participants felt that the main benefit of the default settings for content filtering being on is that it would help to ensure a safer online experience for **all users**, particularly as they might not realise the option exists on the platform.
- However, the main drawback was that it would give users **less control** over **whether they want to restrict the content.**
- There was also recognition that users **might not necessarily revisit the settings** after signing up to the platform, mostly due to lack of awareness about what their options are or not knowing where to find them.
- Consequently, participants highlighted the importance of maximising the users' choice. This involves having the **choice** to select the relevant settings when signing up to the platform, with **clear explanation** about what these settings entail and **signposting** for how the settings can later be changed.

"It should be automatically be switched on because not everyone finds it easy to look for the settings and put on the restricted mode… at least it is by choice not force-feeding content you never wanted to see in the beginning." *(Female, 29, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – high)*

"I think the default should be that everyone sees everything unless they want to opt out, because we should show the range of views in the world whether we like them or not." *(Male, 37, White, No disability, No religion, Trans, Bisexual, Online usage – light)*

"It should be an option when you sign up, clearly displayed, with a list of things that would be hidden." *(Non-binary, 31, White, Disabled, No religion, Prefers to identify as queer, Online usage – high)*

4.3 Attitudes towards other features for controlling online experiences

During the research, participants were asked to provide feedback on a wider range of online tools and features

Online tools
Content filtering tools: A range of versions, each filtering out different kinds of sensitive and potentially harmful content.
Content preference tools: Topic Preferences and Recommendation Resets.
Global Blocking tools: 3 versions that limit which accounts the user sees content from.
Content overlays: warnings that require clicking to proceed.
Actions against individual users: blocking and muting.
Actions against individual pieces of content and content types: • Reporting • ‘See less of this’ • Community Notes • Muting words/hashtags

Content preference tools were seen as potentially useful when the users' feed contained too much content that was not relevant

Topic preferences

The tool to set topic preferences is seen as potentially helpful in highlighting preferred content and consequently **avoiding unwanted content**. However, it is also seen as restrictive. Users worry that this **will limit new content discovery and broad topic categories** may lead to irrelevant suggestions, while also being uncertain about the tool's effectiveness in protecting from sensitive content.

Recommendation resets

The tool is **helpful** in scenarios where users are seeing **mostly irrelevant content to their interests**, providing them with the option to 're-start' their algorithm. However, this is **not felt to always be effective**, with the **algorithm** potentially **reverting** back to providing irrelevant, as well as potentially sensitive and harmful content.

In general, participants felt that the content preference tools could **be useful for feed optimisation,** particularly if their feed is not relevant or personalised to their interests. There were also examples of people making existing use of these tools on platforms such as Pinterest to curate their feed.

However, there were **concerns about whether the tools would be effective** in helping to achieve this. They worried that may still be shown irrelevant suggestions or content that they would prefer to avoid.

Global blocking tools were generally seen as too restrictive, although useful in situations where users might prefer a highly curated online experience

Only seeing content posted by users you 'follow' or are 'friends' with

This **tool** was seen to be beneficial in scenarios where users wanted a highly curated **and safe experience** and only **engage with accounts they know and trust**. However, most felt that it would be too restrictive and limit the exposure to varied content outside of their own circle.

Filtering out potential spam

The tool is seen to be helpful for users to keep their feed focused on '**genuine' content only**. However, concerns around the subjective interpretation of spam were present.

Filtering out content from users without verified status

Participants **expressed distrust** for the 'verified' or 'trusted' status. This is driven by users **paying to obtain it,** and a general sense that 'verified' does not imply their views are impartial or correct[1].

Global blocking tools were seen as potentially useful in certain scenarios, such as when users want a highly curated online experience and only engage with accounts that are legitimate or that they trust.

However, the majority had concerns. Participants generally **did not want to limit their opportunities to discover new content**. They also felt that these tools would not prevent users from encountering potentially sensitive, hateful or harmful content, as these accounts can still post such content. **'Verified' status was distrusted** among the participants, some of it being linked to distrust towards platforms and how the status was being obtained.

1. See section 5 for more detailed feedback on filtering and verified status

Most participants considered content overlays valuable due to giving users more control over whether they want to view individual pieces of content

Content overlays

The content overlay tool that was explored was ‘warnings that require clicking to proceed’, which are shown before users are able to see the content that may be sensitive.

Some participants seemed familiar with overlays in the form of content being ‘blurred’ out on services like Reddit and TikTok.

Participants found **the tool highly valuable** in giving **users the control to decide if they wanted to view the content**. While some concerns existed around content **being inaccurately flagged,** it was generally felt to be effective at protecting users from sensitive and harmful content. Exceptions to this view included participants talking about blurs being ineffective at preventing them from seeing some sensitive content in their current online experiences.

Participants noted having a preference for **detailed descriptions** to help users make **informed choices.**

Many participants said they preferred the idea of having an effective warning about content, rather than it being invisible in their experience.

Participants valued the tools that allow to take actions against individual users, as they were felt to provide more granular control over their feed

Blocking individual users

Participants felt the tool to block individual users valuable in **protecting** their online space from content and interactions (e.g., messages, comments) that they **disliked or considered** to be sensitive and harmful, as well as in cases of being targeted online.

Muting individual users

Muting individual users was perceived to be a useful tool in situations **that did not warrant blocking** a user or an account, but whose content they would prefer to avoid.

Tools that allowed users to take actions against individual users were seen as valuable **in various scenarios**, depending on whether they entirely wanted to **stop any interactions** with the user or if they **simply did not want to see updates and posts from them.**

Participants saw value in using these tools, as they felt they are helpful in **limiting content** that they do not want to see, as well as **protecting** themselves from **sensitive, harmful or otherwise inappropriate content or direct harassment.**

While actions against individual pieces of content and content types were seen as valuable, participants still question their effectiveness

Reporting users and/or posts to the platform

The tool is considered **essential and important for keeping social media platforms safe**. However, whilst useful in principle, in practice it is often seen to be **ineffective**, with participants mentioning instances of sensitive, harmful or hateful content not being taken down.

Reporting users or posts and informing other users about content issues

The tool is **valued for helping to restrict sensitive and harmful content or misinformation** and utilising the community in flagging such content. However, concerns around **subjectivity and misuse** are high, alongside questions about the overall effectiveness of the tool in keeping users safe.

Choosing to 'see less of this' or muting individual words and hashtags

The tools that allow to click 'see less of this'/'not interested'/'hide'/'dislike' and to mute individual words and hashtags were seen as valuable. However, there was uncertainty about the tools' impact on algorithms. Users worried the algorithm may misinterpret preferences, and that the tool to mute words may be impractical, due to too many variations for certain words.

While tools that allow users to take actions against individual pieces of content and content types are seen to be valuable in principle, in practice participants express a range of concerns about their effectiveness.

Lack of trust in platforms is often expressed, due to content that is sensitive, harmful or hateful not being taken down even after being reported. There is also a lack of understanding about the ways in which algorithms work reduces the perceived usefulness of tools, such as clicking to 'see less of this' or muting individual words.

Participants make use of various approaches to tailoring their feeds and interactions. However, many want more customisable tools, alongside making existing tools easier to access

Current approaches taken to customisation

Outside of the tools discussed as part of the research, participants mentioned taking a range of approaches to help customise their feeds and interactions on social media or video sharing platforms:

- **Turning off comments for posts**, particularly if receiving hateful comments. However, there are concerns about silencing users when the option is on for high-profile accounts.
- **Avoiding engagement with content.** It was noted that *not* looking at content was often necessary in the moment to help the algorithm better personalise their feed.
- **Hiding their content from others** was felt to be helpful in avoiding hateful comments (e.g., some mentioned using the tool after experiencing cyberbullying).

Desired tools and features

Participants mentioned a range of options that they would find helpful if implemented by social media platforms:

- Making sure the **existing tools are easy to find.**
- Existing tools could be more **customisable** (e.g., certain topics could be set to contain a warning, rather than being removed from the feed). Alongside content overlays, adding more detailed **content description** to help users make informed decisions.
- A **feed split** (e.g., family/friends vs news) to have different types of content split across feeds, which could help to only engage with the content they prefer at any given time. This could help to temporarily avoid certain content if they only want to see updates from those they follow.
- A tool to **avoid some content for a set period of time** (e.g., when feeling 'low').

"On platforms like Instagram you can turn off comments. I think this is good if comments on a post are harmful and upsetting." *(Female, 20, White, Disabled, No religion, Cisgender, Bisexual, Online usage – light)*

"I would like to have a feed split - friends and family content; influencer content and news and current affairs content. That way, most of the sensitive content (hopefully) would be in the second two sections." *(Female, 41, Other ethnic group, Disabled, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – light)*

5 Attitudes to verified status and identity verification

5.1 Online Identity

Online identity on social media and related platforms was agreed to be a selective and filtered version of real-life identity

- Participants commonly felt that people's **profiles are not a true or full reflection of who users are,** and they personally presented only a part of their identity online.
- For the majority of participants, **an online persona** is often a **more curated version**, focusing on certain interests and traits. It is about being **selective rather than presenting false or misleading information.**
- **Online profiles** were often felt to be exaggerated and **untrue representations of real life,** and can put pressure on participants to live up to these expectations. For some platforms (e.g. to consume content rather than communicate), participants indicated they didn't necessarily include personal information in their user profile.
- **Online identities** were perceived to **differ depending on the platform**, the context of content/group they were part of or engaged with (i.e. online groups focused on a particular interest) and if the platform allowed for anonymous profiles.
- Having an **altered online identity was sometimes seen as the norm for compartmentalising aspects** of life, allowing people with an online presence freedom to build their distinct profile.
- There were a few comments that influencers, including those promoting products online, were often more authentic in the persona they display online due to the need to demonstrate business credibility and gain trust.
- Those with children also discussed how their online persona allowed them to be selective about **how they share family life** online, whilst protecting their children (e.g. not sharing photos of their children).
- There were **neurodivergent participants who** felt that their online identity **allowed them more freedom,** and they were able to express themselves more openly via their online persona.

"For me, my online identity is a bit more polished than my offline one. I tend to share the highlights, good photos, achievements, or fun moments, while in real life, I'm more casual and low-key. It's not fake, just a more filtered version of myself." *(Male, 31, Caribbean, No disability, Christian, Cisgender, Straight, Online usage – high)*

"I suppose people have come to accept that a great many of us have two selves now. Well, maybe three. Private, public IRL and online." *(Female, 20, White, Disabled, No religion, Cisgender, Bisexual, Online usage – light)*

"It's almost more acceptable online to show autistic traits online, I think. The 'real' world isn't geared up to neurodiversity in social settings most of the time." *(Gender fluid, 33, White and Black African, Disabled, Evangelical Christian, Straight, Online usage – light)*

Online identity was also associated with risks, including users being more able to post harmful content, fake profiles, and accounts being hacked

- There were concerns that online, **users felt more comfortable with voicing views that wouldn't be socially acceptable in real life,** for being obnoxious, rude or ignorant.

- Some felt **anonymous online profiles** can be used **to hide** behind for bullying, trolling or sharing extreme views – particularly on platforms where anonymity is common.

- **Participants felt growth of AI and bots** was making it increasingly difficult **to check online profiles** and distinguish between real and unreal identities. **Account hacking and cloning** were seen in the same light.

"It is near impossible to know if someone is who they say they are online too particularly with account hacking. Lots of people will set a profile up with their identity, only to have the content cloned and that profile used poorly." *(Gender fluid, 33, White and Black African, Disabled, Evangelical Christian, Straight, Online usage – light)*

"Some people present a totally perfect life, which can't be real - nobody has only good bits in their life. Other people can be aggressive and unpleasant online - feel free to say things they probably wouldn't say or own in real life." *(Female, 53, White, No disability, No religion, Cisgender, Straight, Online usage – light)*

Participants perceived that people present their online identities in different ways on different platforms

- **Facebook and Instagram**: Online profiles were perceived to be personal and clearly connected to real-life identity, but nevertheless highly curated.
- **X:** Participants felt that it's a platform where it is common for users to hide their personal identity using only a username. They said the platform had a reputation for users not acting as they would in real life, e.g. being ruder.
- **LinkedIn:** Participants noted that, given the platform's emphasis on professionalism and networking, users are inclined to share a version of themselves that is accurate but intentionally curated, especially since they may interact with professional contacts offline.
- **Discord, Reddit, Telegram**: As these platforms allow for greater anonymity and rely primarily on monikers and usernames, participants felt that online identities were less reflective of individuals' offline selves.
- **Online dating:** Participants widely believed that dating platforms encourage users to portray identities that more accurately align with their offline selves, largely because interactions have the potential to move into offline, face-to-face settings.

Finally, participants noted that **sometimes there was no need for someone to include information about their identity**, e.g. on a platform where they mainly consumed content, and didn't interact with other users, or on gaming platforms where genuine identity was not seen to be as important as on other services.

"Different platforms are for different things,…What you say on Facebook is more connected to you in real life than what you post on Bluesky or Discord." *(Male, 35, White, Disabled, No religion, Non-binary, Bisexual, Online usage – high)*

"On Facebook, I'm more myself and more likely to share more information because it's targeting people who know me." *(Female, 58, White, No disability, Christian, Cisgender, Straight, Online usage – medium)*

"On LinkedIn, it's about professional authenticity." *(Male, 31, Caribbean, No disability, Christian, Cisgender, Straight, Online usage – high)*

"I think this may vary with different types of platforms. If for example on a dating site, I'll expect people to be themselves." *(Male, 28, Caribbean, No disability, Islam, Cisgender, Straight, Online usage – high)*

5.2 Verified Status

Participants were familiar with verification status, which they said helped them trust in the authenticity of an online profile

Blue tick verification was widely known and associated with influencers, celebrities, public figures, and brands.

- A verified status was **associated with higher trust** that the user is who they say they are – e.g., "official" or approved account for celebrities and brands.
- The majority of participants saw verification as broadly positive, **adding legitimacy and increasing engagement** with content on some platforms like Facebook and Instagram.

Participants were **more inclined to feel positively** about verified status **for platforms with a narrower purpose for example, dating apps,** mentioned in particular by female participants.

- This was due to their status as apps that have the specific function of potentially **facilitating in-person meetings of people they potentially haven't met.**
- On **LinkedIn,** verification was associated with professionalism and official business accounts, and seen as relevant for job-seeking, e.g. so that users know that the job advert is legitimate.
- Identity verification enhanced perceptions of safety on dating apps in particular and was felt to be useful to minimise the potential for deceptive practices like catfishing and scamming.

"I tend to trust people more if they have verified status and therefore tend to engage more i.e. I like their posts more"
(Female, 28, White, No disability, No religion, Cisgender, Bisexual, Online usage – medium)

"If I'm looking at news… I look for the blue tick just to make sure I'm following the real person — for example, Martin Lewis." *(Gender fluid, 33, White and Black African, Disabled, Evangelical, Straight, Online usage – light)*

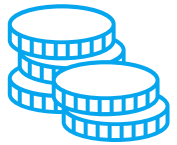
"Dating sites are more likely to result in face-to-face meetings so safety through verification is very important" *(Male, 21, Chinese, No disability, Buddhism, Cisgender, Straight, Online usage – light)*

"For dating, it's about personal safety and maintaining trust and reliability" *(Female, 32, Indian, No disability, Cisgender, Straight, Online usage – light)*

"On professional networking apps, it's very important for trust and credibility" *(Male, 31, Caribbean, No disability, Christian, Cisgender, Straight, Online usage – high).*

However, being able to purchase verified status also eroded trust in this feature

- Verification was perceived to have become **less trustworthy in recent years**.
- **Verification obtained via payment** (e.g. on X) had undermined trust in verified status; as the status was seen as being gained by payment rather than through a more legitimate process.
- The majority of participants are **uncertain about how verified status is obtained** and how this differs between platforms, e.g. payment (X), number of followers and content type (Instagram/Facebook). It was unclear to participants what, if any, documentation is required.
- Verification **did not always guarantee authenticity in the participants' eyes.** There was an assumption that inauthentic accounts may 'slip through the net'.



"[in reference to X's pay for verification system] It allowed scammers to pass themselves off as verified users and it meant anyone without a tick would not get seen as often as before"
(Male, 41, White, Disabled, No religion, Cisgender, Straight, Online usage – light)

"It's no longer a verification process, it's a prestige symbol. If you can buy a blue tick then it's not an "independent" thing, it's a way to make money."
(Male, 35, White, No disability, No religion, Cisgender, Bisexual, Online usage – high)

"I would not want at all to be verified, already these sites know so much about us and some even sell our data, getting verified just gives them even more information, it doesn't feel right" (Female, 29, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – high)

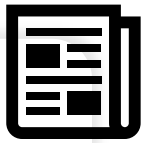
Verified status became more valuable in the context of health and news-related content

- Across groups, participants wanted credible information. They felt that **status verification could be a driver of creating trust** in content, as the poster could be identified and the user could determine for themselves whether the poster is credible. This belief varied depending on the type of content:

Health – high importance

The consequences of disinformation and unreliable health sources meant that health was perceived to be the top content area where verification could reassure.

"I'd want to know that I'm hearing from an organisation or expert…there's been so much misinformation about health, the rise of anti-vaxing and denialism" (Male, 35, White, Disabled, No religion, Non-binary, Bisexual, Online usage – high)



News/political content – high importance

Due to lack of trust in politics and news sources, verification could signify a trusted and checked source of news and information.

"If I'm looking at the news, or for factual information, then I will pay a bit more attention to the verification" (Gender fluid, 33, White and Black African, Disability, Evangelical Christian, Gender fluid, Straight, Online usage – light)

Holidays, travel – low importance

Travel and holiday content verification was seen as less important. Photographic posts were a way of determining trust in the post. While review sites were seen to be more valuable than verification to check information.

"Holidays– not much difference, I trust photos or reviews more" (Male, 31, Caribbean, No disability, Christian, Cisgender, Straight, Online usage – high)

Participants were asked to reflect on how they would feel if verification was available to all internet users for free

- The majority felt that **more readily available verification** would **make little difference** to themselves personally.
- They expressed concerns that it **could make it harder to differentiate between more and less legitimate** and trustworthy user profiles under existing approaches to verification status - for example, that someone who they previously wouldn't have trusted might be given a 'verified' label.
- The majority were **unclear of the benefits of having more verified users** amongst the general population and needed convincing as to why it is needed.
- Verified status was **associated with influencers, celebrities and brands, used to grow followings** – rather than casual users. This association **created a lack of interest** and perceived benefit for casual users, unsure why/how widespread verification would benefit them, as they mainly used it for light entertainment and following people they know.
- Participants were suspicious about whether a rollout of verification would be designed to **benefit social media companies** to get more personal data from their users. This was due to the **lack of transparency and trust in social media companies current operations** which made participants sceptical.
- Only a minority of participants saw verification as a potential online safety solution, and thought it could help them identify authentic profiles and avoid fake profiles, who they associated with potentially malicious activities (e.g. fraud).

"It already means very little, so I guess it would mean even less?" (Non-binary, 31, White, No disability, No religion, Prefers to identify as queer, gay or lesbian, Online usage – light)

"It would worry me - we see why famous people need it but to have regular users verified i.e. the masses. I would wonder what they are doing with our personal data" (Female, 54, Mixed ethnicity, Disabled, Islam, Cisgender, Straight, Online usage – medium)

"I do not trust social media platforms enough to keep this information secure." (Female, 28, White, Religion- Other, No disability, No religion, Cisgender, Bisexual, Online usage – medium)

"I'd be somewhat likely to verify myself on platforms I use often, like Instagram or LinkedIn, because it would make my account feel more trustworthy and secure." (Male, 31, Caribbean, No disability, Christian, Cisgender, Straight, Online usage – high)

There are mixed views about what information would be acceptable to share in order to verify

Participants wanted the **process to be clear, transparent and simple,** with communication about**:**

- What information would be needed, how would it be held/shared, what it would be used for
- Reassurances that data is not being used for other purposes by online platforms
- Simple and secure ways to verify information e.g. using MFA

Most acceptable to share

- Most would be **willing to prove** that they are **not a 'bot'** (e.g. using Captcha style test)
- **Broad location** (e.g. city and region)
- **Full name**
- **Job title** – if a professional /job-related website, otherwise less relevant
- **Sharing an image of themselves** (e.g. on dating sites to match the profile picture)

Divided opinion

- Biometric data **felt familiar** to many, as they already use it on banking or mobile apps.
- **Some** felt biometrics (e.g., a photo of their face) were **safer and easier** to share than home addresses or official documents, seeing them as lower-risk.
- Others drew a firm line, unwilling to share biometrics with social media companies for being **too invasive and personal.**

Least acceptable to share

- For the majority, sharing details such as **home address, or ID such as a driver's license or passport** was less appealing, and they would like reassurances on how it would be used, stored and accessed.
- For the majority, **sharing official ID was a step beyond what is expected** for using platforms.

"The general "prove you're not a robot" Captcha things are fine, the rest, absolutely not!" (Non-binary, 26, White, Disabled, Religion – other, Bisexual, Online usage – medium)

"Not home address I would never provide my biometric info, job title or official ID. It can be done via MFA, ideally via mobile phone." (Male, 39, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)

5.3 Filtering users without verified status

There were concerns that filtering out users without verified status overly limits access to content

Participants were asked to share views on the idea of an online tool enabling them to only see content from users with verified status. We explored participant attitudes towards this both without stating how verified status was obtained, as well as asking for their views if verified status involved users verifying their identity.

There were various negative reactions to filtering out users without verified status:

- The main concern expressed across the group was missing out on interesting, engaging and informative content by users without a verification symbol, including friends and family (who could be unverified).
- There were concerns that this would create a '**two-tier' system** between those who are verified and those who are not. A number of participants were **concerned about 'censorship' with verified accounts being promoted over others in feeds** and the impact on their freedom online. There was a perception that people might have to hand over personal information they didn't want to share to ensure they were still visible online, that could result in people being 'left out'.
- Women and minority groups were more likely to be concerned, fearing that this may **limit minority views** and **restrict a range of perspectives.**

Some lighter social media users were more positive about the scenario:

- They felt reassured that this feature might result in less misleading or harmful content, as verified status came with a sense of authority.
- It was also mentioned by lighter users that **providing ID has become the norm (e.g. online banking)** in other places and, therefore, not necessarily an issue for social media sites to adopt the practice.

"The risk would be that we would most likely have one narrative or one view on social media which can be justified by a verified status." *(Female, 21, African, Disabled, Cisgender, Straight, Online usage – light)*

"Plenty of people will lose their data, give over information they shouldn't, think that the internet is safer than it is, and may end up as part of a two-tier system of visibility online." *(Gender fluid, 33, White and Black African, Disability, Evangelical Christian, Straight, Online usage – light)*

"I'd rather have things verified than receive a lot of false or harmful content" *(Female, 32, Indian, No disability, Cisgender, Straight, Online usage – light)*

There was a little interest in seeing less from users without verified status, compared to not seeing them at all

The participants were presented with other options for how filtering could work:

a) **Seeing less content in general from users without verified status** (e.g. in recommendations, search results) but they can still see your content or interact with you.
   - **Seeing less content seen by some as beneficial** to the browsing experience, making recommendations and results less broad, and reducing interactions with content from more 'random' accounts.

b) **Muting all users without verified status so you don't see their content at all**, but they can still see your content or interact with you.
   - Very **unappealing** for the majority, who were unable to see the benefit and felt this was too restrictive.

c) **Blocking all users without verified status from seeing your content or interacting with you** – for example, commenting on your posts or sending you DMs.
   - **Too extreme for most** but could be useful if the user was experiencing online abuse, e.g. trolling.

- Participants could see potential benefits of filtering out some accounts e.g. bots, fake or abusive accounts, but most wanted to do this on a **manual or case-by-case basis.**
- **Manually blocking accounts** was felt **to be sufficient** without applying a blanket filter to accounts without verified status, which could restrict access to interesting content.

"I can see the appeal as to why people might use them (filtering out fake profiles and bots) but I don't see the need for me personally to do this , and it feels too restrictive." (Female, 28, White, No disability, No religion, Cisgender, Bisexual, Online usage – medium)

"It would feel like the entire world wasn't real anymore. There would be no contact with a group of people purely because they don't want to give their ID over [in reference to C]. Absolutely horrific idea." (Gender fluid, 33, White and Black African, Disabled, Evangelical Christian, Straight, Online usage – light)

"A) seems like a good way to try and limit the random accounts you come across to at least ones who are verified / more likely to be genuine. C) could be useful if you are somebody who posts a lot and has a lot of problems with trolls." (Female, 53, White, No disability, No religion, Cisgender, Straight, Online usage – light)

Filtering out users without verified users felt more relevant for victims of anonymous bullying, than for everyday internet use

Participants were presented with the following scenario and asked how that impacted their view of this feature:

"Imagine that one of your friends or relatives unexpectedly began receiving a lot of abusive comments online and they are looking for ways to stop seeing these. They've told you that some of the comments are coming from people that seem to be anonymous, as there aren't 'real' names on their profiles."

- **Filtering was seen as more relevant in the context of online bullying and harassment,** as a measure to reduce harm from unpleasant behaviour online. Participants thought that losing out on content was acceptable in this context.
- Filtering was seen as more useful in extreme situations, for those experiencing significant harm (e.g. constant harassment from multiple anonymous users).
- However, **there was scepticism about the effectiveness of this,** with concerns that bullies would verify their accounts and continue in their bad behaviour.
- Participants felt that **platforms should prevent bullies from being able to act**, rather than providing retrospective measures.

"If this was a world where everyone was verified, then yes restricted those who are not verified would help if these were bots or not 'real accounts' but in the current social media I don't think any of these would help."
(Female, 20, White, Disabled, No religion, Cisgender, Bisexual, Online usage – light)

"I don't think these options are the solution. You can't blanket approach many for the few."
(Female, 41, White, No disability, No religion, Cisgender, Straight, Online usage – light)

Participants' key concerns for how verified status should work related to transparency, security and accessibility

Transparency

- Participants felt that platforms should acknowledge that users are **starting from a position of low trust in social media companies** and their power.
- Participants **feared that verification meant giving unnecessary information** that would be used for the company's benefit rather than the user's benefit. Participants wanted transparency; explained in clear language what information would be needed, why it is needed and how it will be used.
- **To overcome scepticism**, participants thought that platforms should give users clear reasons why verification could benefit them, **focusing on safety, fewer bots and less harmful content** impacting themselves and their family/friends.
- Participants thought that platforms should frame **verification around safety rather than restricting access** to content or making content more uniform and prescriptive.

Security

- **Participants trusted independent companies and processes** to handle verification, such as MFA.
- **Participants wanted to see** a commit to not sharing data with third parties, minimise storage and delete verified data once used.

Accessibility

- Participants wanted alternatives to **using official documentation, such as passports and driving licences,** which they felt **could create a 'two-tier' system** limiting access for those without these official documents.

6 Appendix:

Other feedback on features for navigating online content

Insights within this section pertain to when participants were asked their opinions on specific filtering methods and tools and within specific contexts. Where appropriate, feedback from other parts of the research has been incorporated.

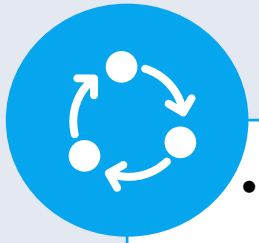
Note on language: *Within this section, where the methods and scenarios refer specifically to 'sensitive' content, in findings, it is referred to as 'sensitive and harmful content'. If participants specifically mentioned 'hateful' content (e.g., racist, sexist), 'hateful' phrasing is utilised instead to draw distinction.*

During the research, participants were asked to provide feedback on a wide range of online tools

Online tools
Content filtering tools: A range of versions, each filtering out different kids of sensitive and potentially harmful content.
Content preference tools: Topic Preferences and Recommendation Resets.
Global Blocking tools: 3 versions that limit which accounts the user sees content from.
Content overlays
Actions against individual users: Blocking and muting.
Actions against individual pieces of content and content types: • Reporting • ‘See less of this’ • Community Notes • Muting words/hashtags

6.1 Content filtering tools

Participants were asked for their views on the following content filtering tools...



Versions of tools shown to participants during the online community:

- **Version 1 – Filtering sensitive content:**
  Described to participants as 'Choosing to "Hide sensitive content" from your online experience/browsing'
- **Version 2 – Filtering potentially hateful content:**
  Described to participants as 'An option to avoid seeing content flagged as 'potentially hateful or discriminatory'
- **Version 3 – Filtering content not suitable for children:**
  Described to participants as 'An option to 'only see content suitable for under 18s', even if you are over 18.
- **Version 4 – Filtering content that is hateful to the user's identity:**
  Presented to participants in a range of different ways depending on their identity – e.g. sexist, homophobic, racist, transphobic, ableist
  (Note: for participants without an identity commonly targeted in hate, we asked about filtering hateful content).

The option to hide sensitive content had numerous benefits, but many were concerned about its effectiveness

Feedback on Version 1 – Filtering sensitive content

Benefits

- The method was seen as helpful for users who want to **avoid <u>any</u> type of sensitive and harmful content,** giving them confidence to use social media without fear of encountering unwanted content, and without affecting what others see online.
- Participants **valued the control** it provides for ensuring a **safe** browsing experience, noting that it could be especially useful for **protecting children** or **vulnerable users** (e.g. those in recovery, experiencing trauma, or poor mental health).
- They liked that the method covers **a wide range of content,** without needing to do a granular level filtering, making it **easy to use**.
- There was an expectation that it would be **flexible and simple to turn on/off** based on users' needs (e.g. enabling it temporarily to safeguard mental health).

"This allows you to use in-app settings to restrict what content you can see, it's good because it minimises the chances of bumping into sensitive content unexpectedly. It allows for filtration of what you don't want." *(Female, 29, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – high)*

Drawbacks

- Participants expressed a **lack of trust** in the implementation of the method, as it relies on the platforms' definitions of sensitive and harmful content. This raised **concerns about censorship, bias** and informative content falling under this definition (e.g., news or politically controversial content), potentially leading to ignorance about important topics. Participants from minority gender identity and sexual orientation in particular highlighted examples of informative LGBTQ+ content often being flagged as sensitive when it was not. The importance of **transparency** was therefore often highlighted, as participants would want to know why content is being flagged and what type of sensitive material it contains.
- Similarly, there was a concern about the method being **unreliable**, with content 'slipping through' the filters (e.g., if it was missed by the platform moderators, or if users utilised methods to avoid their content being flagged).
- Others noted that the definition of 'sensitive' is **subjective,** and people may have different views over what topics they consider sensitive and harmful.
- A few disliked that the tool **does not address the 'root issue',** i.e., the presence of harmful content on social media platforms.

"It might hide content users actually want to see if the system flags it incorrectly, like educational or news-related material that's labelled as sensitive. It could also create a slightly filtered or "incomplete" experience, where users miss important discussions or context." *(Male,31, Caribbean, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – high)*

While this method offered a safer online experience, the possibility of missing out on valuable content reduced participants' interest

Interest in usage – Filtering sensitive content

More interested in using

- Those who would be likely to hide sensitive and harmful content would do so to make their online experience **more enjoyable** and **less stressful**. The views were mixed across groups; however, most Muslim participants mentioned they would be likely to use it to avoid disturbing content, with a few ethnic minority participants noting that they would specifically do so to avoid hateful and racist content.
- A minority would use the option **when accessing social media around their children.**
- Others mentioned they might use it if it could be **easily switched on/off** to allow them to 'take a break', particularly if the content got too overwhelming or was affecting their mental health.
- A few would consider using it if they had a **clear understanding of what 'sensitive' included**.
- Those who would use it would likely **enable it on apps that they access most frequently, or apps they perceive as having more sensitive** and harmful content in general.

"It makes a massive difference because it allows me to control what I don't want to see which protects my wellbeing and makes my online experience more comfortable and enjoyable."
(Female, 28, White, Disabled, No religion, Cisgender, Bisexual, Online usage – medium)

Less interested in using

- Participants who would be unlikely to use the option wanted to **avoid missing important information**, and expressed concerns about informative posts, news or activist content being flagged as sensitive. There was a preference to gain a full picture of the content available online, **even if it meant exposing themselves to potentially upsetting and hateful posts** (e.g., racist or xenophobic). The view was expressed across groups, including ethnic minority participants.
- A few expressed a preference to **define what sensitive content to avoid** over a 'blanket' option.
- Men and participants from non-minority groups noted being less likely to use it due to **concerns about censoring**, or because they did not feel personally affected by sensitive and harmful content, preferring instead to **manually block** concerning content.

"I don't trust the social media sites to curate content the way I would like it to be… I wouldn't use it, because I don't trust how it would be, or is, implemented."
(Male, 35, White, Disabled, No religion, Cisgender, Bisexual, Online usage – high)

While hiding potentially hateful content is widely seen as beneficial, some are concerned about platforms' definitions

Feedback on Version 2 – Filtering potentially hateful content

- In this scenario, as the content is flagged in the context of being potentially hateful and discriminatory, this is seen as mostly **beneficial in helping to create a more comfortable online environment**. Participants value that it gives a clear idea about the type of content they will not see, centering around content that has the potential to upset.
- The benefits of the option were **particularly mentioned among ethnic minority participants** who said they would most likely use the feature on platforms known for hateful (e.g. racist) content. Participants felt that the feature could be beneficial for X, Instagram, TikTok and YouTube.
- However, some expressed **distrust for the platforms' definition** of hateful or discriminatory content, as well as how algorithm would distinguish it from informative material, fearing the setting may lead to **missing out on valuable information**.

Default settings

- Having the option switched on by default when signing up to a platform to prevent exposure to hateful or discriminatory content was seen to be **potentially beneficial** in some cases, particularly for **vulnerable** users or those under the age of 18.
- However, there was a general sense that users must retain **full control**, expressing **distrust** in platforms due to potential **political or personal bias** in determining what is 'hateful or discriminatory'. This was particularly concern among participants with a minority gender or sexual identity.

"I think this is a brilliant idea that will make my online experience more enjoyable… this scenario would affect my online experiences positively. It gives me power over what I do online and makes me feel like I am in charge of my online activity."
(Female, 29, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – high)

"I believe this should be up to the individual to set the settings as I don't always trust the social media apps to rightfully pick what is hateful or discriminatory."
(Female, 31, White, No disability, Religion – other, Cisgender, Straight, Online usage – medium)

Only seeing content suitable for under 18s is generally seen as too limiting for adults' online experience

Feeback on Version 3 - Filtering content not suitable for children

- In general, participants recognised that only seeing content suitable for under 18s could create a **safer and more emotionally comfortable online space**.
- For a minority, the setting would be **useful in specific scenarios** (e.g. during shared screen time with children and family). As a result, they would find useful being able to turn the setting on and off easily.
- However, the majority of participants across groups felt restricting themselves to under-18 content would significantly **reduce the relevance** of their online experience.
- They worried that such an option would potentially impacting users' **awareness of topical issues** and **hide educational discussion, world issues,** as well as **body-related topics** (e.g., pregnancy, body positivity) **and other mature themes** that help adults stay informed about world affairs that could be deemed as inappropriate for an under-18 audience.

"It would also limit exposure to some news, educational content, or discussions intended for adults, so it would be more useful in specific situations, like when I just want a "cleaner" feed for casual browsing."
(Male, 31, Caribbean, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – high)

"I would feel like I'm only seeing half of reality … There are some things of a sensitive nature, it's important to see as an adult in order to better understand people, society and contexts."
(Female, 32, Indian, No disability, No religion, Cisgender, Straight, Online usage – light)

Feedback on filtering out sexist and misogynistic content from female participants (Version 4)

- **Female participants** were concerned about the rise of **online misogynistic abuse, particularly appearing in comment sections**.
- **There were concerns that filtering might hide rather than address the issue,** and a preference to keep such content visible so it can be challenged and to avoid being in a “social media bubble”.
- They **favoured targeted keyword filters** such as ‘misogynistic’ or ‘racist’ over blanket bans, due to concerns about **accuracy and coded language.**
- Participants questioned whether filters would apply to **comments**, where they witness most misogyny, racism and abuse occurring.
- There was some interest in using content filters to avoid the risk of receiving sexist abuse via individual direct messages.

> *“I think social media websites like TikTok and Instagram have over-applied their sensitivity filters to the point that content creators have to use words like 'unalive' and 'grape' so they won't get flagged. I think these words are stupid because they over-euphemise the word to the point that they lose their effect and are said so casually.”*
> *(Female, 19, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – medium)*

> *“I might use a technology to completely avoid misogynist content online, but I'd probably toggle it on and off e.g., for days when I'm already stressed and don't want to see it, I'd happily toggle that on for Twitter/X especially.”*
> *(Female, 37, Mixed ethnicity, No disability, No religion, Cisgender, Straight, Online usage – light)*

Feedback from LGBTQ+ participants on filtering out transphobia and hateful content about non-conforming gender identities (Version 4)

- Participants with **minority gender identities** saw some value in filtering content, though sometimes preferred the idea of alerting users to such content (e.g. blurring abusive comments from strangers).
- There were strong concerns about how platforms would decide what to filter out and whether they could be focusing on stronger moderation. Participants referenced cases where non-harmful or educational **LGBTQ+ content was restricted** when using similar filters, which limited access to vital community information and support, **raising fears of censorship of LGBTQ+ voices**.
- There was also acknowledgement that there are diverse views about how to define, for example, what is transphobic.
- They felt it was important to retain visibility of hateful content towards the LGBTQ+ community to **identify those who need support** and remain aware of hostility to them.

> *"I think my answer would be dependent on whether such a filter would apply to news sources, but also if it applies to trans-friendly users/pages who discuss issues of transphobia. As myself and other users have discussed earlier, it is important that as a community we remain informed on such issues so that we can provide support to each other. So if this filter would *not* apply to factual content as I've mentioned above, then yes, I would use it. "*
> *(Non-binary, 26, White, Disabled, Religion – other, Bisexual, Online usage – medium)*

> *"I'm not a big fan of tools that just blanket block 'sensitive content' because that leaves a lot of the decision-making as to what is considered sensitive to the company running the social media platform, and it can result in them hiding more things than you might have expected. For example, LGBTQ+ content of all kinds often gets hidden when sensitive content filters are turned on, whether or not it's sexual content. "*
> *(Male, 35, White, Disabled, No religion, Cisgender, Bisexual, Online usage – high)*

Feedback from LGBTQ+ participants on filtering out homophobic content (Version 4)

- **Participants** were generally hesitant about how to implement filters for homophobic hate.
- They worried filters might **suppress LGBTQ+ voices, as well as important discussions and support.** They noted that subtle hate often bypasses filters, making reporting and blocking individuals more effective.
- Those who were more interested in filtering out content talked about wanting to avoid unnecessarily upsetting content, while continuing to stay informed.

> *"I am in support of an age filter for filtering out this type of hate speech but am doubtful about its implementation. I find it likely that this filter would just censor all that is labelled as "gay" entirely, thereby erasing gay representation… I do not think algorithms can sort this as an algorithm would pull from a database that, due to the societal nature of homophobia, would skew towards homophobic bias."* *(Male, 21, Chinese, No disability, No religion, Cisgender, Gay, Online usage – high)*

> *"Instagram and Facebook comments can be brutal against gay people. I have not tried to see if filtering out certain words works on Instagram for example. I would wish to filter this type of hate speech, for example if I was not having a good day. I want to be able to have more control over seeing this type of content. I cannot think of a scenario where I would not wish to filter out homophobic content."*
> *(Male, 45, Caribbean, Disabled, No religion, Cisgender, Gay, Online usage – light)*

Feedback from participants from minority ethnic groups on filtering out racist content (Version 4)

- There were mixed views on filtering out hateful and racist content. Participants wanted to see “the good, the bad, and the ugly” **to stay informed** and avoid shutting down constructive dialogue.
- Some participants talked about being more likely to use a filter on platforms where discussion between users was more intense and might include racism.
- There were concerns that filtering content would mean not being able to report it to the platform, with a desire for platforms to do more to take down racist content.
- As with other groups, there was also acknowledgement that views on what constitutes racist content can vary.

“Knowledge of the world includes knowledge of the good and the bad. If there are people posting on social media with extreme or racist views, we can either filter them out, and they will continue as they are, or we can attempt to engage in dialogue in the hopes of changing their views for the better. If we filter out violent scenes of protests or the consequences of war, we lose the opportunity to educate ourselves and others.”

(Male, 39, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)

“I think my choice might also depend on the platform. For example, on LinkedIn or Pinterest, I’d be less likely to encounter overtly racist posts in my feed, so I might not feel the need to filter. On platforms like X or Reddit, where discussions can be more heated, I’d probably be more likely to use the filter regularly.”

(Male, 31, Caribbean, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – high)

Feedback from disabled participants on filtering out content that's ableist or hateful about disabled people (Version 4)

- Participants were concerned that filtering out hateful and ableist content could unintentionally remove educational or advocacy related content, and limit participants' ability to report hate.
- Participants with disabilities often said they wanted to engage **with ableist or hateful content** in order to **challenge hate and support victims.**
- If they were being personally targeted, there was recognition of a potential need to filter this out, but also a concern about how to report such abuse if you couldn't see it**.**
- Many doubted filters' ability to catch offensive slurs and **emphasised reporting as a more effective approach** so those posting this kind of hateful content would be held **accountable** and hopefully bring about long-term change.

> *"As much as I despise seeing ableist or hateful content, it is sometimes necessary to have it available so you can fight back against it. Reporting hate comments, fighting back, reporting the users themselves, etc. While it would be great if every single person against the hate would just ignore it all and it would go away, but this would only be in an ideal world."*
> *(Female, 24, White, Disabled, No religion, Cisgender, Straight, Online usage – light)*

> *"I think if I was receiving some form of targeted ableist abuse/spam from online users that would be an appropriate time to filter this out. The reasons you might not choose to do so is to support complaints/reporting of responsible individuals… ultimately there is a need to know if people are engaging in this sort of behaviour, particularly in a public forum. I assume the impact on news would be minimal (at least for established media) which have rules to adhere to and are likely to report on such situations sensitively, but I imagine it would more widely affect user generated content and direct messages, particularly those from online trolls, or people who don't see themselves as being constrained online and think they can hide behind an online persona."*
> *(Female, 35, White, Disabled, Presbyterian / Church of Scotland, Cisgender, Straight, Online usage – light)*

6.2 Content preference tools

Participants were asked for their views on the following tools for changing what content users see...

Content preference tools and explanations shown to participants during the online community:

- **Topic preferences:** Choosing which topic areas you want to see content about (e.g. “home feed tuner” on Pinterest, “manage topics” on TikTok)
- **Recommendation resets:** Asking the platform to 'reset' what content is recommended to you. This would mean the algorithm is 'reset' and doesn't suggest things based on what you have looked at in the past.

‘Topic preferences’ offered helpful personalisation, but were seen as potentially limiting

Topic preferences: choosing which topic areas you want to see content about.

- Participants generally saw the content preference tool as **useful for controlling and personalising their feeds**, helping to avoid irrelevant content that does not interest them or sensitive and harmful content that they find upsetting, therefore improving the online experience.
- However, some preferred tools that would allow them to **exclude** sensitive and harmful topics instead.
- Concerns about the tool **included overly broad topic categories**, making **algorithms less effective** at suggesting relevant content.
- The tool was considered **most beneficial on platforms like Pinterest or Reddit,** where content aligns closely with interests. While in principle the idea was perceived positively, participants were **uncertain about using it on broader range of platforms,** such as Instagram or TikTok, fearing missed opportunities to discover new or important content.
- Some participants reported previously using the tool on Pinterest, TikTok, Reddit, and Snapchat, with **some ethnic minority** users **applying it** on Facebook and X **to avoid** sensitive, hateful and **racist content**. Similarly, some noted using it when setting up their account to better personalise their feeds.

“I do use this on Pinterest. I think this method gives full control of what content I wish to see and enable me to filter out content that I find sensitive. The drawback of this tool is that I am limiting myself and may miss content that may be of interest to me.”
(Male, 45, Caribbean, Disabled, No religion, Gay, Online usage – light)

“I may use this, but I would prefer to be exposed to different things and decide if I am interested rather than stick with what I already know I like to see.”
(Gender fluid, 33, White and Black African, Disabled, Evangelical Christian, Straight, Online usage – light)

‘Recommendation reset’ was seen as helpful in specific scenarios

Recommendation reset: asking the platform to ‘reset’ what content is recommended to you.

- The **tool was seen as potentially valuable,** with opinions being split based on how satisfied participants were with their feeds. For instance, participants felt it would be **useful** if their feed became overly focused on a particular topic that was **not relevant** to them or that did not offer diverse perspectives, as well as if it brought up too much content that they **disliked** or found sensitive and harmful, giving them the option to restart the algorithm.
- However, the main downside was felt to be **the new algorithm not being relevant to them**, alongside potentially losing some personalised suggestions that they enjoyed or found useful.
- Some noted that despite their current algorithms being mostly personalised to their interests, they **still came across sensitive and harmful content**. As a result, they feel a reset would not reduce the chances of encountering such content in the future.
- Only a few participants mentioned using **the tool when repeatedly coming across similar content** that they no longer wanted to see.
- Within this minority, some used it either because they were **not interested in the content** that was being shown to them or because they frequently came across **sensitive and harmful content,** for example, relating to war.

“Likely if the content in my feed becomes too focused on one topic and does not have a wide variety of creators.”
(Male, 22, African, No disability, Pentecostal, Cisgender, Straight, Online usage – light)

“Could be useful if you've accidentally clicked on unsavory content previously and got recommendations, though as it's a complete reset you will lose anything you want to see too.”
(Non-binary, 26, Chinese, No disability, No religion, Sexuality – prefer to use another term, Online usage – light)

6.3 Global Blocking Tools

Participants were asked for their views on the following examples of global blocking tools...

Tools and explanations shown to participants during the online community:

- **Version 1 - Only seeing content from users in your network:** Choosing to only see content posted by users you 'follow' or are 'friends' with.
- **Version 2 - Only seeing content from some 'trusted' users:** for example, 'trusted' users (like Crowd Control on Reddit), or users that are verified / have a tick next to their profile.
- **Version 3 - Filter out potential spam:** for example, Reputation Filter on Reddit.

The option was felt to offer a safer and more predictable online experience, but limiting exposure to new content was a concern

Version 1: Choosing to only see content posted by users you 'follow' or are 'friends' with

Benefits

- Gives users **control to curate their online experience** so that they are only seeing content from individuals and accounts they **choose** and **trust**.
- Ensures a **safer, more predictable online experience**, helping to avoid content that users do not want to see or that may be **sensitive and harmful**.
- **Helpful** for users who only want to use social media to stay in contact with friends/family.
- Provides the **option to 'unfollow' accounts** if users want to **stop seeing content they dislike**.
- **Useful for children**, allows parents to ensure their feed is curated to **safe** and **trusted content only**.

"You only see content you are genuinely, 100%, interested in, so your internet experience is overall more entertaining, a more valuable use of your time, and overall less frustrating."
(Non-binary, 26, White, Disabled, Religion – other, Bisexual, Online usage – medium)

Drawbacks

- The method was felt to be potentially too prescriptive – seeing **less diverse content** was a concern, due to a potential **lack of exposure to content outside of their immediate circle**, including different perspectives, ideas, as well as other content they may enjoy and find beneficial.
- The option **does not entirely prevent coming across sensitive and harmful content**, as the individuals / accounts they follow could potentially post it.

"Your exposure is limited [to] just those people...and so your outlook on subject matters will likely be limited by only their opinions and politics."
(Male, 39, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)

While some would be likely to use this option to create safer online experience, most found it too restrictive, as they would prefer being able to discover new content or accounts

Choosing to only see content posted by users you 'follow' or are 'friends' with: Interest in usage

More interested in using

- Those who would be likely to use the option would choose it to have a **safer and more controlled experience online**, focusing only on the content that they want to see posted by friends, family and accounts they follow, and avoiding potentially hateful or sensitive content.
- While the views were mixed across groups, this sentiment was **prevalent among ethnic minority and muslim participants.** In general, there was a sense that they want to be able to presented with a choice and to stay informed from a range of sources, while simultaneously being able to narrow the content to their circle when it is useful (e.g., temporarily enabling the option for mental health purposes).
- Some felt it would help to **manage time spent online**, due to the number of posts being limited.
- Participants would mostly use it on platforms like X, Instagram, and TikTok, where they feel exposed to a wide range of content from accounts that they don't follow. A minority used the option and found it **helpful** in keeping their feed more **enjoyable;** however, some said the feed still occasionally brought up other content.

Less interested in using

- Those who would be unlikely to use the option mentioned that it feels **restrictive** and would **prevent them from discovering new content and accounts** they might be interested in or find beneficial. Similarly, they feel it would limit coming across a broad range of opinions and perspectives, as well as keeping up with new trends, interests (e.g., creative content), or important topics and news.

"More likely, as I think it will give more reason to be offline more, I would be more considerate to who and what I follow so it's more fulfilling rather than void filling."
(Female, 30, Caribbean, No religion, Cisgender, Straight, Online usage – light)

"I would not use this option because it shuts you off from the world and it's very limited to what you can see on there. It would be very boring."
(Male, 38, Chinese, No disability, Buddhism, Cisgender, Gay, Online usage – light)

There was a preference for users to have the choice if they wish to use this option, as opposed to it being applied by default

Choosing to only see content posted by users you 'follow' or are 'friends' with: Attitudes towards default settings

On

- Those who felt this filtering method should be switched on by default mention that it would allow users to experience the platforms with only the content they **trust** and **want to see**, before deciding whether they want to engage with the wider content.
- Participants highlighted that **the setting should be automatically switched on for children / younger people** to ensure safer online experiences.

Off

- Those who did **not** think that this option should be on by default, felt that it would **limit online experience** and stop users from discovering content and aspects of social media that they might enjoy.
- Instead, they preferred being **given a choice when signing up**, allowing them to have **control** over deciding how they want to use social media. Alongside that, having clear guidance on how to later change the settings would be important.

"This should be automatic, it will allow people to experience online space the way it initially was, engaged with what you know or already like. If someone really wants to be online and see all sorts, they have time to consider before it's bestowed upon them."
(Non-binary, 26, White, Disabled, Religion – other, Bisexual, Online usage – medium)

"There should be nothing automatic; instead, a mandatory prompt should be given to selected this setting should the customer chooses.."
(Male, 38, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)

"There should be an option I think that automatically for younger people… as it's safer for them online and for adults, they should have the option."
(Female, 60, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)

Seeing content only from ‘trusted’ or ‘verified’ users did not appeal due to a desire to see a wide range of content

Only seeing content from some users, for example, ‘trusted’ users, or users that are verified.

- The tool was seen as **useful** in allowing users to choose whom they want to see content from. In addition, engaging only with the accounts they trust and that are familiar and safe was felt to be potentially helpful in **avoiding misleading** or **sensitive and harmful content.** They felt it could be potentially useful for protecting their mental health by sticking to ‘safe’ content. Some expressed that the tool would be especially useful to limit hateful content, such as sexist language,.
- However, most were **unlikely to use it,** fearing it would restrict options and prevent the discovery of new or important content.
- **Trust in verification status was low as it can be purchased,** particularly on platforms like X, so it doesn’t guarantee accuracy of their posts and verified users might still share harmful views.
- Only one participant reported previously using the tool but found it limiting.

“This would be helpful and I would most likely use it at some point; especially when my mental health is suffering and I want to stick within the 'safe' compounds of those I follow.”
(Gender fluid, 33, White and Black African, Disabled, Evangelical – independent / non-denominational, Straight, Online usage – light)

“I wouldn’t use this because verified users are not always trusted… and in the past, we’ve seen users essentially pay to become verified.”
(Non-binary, 26, White, Disabled, Religion – other, Bisexual, Online usage – medium)

While spam filters were seen as potentially valuable for keeping the feed focused on genuine content, some were concerned about genuine accounts being inaccurately categorised as spam

Filter out potential spam.

- The feature was felt to be **useful** for keeping **the feed clean** from spam and bots, and focusing on **genuine content**, particularly on Reddit and Meta platforms, where spam accounts were felt to appear frequently.
- However, participants noted that users can be downrated for a range of factors or **platforms could mistakenly filter them as spam**, raising **concerns** about disingenuous reporting and the overall effectiveness of the tool.
- Since the tool would filter out spam before users could see it, there were concerns about missing out on interesting posts, particularly in cases where content was mistakenly categorised as spam.
- Only a minority reported using the tool on Facebook and Instagram.

"This seems like a good tool, I'm more likely to trust the community as a whole rather than a site's management."
(Male, 35, White, Disabled, No religion, Cisgender, Bisexual, Online usage – high)

"It would be useful for filtering out bot accounts and also accounts that are posting spam content. Although genuine accounts may get filtered as a result."
(Male, 45, Caribbean, Disabled, No religion, Cisgender, Gay, Online usage – light)

6.4 Content Overlays

Content overlays were one of the most valued features in the research, as they allow users to protect themselves while also giving them the choice to see individual pieces of content

Content overlays

- Most participants found the tool **useful**, giving users the **option to decide** if they want to view the content, whilst also acting as a **warning** and mentally preparing them to see the content.
- Participants expressed that **adding a description** about the content being hidden would be helpful to enable users to make more informed choices.
- However, a few noted that the **'warning' may have an unintended effect** of intriguing some to view the content, which could be triggering, sensitive or harmful. A few also worried that some users may click out of curiosity and in turn this would prompt the algorithm to show more similar content.
- There were also concerns among a minority that the tool may be **misused** to censor certain types of content (e.g., due to bias or for political reasons), as well as that the warnings could be **inaccurate**, with some reporting seeing them for content that they felt was not sensitive.
- A minority felt it could **interfere with a seamless online experience** (i.e., due to needing to take an additional action to view), particularly if they did not find the content sensitive and harmful, making the tool inconvenient and unhelpful.
- Participants reported previously coming across this feature on Instagram, X, Facebook, Reddit and TikTok. However, a few noted that on X, they were still seeing sensitive and harmful content without warnings.

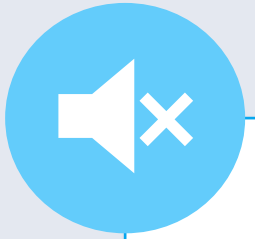
"I like that, it prepares me that some content will be extreme and I can choose to see it or not see it based on the emotional state I am in."
(Female 28, Indian, No disability, Hinduism, Cisgender, Straight, Online usage – medium)

"I have used this and I think it's a really good tool, gives the user the control on whether or not they want to view the content."
(Male, 39, White, No disability, No religion, Cisgender, Straight, Online usage – medium)

"I have used this tool on Instagram and TikTok and I appreciated the warning before clicking onto the content as opposed to seeing it right away. This gave me a choice to consider the overall topic of the content and whether or not I wanted to risk seeing potentially upsetting content."
(Female, 28, White, Disabled, No religion, Cisgender, Lesbian, Online usage – medium)

6.5 Actions against individual users

Participants shared their thoughts on the following tools to take actions against individual users...



Tools and explanations shown to participants during the online community:

- **Blocking individual users:** They can't contact you and you don't see their content.
- **Muting individual users:** You don't see their content, but they could still contact you.

Blocking individual users is seen as beneficial for those who were victims of online harassment

Blocking individual users so that they can't contact you and you don't see their content.

- Although the tool was seen as more manual, it was generally seen to be **useful,** as it provides **granular control over who can engage with them**. Most saw value in using it, as they felt it is a **convenient way to protect themselves** from sensitive, harmful or otherwise inappropriate content appearing on their feed that is posted and/or shared by specific users, as well as harassment via messages or comments. A few noted that it may be **beneficial for victims of online harassment** targeted at their characteristics (e.g., gender, sexual identity, ethnicity), as well as in cases of cyber bullying.
- However, some noted that the feature **may <u>not</u> be sufficient to protect themselves** online – similar content or messages might still appear from other accounts, or users creating new accounts to target someone.
- Participants reported previously using the tool on Facebook, Instagram, X, Reddit, TikTok and Tumblr to block users they didn't want to engage with, or if users posted or sent sensitive, harmful or hateful content, or if they were a suspected bot or scam.

"This is necessary already due to bullying and harassment that is so prevalent online. On apps where there are less stringent rules on content, this feature is a must."
(Female, 41, White and Black Caribbean, No disability, No religion, Cisgender, Straight, Online usage – light)

"This is common on almost all platforms I use. I have had to block people and pages with questionable / sensitive / immoral content."
(Male, 38, Pakistani, No disability, Islam, Cisgender, Straight, Online usage – light)

Muting individual users is valued in cases where users want to avoid seeing content from specific accounts, without blocking them

Muting individual users so you don't see their content, but they could still contact you.

- It was generally considered to be a **useful tool**, particularly when **they <u>do not</u> want to block another user** (e.g., friends, family), but do not wish to see their content, either because they were uninterested, or because the content was upsetting, sensitive or harmful. Those who did not relate to the situation of having someone (e.g. a friend or family member) whom they want to remain following, but whose posts or content they do not want to see, did not see value in the tool or how it could be more useful than blocking/unfollowing.
- Participants previously used the option on Discord, Facebook, Instagram, X and Bluesky when they did not want to see posts of those they followed / in their friends list, but whom they did not want to block. This was mostly because they were **uninterested in content,** or if they posted content that was upsetting.

"I have also used this on [X] for friends I have that I don't always share the same interests with."
(Female, 31, White, No disability, Other religion, Cisgender, Straight, Online usage – medium)

"Sometimes useful and less extreme than blocking them. Sometimes you just want a break from seeing their posts."
(Male, 41, White, Disabled, No religion, Cisgender, Straight, Online usage – light)

6.6 Actions against individual pieces of content and content types

Participants shared their views on the following actions against against individual pieces of content and content types...

Actions and explanations shown to participants during the online community:

- **Reporting users and/or posts to the platform**.
- **See less of this:** Clicking 'see less of this'/'not interested'/'hide'/'dislike' for individual content.
- **Community notes:** This might involve you posting a ‘community note’ about the accuracy of someone else’s post (like on X and Instagram).
- **Muting words or hashtags:** This means filtering out content with specific words/hashtags you don't want to see.

Reporting tools were seen as helpful, however, were often ineffective, due to not leading to the desired results

Reporting users and/or posts to the platform.

- The tool was felt to be **important for making social media safer** and **easier to use**. The majority would be likely to report the content that is sensitive (e.g. violent, graphic), harmful or breaks the platform's rules, with a few participants highlighting that the option is important, given that women and minority groups are often targeted online.
- While in principle most liked the tool, it was felt to be **ineffective** among some who had previously used it, due to the platforms' moderation processes, and often no action being taken or the review process taking too long. Participants specifically mentioned examples of reporting hateful, racist or violent content and receiving a response that it did not violate guidelines.
- Another concern that participants expressed and that they have previously seen happen, is that the tool **can be misused** to report content that some users dislike or disagree with, rather than it being harmful or otherwise inappropriate.
- Participants previously used the tool on Facebook, Instagram, X, Reddit and TikTok when seeing content, messages or comments that were inappropriate, harmful, hateful, spreading misinformation or were suspected posted by spam or bots.

"This is a good idea… My experience of this is not good though, I have reported stuff that clearly shouldn't be on there such as acts [of] extreme violence etc. and you get… an automated message back saying… it doesn't break our guidelines."
(Female, 32, White, No disability, No religion, Cisgender, Straight, Online usage – light)

"This will always be necessary for the most extreme content and behaviours although reporting is also used maliciously/vexatiously."
(Female, 41, White and Black Caribbean, No disability, No religion, Cisgender, Straight, Online usage – light)

‘See less of this’-style tools were seen as potentially beneficial; however, there were questions around long-term effectiveness

Clicking 'see less of this'/'not interested'/'hide'/'dislike' for individual content.

- It was felt to be **a useful and easy way to improve the algorithm of their feed**, with most participants noting that it would be potentially useful for content that is either irrelevant, upsetting or harmful.
- However, the longevity of these actions on the algorithms was questioned. Based on their experience, it only worked for a **short period of time or not at all,** with unwanted **content still appearing in their feed.**
- There was **uncertainty** about how the tool works in practice, which was raised by both, those who previously have and have not used the tool. For instance, participants were unsure how the tool would affect the content that appears on their feed. They were uncertain whether the algorithm would correctly **interpret the type of content** that they don’t want to see, and whether the tool would stop showing similar content completely or only some of it.
- Some participants reported previously using the tool on TikTok, Instagram, Facebook, X, Reddit and YouTube for content they disliked and wanted to avoid, but that did not warrant blocking or reporting (e.g. not harmful content).

“It is useful for making my experience more personal. Not so useful as unwanted content may still show up.”
(Male, 45, Caribbean, Disabled, No religion, Cisgender, Gay, Online usage – light)

“My algorithm is pretty tailored to my interests, but I would definitely use this if I started seeing more content I didn't like, that was either sensitive or AI slop.”
(Female, 34, White, Disabled, No religion, Cisgender, Straight, Online usage – medium)

Community notes were favoured for protecting users from misinformation, but were felt to be open to subjectivity and misuse

Informing other users about potential issues with content.

- The tool was felt to be **useful** to **keep social media safe** by warning others and keeping accountable those who post harmful content, misinformation or who break the platform's rules.
- However, participants noted a range of factors that would **discourage** them or others from posting the community notes, including it being time-consuming, feeling that **they are not an 'expert' to comment** or being **worried** about **their own subjectivity**.
- There was a concern about **subjectivity** and potential **misuse** of the tool (e.g., leaving community note for content users disagreed with, rather than it being harmful), making them apprehensive about its accuracy.
- Some were concerned that it could be potentially **inefficient** in preventing harm or misinformation, due to the content circulating online before users are able to spot the issues and warn others, as well as users potentially ignoring the warnings.
- Only a few reported previously using the tool and posting community notes on X and Facebook when coming across misinformation or potential scams. Some also mentioned coming across and reading the notes, which they found helpful in identifying potential misinformation.

"It's also especially good that if you like the post on X that it'll alert you later that changes have been made to it/someone's noted it for being misleading or not wholly accurate."
(Female, 32, White, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – light)

"Benefit is community policing. Drawback is that the majority isn't always right."
(Male, 44, White, Disabled, No religion, Cisgender, Gay, Online usage – light)

Muting individual words and hashtags was seen to be valuable in giving users control over the specific topics they want to mute, however, there were concerns about how well it works in practice

Muting individual words and hashtags.

- It was seen as helpful **for filtering harmful, upsetting or irrelevant topics**, giving users more control.
- However, most felt it **could be ineffective due to numerous related hashtags and words**, making muting time-consuming.
- Users also noted **risks of circumvention** through altered spellings or mislabeled content. A few mentioned that they would avoid using the tool, fearing it **might overgeneralise** and block content they're interested in when similar words appear in different contexts.
- A minority reported previously using the tool on X, TikTok, Instagram, and Tumblr to mute hateful or irrelevant content.

"It's a simple way to make my feed more enjoyable and focused on what I actually want to see. The only downside is that sometimes useful content might get filtered out by accident, but overall, it seems really helpful."
(Male, 31, Caribbean, No disability, Church of England / Anglican / Episcopal, Cisgender, Straight, Online usage – high)

"While it is helpful in filtering and reducing the appearances of certain topics I do not wish to see, it is a long shot and cumbersome way of doing things. New hashtags emerge daily, it's a job trying to keep up with muting…"
(Female, 29, African, No disability, Roman Catholic, Cisgender, Straight, Online usage – high)

July 2026

Exploring how internet users navigate sensitive and harmful content

A qualitative study.

Report prepared by YouGov Qualitative.