Twitter UK
20 Air Street
London
W1B 5AN

twitter.com

Response to Ofcom Call for evidence: Second phase of online safety regulation: Protection of children

About Twitter

Twitter is what’s happening in the world and is a public, real-time global information service, where people can see every side of a topic, discover news, share their perspectives, and engage in conversation. Twitter is available in more than 40 languages around the world, and can be accessed via Twitter.com, and an array of mobile devices via Twitter owned and operated mobile applications (e.g. Twitter for iPhone and Twitter for Android).

The service allows people to consume, create, distribute, and discover content. From breaking news and entertainment, to sports, politics, and everyday interests, people on Twitter provide insight into every angle of a story.

On Twitter, you can join the open conversation, including with photos, video clips, and live-streaming. Twitter requires people using our service to be 13 years of age or older.

Twitter’s users may choose to use the platform pseudonymously (i.e. the username and profile picture of the Twitter account do not need to identify the account holder); this may be important for users who seek to disclose whistleblowing, protect minorities etc. Other Twitter users may choose to verify themselves through Twitter Blue to improve reach and impact. However, Twitter prohibits individuals from misleading, confusing or deceiving other users to preserve the authenticity of conversation on the platform.

Twitter’s approach to protecting younger users

In accordance with regulatory requirements in the US and UK, Twitter requires its users to be aged 13 or over to create an account. Twitter is not targeted at younger users and studies consistently show that platforms other than Twitter are favoured by younger UK users.This is evidenced by Ofcom research 'Children and parents: media use and attitudes report 2022' which shows that of social media sites used, just 21% of 12-15 year old respondents use Twitter. Further, only 15% in the same category have set up a profile meaning that they use Twitter as a logged-out user and therefore there are controls protecting them from seeing sensitive media (e.g. adult content, violence) and age-restricted advertising (e.g. alcohol).

Twitter is a client of Comcore and using data science and audience insights across viewership, Comscore provides a third party source for reliable measurement of cross-platform audiences. Comscore data from December 2022 shows that 98% of Twitter users are over the age of 18.

Twitter has designed its platform to provide a different experience for younger users as well as deploying a number of safety features in order to keep all users safe. By way of example:

1. *Age restricted content* **–** Twitter automatically restricts users who are under 18, or who do not include a birth date on their profile, from viewing sensitive media content (as set out in our sensitive media policy). In addition, a different approach to advertising is taken for users who are either under 18 or who do not include a birth date on their profile. Twitter prohibits marketing or advertising of a number of products and services to minors, including alcohol, weapons, weight loss products, health supplements, gambling products, sexual products and services, permanent cosmetics and other forms of body branding. These age restrictions are in addition to complete bans on advertising certain products on Twitter, including any advertising of controlled substances, tobacco and projectiles.
2. *Safe Search* – users of the Twitter platform have control over what they can see in search results through selecting the Safe Search settings. Safe Search is automatically enabled for anyone with a birth date under 18 years of age. Once enabled, these filters are designed to exclude from search results any potentially sensitive content (such as content which is excessively gory, violent, or of a graphic sexual nature) along with accounts a user has muted or blocked (for whatever reason).
3. *Sensitive Tweet Warnings* – Twitter's sensitive media policy prohibits users from including graphic content or adult nudity and sexual behaviour within areas that are highly visible on Twitter, including in live video, profile, header, List banner images, or Community cover photos. If a user shares this content on Twitter, the policy requires the user to mark their entire account as sensitive or to add sensitive content warnings to individual photos of videos. Doing so places an interstitial warning message on images or videos they post which contain sensitive media. Twitter may also place an interstitial warning message on some forms of sensitive media. An interstitial warning alerts a user that a Tweet contains sensitive content such as nudity, violence or sexual content and means other users can only see the media if they actively click to "show" the Tweet; it cannot be viewed by accident.
4. *Controlling replies* – users can choose who will be able to reply to their Tweets when posted. The default position is that everyone can reply but options are available to turn off all replies or only allow the accounts mentioned in the Tweet to reply. A user can also change who can reply to their Tweets, or turn off replies, after the Tweet has been posted.
5. *Protected accounts* **–** when an adult user signs up for Twitter, they can choose to keep their Tweets public or to protect them so that only approved followers can see and interact with them. By contrast, when a user signs up for Twitter with a date of birth under 18 years of age, the account is automatically defaulted to protected mode.
6. *Account filters* **–** users can filter the types of accounts they see in their notifications timeline. This feature allows users to mute notifications from certain categories of users, such as those with accounts who have not confirmed their phone number or email address, new accounts, accounts who have a default profile photo, accounts that the user does not follow or accounts that do not follow the user.
7. *Block and mute* **–** users can block accounts instantly if they do not want that account to see their Tweets and/or the user does not want to see the account's Tweets. Users can also mute an account if they don't want to see their Tweets, but don't want to unfollow

the account. Particular words, conversations, phrases, emojis and hashtags can also be muted to ensure those words or phrases do not appear on the user's timeline.

8. *Safety Mode:* Twitter introduced 'Safety Mode' in September 2021, which allows users to temporarily block accounts for using potentially harmful language or sending repetitive and uninvited replies or mentions.

Twitter's approach to age assurance

Twitter is committed to protecting child safety online and has launched a range of age assurance measures to seek to ensure that, in the UK, only users aged 13 and over are permitted to access the Twitter platform. Twitter approaches the challenge of age assurance by combining self-declaration (i.e. users providing their date of birth) with additional technical measures (as described in the ICO's Age-Appropriate Design Code) which together aim to ensure that the account holder's self-declared age is genuine and that appropriate controls are in place to protect teenagers.

Twitter first collects the user's age through the neutral presentation of a date of birth prompt. Once a date of birth is entered, Twitter then determines the user's age. At this stage, new users are informed that Twitter uses their age to customise their experience, including advertising, and provides options as to the visibility of the user's date of birth to others.

Users who enter a date of birth that indicates they are under the age of 13 are not permitted to go any further in the account opening process. There is an account restoration appeals process for those who erroneously enter the wrong date of birth and are not permitted to proceed with account opening or who have their account off-boarded as a result of an indication of being under 13. As part of the account restoration appeals process, the user is required to provide ID documentation proving that they are over the age of 13. These appeals are subject to human review. If Twitter cannot verify the user is over the age of 13, the account is not restored. This appeals process is often used by business accounts who enter the date of incorporation, rather than children seeking to attempt to gain access to Twitter.

Users are also able to report accounts which they believe are operated by someone who is underage and Twitter will take action if appropriate.

Users who enter a date of birth that indicates they are over 13 but under 18 are prevented from seeing sensitive content, such as adult content on any surfaces (e.g. their timeline or search results) in line with Twitter's sensitive media policy and the automatic application of 'Safe Search' for such users. Any sensitive content contained in the account holder's page will be obscured by a sensitivity screen.

Twitter prohibits marketing or advertising of a number of products and services to minors, such as alcohol. In respect of advertising, users who have not registered a date of birth to their account will a) have their age inferred based on their interactions with Twitter to prevent them from seeing advertising in breach of the Prohibited Content for Minors advertising policy; and b) will be asked to enter their date of birth before being permitted to follow the accounts of certain brands (eg. alcohol). More information on our advertising policy, including prohibited content for minors can be found here.

In addition to the measures above, Twitter has been working with experts to research further age assurance measures that incorporate 'privacy by design' principles (required by the GDPR) and work in a global context. These measures also need to account for the importance of online anonymity for minorities and disadvantaged communities around the world and the use of Twitter as a platform for whistle-blowers and human rights advocates.

Furthermore, there are currently a range of projects which are underway and actively being examined by Twitter with these considerations in mind, focused on the best interests of children. In January 2023, we updated our Sensitive Media Policy to prohibit posting media that is graphic or share violent or adult nudity and sexual behaviour within live video or in profile header, List banner images, or Community cover photos. Media depicting excessively gory content, sexual violence and/or assault, bestiality or necrophilia is also not permitted. Information on sensitive media settings on Twitter is available here.

Additionally, our systems and teams may add notices on Tweets to give you more context or notice before you click through. Some of the instances when we may add notices on Tweets include:

- *Placing a Tweet behind an interstitial:* We may place some forms of sensitive media like adult content or graphic violence behind an interstitial advising viewers to be aware that they will see sensitive media if they click through. (Note: you cannot click through on Twitter for iOS.) This allows us to identify potentially sensitive content that some people may not wish to see. Learn more about how to control whether you see sensitive media.
- *Age restricted content:* We restrict viewers who are under 18, or who do not include a birth date on their profile, from viewing adult content. (Note: people over 18 can opt out of viewing sensitive media on Twitter by updating their settings here).
- *Placing a Tweet in violation behind an interstitial:* We may allow controversial content or behaviour which may otherwise violate our rules to remain on our service because we believe there is a legitimate public interest in its availability. When this happens, we limit engagement with the Tweet and add a notice to clarify that the Tweet violates our rules, but we believe it should be left up to serve this purpose.

Twitter’s approach to harmful content

Developing a policy or a policy change requires in-depth research around trends in online behaviour, developing language that sets expectations around what’s allowed, and reviewer guidelines that can be enforced across millions of Tweets. We gather input from around the world so that we can consider diverse, global perspectives around the changing nature of online speech, including how our rules are applied and interpreted in different cultural and social contexts. We then test the proposed rule with samples of potentially abusive Tweets to measure the policy effectiveness and once we determine it meets our expectations, build and operationalise product changes to support the update. Finally, we train our global review teams, update the Twitter Rules, and start enforcing the relevant policy.

Twitter's purpose is to serve the public conversation. The Twitter rules are in place for all users, including younger users, in order to ensure all people can participate in the public conversation freely and safely as outlined below.

Safety:

- *Violent Speech:* You may not threaten, incite, glorify, or express desire for violence or harm.

- *Violent & Hateful Entities:* You can’t affiliate with or promote the activities of violent and hateful entities.
- *Child Sexual Exploitation:* We have zero tolerance for child sexual exploitation on Twitter.
- *Abuse/Harassment:* You may not share abusive content, engage in the targeted harassment of someone, or incite other people to do so.
- *Hateful conduct:* You may not attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.
- *Perpetrators of Violent Attacks:* We will remove any accounts maintained by individual perpetrators of terrorist, violent extremist, or mass violent attacks, and may also remove Tweets disseminating manifestos or other content produced by perpetrators.
- *Suicide and self-harm:* You may not promote or encourage suicide or self-harm.
- *Sensitive media:* You may not post media that is excessively gory or share violent or adult content within live video or in profile or header images. Media depicting sexual violence and/or assault is also not permitted.
- *Illegal or Certain Regulated Goods or Services:* You may not use our service for any unlawful purpose or in furtherance of illegal activities. This includes selling, buying, or facilitating transactions in illegal goods or services, as well as certain types of regulated goods or services.

Privacy:

- *Private Information: Y*ou may not publish or post other people's private information (such as home phone number and address) without their express authorization and permission. We also prohibit threatening to expose private information or incentivizing others to do so.
- *Non-Consensual Nudity:* You may not post or share intimate photos or videos of someone that were produced or distributed without their consent.
- *Account Compromise:* You may not use or attempt to use credentials, passwords,tokens, keys, cookies or other data to log into or otherwise access, add, delete or modify the private information or account features of any Twitter account other than your own (or those you have been directly authorised to do so via Twitter’s Teams authorization, OAuth authorization or similar mechanism).

Authenticity:

- *Platform Manipulation and Spam: Y*ou may not use Twitter’s services in a manner intended to artificially amplify or suppress information or engage in behaviour that manipulates or disrupts people’s experience on Twitter.
- *Civic Integrity:* You may not use Twitter’s services for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process.
- *Misleading and Deceptive Identities:* You may not impersonate individuals, groups, or organisations to mislead, confuse, or deceive others, nor use a fake identity in a manner that disrupts the experience of others on Twitter.
- *Synthetic and Manipulated Media:* You may not deceptively share synthetic or manipulated media that are likely to cause harm. In addition, we may label Tweets

containing synthetic and manipulated media to help people understand their authenticity and to provide additional context.

- *Copyright and Trademark:* You may not violate others' intellectual property rights, including copyright and trademark. For more information, please see our trademark policy and copyright policy.

Accessibility of terms of service

Upon signing up to Twitter, users are required to agree to our terms of service. While detailed information is provided on each of these rules by tapping on the headings, this page is intended as a short, accessible and clear summary of behaviours that are not permitted on Twitter.

One of Twitter's core missions is to ensure that it is accessible to all users, including all those that satisfy minimum age requirements. Our Accessibility policy ensures that established guidelines and best practices are incorporated into the platform to allow everyone to share their ideas and experience Twitter in the best possible way.

Twitter's Trust and Safety Team is dedicated to advocating for the safety of its users and protecting their rights, and therefore engages with experts to ensure Twitter offers the most appropriate solutions for persons of all ages. To support parents and guardians with children using Twitter, we have collaborated with Internet Matters, an organisation launched with the specific intention of supporting parents and carers to navigate the digital landscape, to develop a parental controls guide. This guide provides step-by-step instructions for parents to manage their child's account. These instructions allow parents to protect their child's Tweets and prevent children from receiving abusive or inappropriate content. It also gives the parent control over who can contact their child and what personal data is shared. The controls also allow parents to limit who can see their child's Tweets, who can contact them and who can tag them.

Supporting user safety and wellbeing

A number of product changes we have introduced in recent years have had a positive impact on harmful behaviours. We've previously partnered with local mental health authorities and non-profits in each market to offer #ThereIsHelp - a notification service that provides valuable information and resources via Twitter and email.

One such prompt has been on the topic of suicide and self-harm. When someone searches for terms associated with suicide or self harm, the top search result is a notification encouraging them to reach out for help. Depending on user location, we'll provide the most appropriate partner organisation. In the UK for example, we have partnered with the Samaritans on this.

In 2020, we tested prompts that encouraged people to pause and reconsider a potentially harmful or offensive reply before they hit send. While it was clear that prompts cause people to reconsider their replies (about one third of the time), we wanted to know more about what else happens after an individual sees a prompt. To understand this, we conducted a follow-up analysis to look at how prompts influence positive outcomes on Twitter over time. In June, we published a peer-reviewed study of over 200,000 prompts conducted in late 2021. We found that prompts influence positive short and long-term effects on Twitter. We also found that people who are exposed to a prompt are less likely to compose future offensive replies. Data and further information is available here.

Twitter's use of algorithms

Twitter, along with the majority of online services, uses algorithms in some form to suggest relevant content to users, which helps improve the usability and accessibility of online services.

Twitter uses algorithms to help provide content to users. The content that Twitter provides to users is based on the choices they make when using Twitter (e.g. accounts that users follow and the topics or interest users choose) and recommendations that Twitter makes to users based on a variety of signals, including Tweets the user has engaged with or content that is popular in the user's network. On March 17th it was announced that Twitter's algorithm will open source all code used to recommend tweets on March 31st 2023.

Algorithms help show users content on many of the surfaces that they may access on Twitter, including Notifications, Topic Landing Pages, Explore, Spaces Tab, and the Who to Follow module. Users will typically see a main feed of content when they open Twitter (referred to as the Home Timeline in Twitter) and it is sub-divided between a 'Following' tab (which shows Tweets posted or Re-Tweeted by accounts a user is following in reverse chronological order) and a 'For You' tab (which suggests more Tweets from accounts and topics a user follows as well as recommended Tweets). Users may also see content such as Promoted Tweets or Retweets in their Home Timeline. Twitter outlines how it recommends content to users across various areas (including the Home Timeline, Explore, Notifications) and how users can control those recommendations in a Help Centre article - here.

Promoting healthy conversations is one of Twitter's core principles; "freedom of speech is a fundamental human right — but freedom to have that speech amplified by Twitter is not." While our enforcement philosophy empowers people to understand different sides of an issue by allowing many forms of speech to exist on our platform, we also work hard to prevent the amplification of harmful content on Twitter, particularly for younger users. In addition to the setting we set out in the Protecting Younger Users section above, the measures that Twitter takes to help prevent amplification in certain areas (e.g. as a recommendation in Home timeline from someone you don't follow) include:

- Neither the Following tab or the For You tab permits sensitive content or inappropriate advertising to be surfaced for accounts under the age of 18. Twitter's policies and enforcement measures seek to reduce the risk that illegal or potentially harmful content could be shown to users.
- During events such as civic processes and certain crises, we may proactively recommend informative messages or updates to add context to disputed or potentially misleading narratives that emerge.
- Twitter strives to keep certain categories of content out of our recommendations (for example: content that violates any of the Twitter Rules, but has been left on the platform due to the public-interest exception; content that promotes the use of regulated substances or weapons; content that is deemed marginally abusive and is ineligible for amplification under our safety policies; and content that automated systems have determined may violate the Twitter Rules, but that has not yet been reviewed by a human and/or may have been identified in error.
- Twitter strives to keep certain categories of accounts out of our recommendations (for example: accounts that recently violated the Twitter Rules; accounts sharing harmful misleading information; accounts that engage in coordinated harmful activity; and accounts that contain graphic violence or hateful imagery in Twitter profile elements)

Algorithms are currently used to provide relevant and appropriate adverts to users based on their age. Twitter uses an additional age inference algorithm for those legacy accounts who have not entered a birthdate. In the UK, entering a birth date has been a mandatory requirement to create an account on Twitter since 2020 (from 2018-2020, Twitter required users to confirm that they were 13 years of age). As a result of the mandatory age gate, the proportion of accounts for which this is relevant is decreasing.

The algorithm works by collecting data from Twitter accounts with a birthdate. Following this, the known age from those accounts allows Twitter to infer the age of other Twitter users without a birthdate based on a number of attributes and interactions with the service (e.g. accounts followed, country and level of activity). This process allows Twitter to infer a single, specific age (e.g. 23) of users without a birthdate and these users will then be assigned into an age bucket.

Twitter is shifting to an approach of content de-amplification (vs. removal) for certain kinds of content that may be widely objectionable, but does not violate the law. This will mean that one's experience on Twitter will mirror their experience in other parts of the internet - where unless you actively seek out unsavoury content, you should not encounter it.

We believe de-amplification - where we reduce the blast radius of this content by ensuring people are only exposed to it when they intentionally seek it out - is the most effective way to maintain the health of the platform without unduly restricting user activity. Twitter will continue to remove content that is in violation of the Twitter Rules, including illegal content.

Content moderation

We use a mixture of proactive and reactive detection in enforcing the Twitter rules. Automated tools include:

- *Hash-sharing:* Our current methods of surfacing potentially violating content for human review include leveraging the shared industry hash database supported by the Global Internet Forum to Counter Terrorism (GIFCT).
- *PhotoDNA and internal proprietary tools:* a combination of technology solutions is used to surface accounts violating our rules on Child Sexual Exploitation. As standard and required by law, we continue to report to the National Center for Missing and Exploited Children (NCMEC) when appropriate.
- *Detection of abusive content:* Now, a majority of the abusive Tweets we remove, we have detected proactively using machine learning.

Our content moderators provide moderation services 24 hours a day, 7 days a week. We have teams spread around the world specifically trained in this work so that we can provide this level of coverage in the languages we serve on Twitter. For safety and security reasons, the locations of these teams are not disclosed publicly.

Updates about significant current events or rules and policy changes are shared with all content reviewers, to give guidance and facilitate balanced and informed decision making. In the case of rules and policy changes, all training materials and related documentation is updated.

Both full time and contract Twitter employees receive information in their onboarding that acknowledges they may potentially come into contact with material that is sensitive and potentially distressing in nature - it's the nature of the work we do to ensure healthy conversation on our service. The wellbeing of those who review content is a primary concern for our teams and our highest priority is to ensure our staff and partners are treated with compassion, care, and respect. We are continually evaluating our partners' standards and remain committed to protecting the well-being of the teams tasked with this important and challenging role. We have a full suite of support services available for our employees, including content moderators.. Some of the measures we take include, establishing resiliency programs across all of our partners, committing to recurring leadership visits and ongoing feedback loops and communications between all of our teams.

In the long term, we believe one of the most valuable investments we can make is in technology. The more we can leverage technology to minimise the exposure to content, the less frequently our employees and contractors will come into contact with it.