



12 January 2021

Committee Secretary
Select Committee on Social Media and Online Safety
PO Box 6021
Parliament House
Canberra ACT 2600

By email: smos.reps@aph.gov.au

To whom it may concern,

Thank you for the opportunity to provide a written submission and feedback to the Select Committee into Social Media and Online Safety (**the Committee**) as part of its inquiry into Social Media and Online Safety (**the inquiry**).

We share the Australian Government's desire to promote online safety, and Twitter remains focused on making people feel safe, secure, and empowered to participate in the public conversation every day.

As we continue to iterate and strengthen our approach to meet evolving contours and challenges surrounding online behaviours, we're moving with urgency and purpose. We remain committed to developing and enforcing a range of policy, procedural, and product changes to help people feel safe, welcome, and able to control their experience on Twitter. We support smart regulation, and our focus is on working with governments to ensure that regulation of the digital industry is practical, effective, inclusive, and feasible to implement to ensure that certain core democratic values are intact while promoting tech innovation, including Twitter's core commitment to an Open Internet worldwide.

Our submission stands together with the respective submissions from the Digital Industry Group Inc. (**DIGI**) and Communications Alliance (**Comms Alliance**), both of which Twitter is a member. For clarity, and to complement and reinforce these statements, we've structured this submission to address the key issues within the Committee's Terms of Reference for this inquiry as they pertain to Twitter operating in Australia.

Twitter is committed to working with the Australian Government, our industry partners, non-government organisations, academics, and wider civil society as we continue to build our shared understanding of the issues and find optimal ways to approach these together.

We trust this written submission will be a useful input to the Committee's consultation process. Working with the broader community we will continue to test, learn, and improve quickly so that our platform remains open, accessible, effective, and safe for everyone.

Thank you again for the opportunity to input into this important process.

Kind regards,



Kara Hinesley
Director of Public Policy
Australia and New Zealand



Kathleen Reen
Senior Director of Public Policy
Asia Pacific



Table of contents

Introduction	4
Protecting the Open Internet	
Twitter’s commitment to healthy conversations and public consultation.	8
Anonymity and pseudonymity on Twitter	9
Twitter’s Trust & Safety Council, partnerships, and mental health	10
Twitter’s Developer Platform and Commitment to Transparency	11
Content Authenticity Initiative	13
Current online safety landscape in Australia	13
Conclusion	14
Appendix	15



Introduction

Twitter's mission is to serve the public conversation, and online safety is a top priority for the company. It is essential to fulfilling that mission. We acknowledge that the changes brought about by the digital age require ongoing consideration and informed debate. Twitter has invested – and will continue to invest – substantial resources to ensure online safety practices remain at the highest level to keep pace with latest societal, technological, and legal developments.

Twitter remains in open dialogue with government partners to create collaborative partnerships and online safety solutions, while protecting vital public expression. We share the Government's goal of keeping people safe online, especially with regards to the mental health and wellbeing of children and vulnerable communities, and aim to help provide the Committee with information about how Twitter is approaching these issues and the range of measures we take to ensure people feel safe when taking part in the public conversation.

This submission will provide further insight into:

- Our commitment to Open Internet principles
- Product and policy expansions and updates
- Driving online safety through collaborative partnerships and consultation
- Our approach to transparency, freedom of expression, and anonymity

As we consider the current online safety regulatory landscape in Australia, we must also understand the context online safety approaches and internet regulation are developed in, including the diverse challenges and competing policy priorities at hand. The potential consequences for hasty policy decisions or rushed legal regimes will stretch far beyond today's headlines, and are bigger than any single company.

How these challenges are defined, understood, and addressed will affect services of all sizes, the ability of millions of people to share and access information, and the future of the digital economy. Regulatory approaches to new industries are often shaped by the policy responses designed in response to new technologies. This approach struggles to adapt to the unpredictable and rapidly evolving nature of human use of technology and expression. Designing regulation around the established online services of today will risk losing the innovation needed to solve future inequities and could result in negative consequences for digital participation.

More broadly, the policy issues addressed are often rooted in complex societal challenges that exist in offline, as well as online, contexts. As we continue to work together in good faith on these complex issues, we emphasise that these challenges will not be resolved by the removal of content online alone. Bad actors seeking to exploit online services to undermine elections, spread disinformation, and harm others will not be deterred by their accounts being removed. Effective solutions demand a whole of society response that recognises the full scope of the problem being addressed. Policy design must also consider the effects, including unintended consequences, that any framework might have on essential democratic rights.

Protecting the Open Internet

Twitter recognises the need to balance tackling harm with protecting a free and secure Open Internet. In our advancements to promote online safety and address the policy challenges ahead, we have focused on five overarching principles and how they intersect with issues of competition, content moderation, and the role and responsibilities of services like Twitter.

With regards to the creation of digital policy, we believe:

1. The Open Internet is global, should be available to all, and should be built on open standards and the protection of human rights.
2. Trust is essential and can be built with transparency, procedural fairness, and privacy protections.
3. Recommendation and ranking algorithms should be subject to human choice and control.



4. Competition, choice, and innovation are foundations of the Open Internet and should be protected and expanded, ensuring incumbents are not entrenched by laws and regulations.
5. Content moderation is more than just leave up or take down. Regulation should allow for a range of interventions, while setting clear definitions for categories of content.

We have outlined and detailed these principles for the Committee below to provide insights into our work and help inform the policy debate regarding online safety. We have also provided supplementary information regarding Twitter's policies and product features in the appendix of this submission.

(1) The Open Internet is global, should be available to all, and should be built on open standards and the protection of human rights.

The Open Internet has been an unprecedented engine for economic growth, cultural development, and self-expression. For these benefits to continue, it must be available to all.¹ The infrastructure of the internet is itself now a geopolitical space.

Governments should prioritise policies, partnerships, and investments at home and abroad that support and defend the Open Internet, both through regulatory and standards bodies, as well as ensuring domestic regulation does not undermine global norms or set dangerous precedents. Open standards championed by these bodies will provide for greater interoperability, connection, and competition

Access is a critical issue. The principle that information should be able to safely and securely move across borders freely as part of a global internet should be core to democratic regulation.

Rhetoric and policies pursuing national data sovereignty should be avoided and scrutinised. Some actors seek to exploit this concept to strengthen control of and access to data, weakening the Open Internet through forced data localisation and limits on the free flow of data.

The principle that data belongs to a person does not mean that all people's data belongs to the state. Policy makers should avoid the use of extra-territorial application of national content standards as this further undermines the global internet and encourages a race to the bottom, with the entire world's open communications imperiled by those actors least committed to freedom of expression.

Both governments and industry should ensure their approaches to addressing online harm are consistent with universally recognised human rights norms, including proportionality and the protection of privacy and freedom of expression.

(2) Trust is essential and can be built with transparency, procedural fairness, and privacy protections.

There's a deficit in trust with respect to many online services and government functions alike. It's essential every sector works to rebuild trust, beginning with greater transparency. People should understand the rules of online services and the way that governmental legal powers are used. Transparency enables accountability for companies and Governments. Without transparency, there can be no accountability.

Just as due process is a core feature of robust judicial systems, procedural fairness should be a core function of online services. These concepts should be at the core of regulation, particularly where governments seek to require services to remove content and companies take action under their terms of service.

¹contractfortheweb.org/principle-theme/access/



Legislators should ensure clear harmonised standards for safeguarding and processing personal data, supplemented by regulatory guidance as new issues emerge, recognising that it's neither feasible nor desirable to legislate for every potential scenario of how personal data is used in primary legislation.

Fragmented and inconsistent frameworks weaken consumer protection and the establishment of industry norms. While many services do collect data to enable them to provide advertising, granular privacy controls balance the functionality of online services with consumer control while serving a desire to allow people who use services to make informed decisions about the data they share. Individuals should know, and have meaningful control over, what data is being collected about them, how it's used, and when it's shared. In the long run, regulation should protect and encourage services based on a range of business models, not just those built on advertising.

Policymakers should protect the ability to use the internet without having to disclose your real identity, legal identifications, or detailed personal information. This is foundational to a universally accessible internet for all, and it's essential to recognise that not all services require the same amount of personal information to be disclosed or verified and nor should they be required to.

(3) Recommendation and ranking algorithms should be subject to human choice and control.

Recognising that content moderation and content organisation are two different spheres of work, particularly when content is recommended without a positive signal to seek it out, policymakers should prioritise empowering people to have control over algorithms they interact with and ultimately drive an ability to make our own choices among algorithms. Choice can also help foster greater understanding and awareness of how algorithms impact people's online experiences, leading to greater digital literacy.

While algorithmic transparency is an important part of deepening understanding of how these systems work, both in terms of process and training data, the focus on source code for algorithms, a literal interpretation of the phrase "algorithmic transparency" offers flawed and unclear benefits. While in a limited context this may provide a small, highly technical audience with insights, it does little to change the experience of people online.

The first step is the ability to control whether an algorithm is shaping a person's online experience. Since 2018, Twitter has provided those using the service the option to turn off our Home timeline ranking algorithm, returning them to a reverse-chronological order of Tweets. This control enables transparency — people can see how the content appears in the two environments. In the long term, as we envision through our @bluesky project, this control will extend to the choice among ranking algorithms, built on an open standard for social media to which we hope Twitter will ultimately adhere. The idea of "Protocols not platforms" is instructive not only for the technological potential for standardisation of ranking algorithms but also for the underlying impact this would have on protecting free expression and driving competition.²

Additionally, in 2021 we announced our Responsible Machine Learning (RML) Initiative, which shared what Twitter has been doing to improve our ML algorithms within the platform, and our path forward through a company-wide initiative called Responsible ML.³ For Twitter, Responsible ML consists of: (1) Taking responsibility for our algorithmic decisions; (2) Equity and fairness of outcomes; (3) Transparency about our decisions and how we arrived at them; and (4) Enabling agency and algorithmic choice. Responsible technological use includes studying the effects it can have over time. When Twitter uses ML, it can impact hundreds of millions of Tweets per day and sometimes, a system designed to help can behave differently than was intended. These subtle shifts can then start to impact the people using Twitter, and so we want to make sure we're studying the changes and using them to build a better product.

This effort is part of our ongoing work to look at algorithms across a range of topics. We shared the findings of our analysis of bias in our image cropping algorithm and how they informed changes in our product.⁴ Our teams

² knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech

³ https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative

⁴ https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping



are also publishing learnings from an in-depth analysis of whether our recommendation algorithms amplify political content.⁵

This research study highlights the complex interplay between an algorithmic system and people using the platform. Algorithmic amplification is not problematic by default – all algorithms amplify. Algorithmic amplification is problematic if there is preferential treatment as a function of how the algorithm is constructed versus the interactions people have with it. Currently, our teams are exploring opportunities this work may unlock for future collaboration with external researchers looking to reproduce, validate and extend our internal research, and will continue to share additional updates.

(4) Competition, choice, and innovation are foundations of the Open Internet and should be protected and expanded, ensuring incumbents are not entrenched by laws and regulations.

A less competitive internet trends towards a less open internet. There's a risk that some regulatory interventions will undermine competition and entrench incumbent services, reducing consumer choice. It's not unique to the technology sector that incumbents often seek to use new regulations to protect their own market dominance, and just because services are online does not mean they depend on the Open Internet. Indeed, in some cases, a less open internet may suit certain businesses.

Competition in the online service space depends on a number of pillars, which are sometimes portrayed as only benefiting large providers. This framing is often misleading, given that these protections currently — and should continue to — benefit services of all sizes and are of most importance to those with fewer resources.

Intermediary liability protection is a foundation of the global, Open Internet and critical to competition online. Without this foundation, the internet as we know it — allowing speech, interaction, and discovery for billions of people — would cease.

Policymakers should avoid mandating technical means of implementation that have the effect of further entrenching services based on those tools and technologies, or of benefiting those that have the financial and technical means to deploy the particular implementation proposed, not to mention the vendors promising a simple solution. Opportunities to expand interoperability and the adoption of open standards will empower people with greater choice and flexibility about how they interact with online services and drive competition.

Finally, the technologies that underpin the ability to address and remove the most harmful content and respond to further harms remain in proprietary silos, becoming exponentially more effective as businesses scale, further enhancing dominance and undermining competition. Content moderation technology is one of the most significant barriers to entry, particularly as regulators set ever stricter requirements on the time within which harmful content must be removed. Policymakers should encourage and facilitate a fundamental change in the availability of proactive technologies, and the data that underpins them, to enable new services and tools to be made more accessible to a greater range of services, including by providing a robust legal framework for information sharing.

(5) Content moderation is more than just leave up or take down.

Legislation and regulation should allow for a range of interventions while setting clear definitions for categories of content they seek to address, with substantive definitions and boundaries and consistent with human rights standards.

Where the content at issue is lawful, but a government believes there's a need to intervene, the regulatory framework should clearly distinguish between these types of content. Government requests for the removal of

⁵ https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent



specific pieces of content based on illegality should be based upon legal process and provide for transparency about how these powers are used. It's a fundamental question of due process that a government agency, not a private actor, is responsible for determining criminality. Except in specific, exigent circumstances consistent with prevailing law, companies should be free to provide notice to people around the basis for actions taken against them or their accounts.

Secondly, we believe the regulatory debate needs to reflect how content moderation is now more than just leaving content up or taking it down. For example:

- Providing users with context, whether concerning an account; piece of content; or form of engagement, is more informative to the broader public conversation than removing content.
- Providing controls to people and communities to control their own experience is empowering and impactful.
- Equally, deamplification allows a more nuanced approach to types of speech that may be considered problematic, better striking a balance between freedom of speech and freedom of reach. In the long term, how peoples' attention is directed is a critical question.

Thirdly, regulatory frameworks that address system-wide processes, as opposed to individual pieces of content, will be able to better reflect the challenges of scale for all modern communications services. The approach is also more flexible and fit for purpose, recognising that the nature of challenges faced changes depending on whether you are seeking to protect a certain group, such as young people, or a particular type of behaviour, such as platform manipulation.⁶

As has been noted by a range of voices, the combination of significant administrative penalties for individual pieces of content and expected removal of content in short time periods — whether one hour or 24 hours — creates a significant corporate incentive to over-remove content, particularly in edge cases. This would more acutely impact small companies and new services that have more limited resources to litigate or pay fines. These frameworks must be underpinned with strong, independent processes that are free from political interference and allow for civil society participation.

To avoid incentivising over-removal, regulation that assesses the system-wide performance of how services enforce their terms of service will provide essential flexibility and reduce incentives to over-moderate content, while incentivising robust appeal mechanisms, and investment in technological solutions, despite the inevitable errors that come from imperfect tools.

Twitter's commitment to healthy conversations and public consultation

Abuse and harassment have no place on Twitter and our teams take action against harmful content under the [Twitter Rules](#). As an open service, our rules reflect the voice of the people who use Twitter. We think it's critical to consider global perspectives, as well as make our content moderation decisions easier to understand.

Twitter hasn't always been a place where everyone felt they could express themselves free from harassment or abuse. Other similar types of behaviour can also discourage people from expressing themselves. That's why, in recent years, we have strengthened our policies, expanded our partnerships, and grown our team dedicated to developing and building the products and features that empower healthier conversations.⁷

Our aim is to have comprehensive policies that appropriately balance fundamental human rights and consider the global context in which we operate. We have recognised the role of deep consultation to appropriately address the complexity of online safety issues. Through our policies, products, and partnerships, we have undertaken coordinated efforts to consult with a range of partners, human rights experts, civil society

⁶ blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

⁷ <https://blog.twitter.com/common-thread/en/topics/stories/2021/about-common-thread>



organisations, academics, and the general public, whose feedback is reflected in revisions to the policy frameworks that govern the Twitter platform.

For example, we have called for public feedback in the development of a number of policies, including our synthetic and manipulated media policy,⁸ our approach to world leaders,⁹ and our hateful conduct policy.¹⁰

With regards to our hateful conduct policy, we are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalised.¹¹ The policy makes clear that no one on Twitter may promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.¹² We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

Over the past two years, we conducted a multi-stage consultation process, to expand this policy and encompass the evolving nature of conversations online. We engaged multiple global stakeholders, including partners in Australia, on the phased updates.¹³ Through this process we continued to expand on the policy and prohibit behaviour that targets individuals or groups with abuse based on their perceived membership in a protected category. The consultation sought a variety of perspectives and voices when considering how to balance harm reduction and freedom of expression, as well as how to avoid any unintended consequences of removing legitimate speech from marginalised groups.

Conversely, Australia continues to utilise a very limited definition of hate speech under the *Racial Discrimination Act 1975* (Cth) that is limited to race-based speech or behaviour, and does not include a number of the aforementioned categories, including sexual orientation, disability-based, religious-based or gender-based speech.¹⁴

We want the Twitter Rules and Terms of Service (**TOS**) to ensure all people can participate in the public conversation freely and safely. We take our role in promoting healthy conversation seriously, and we will continue to constantly work towards making Twitter a space that's safe for people to talk about what's happening.

Anonymity and pseudonymity on Twitter

There's a common misconception that trust begins with knowing who a person 'really' is, starting with knowing their real name. But trust isn't that simple. A recent study conducted by Twitter, among people using the service in the US, found that people assess trust based on multiple signals. For example, knowing the account is a genuine person and not a bot. Recently, we also rolled out new labels that identify bots with an 'automated' designation on their profile, an icon of a robot, and a link to the Twitter handle of the person who created the bot. This helps to understand how interpersonal trust is developed on Twitter and how automated accounts could affect that trust.¹⁵

Twitter works to prevent spam and fake accounts from harassing other people on the service both at the sign-up stage so they won't be able to join, and by removing accounts that have been proven to cause trouble. Twitter detects roughly 25 million accounts per month suspected of being automated or spam accounts. In fact, in the second half of 2020, it deployed 143 million anti-spam challenges to accounts, which helped bring spam reports

⁸ https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback

⁹ https://blog.twitter.com/en_us/topics/company/2021/calling-for-public-input-on-our-approach-to-world-leaders

¹⁰ https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate

¹¹ Twitter Help Centre, Hateful Conduct, 2021 [online] Help.twitter.com. Available at: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy> <Accessed 6 August 2021>.

¹² <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.html>

¹³ https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate

¹⁴ <https://www.legislation.gov.au/Details/C2014C00014>

¹⁵ <https://blog.twitter.com/common-thread/en/topics/stories/2021/the-secret-world-of-good-bots>



— those coming from people who flag Tweets as spam — down by about 18% from the first half of the year.¹⁶ People can also use tools such as muting or blocking to feel more in control of their experience.

Trust is also derived from the content someone posts and how they connect with others.¹⁷ Sharing a personal connection with a fellow person on Twitter, even if that person is posting anonymously, can still stoke empathy. Trust is based on what you do, not who you are. Anonymity provides space for more people to express themselves freely and safely, in ways that actually engender that sort of trust-building connection. Simply put, in addition to providing safety, anonymity and pseudonymity provides people with the agency and control to choose how they present themselves. This has been a core tenet of the internet since its inception and is essential to a society that promotes individual choice and freedoms.

Various academic studies have already shown that anonymity alone does not lead to harassment. In many cases people post harassing, toxic replies under their real name, with a photo of their real face.¹⁸ Research into why people might harass others on the internet, and how to most effectively tackle the behaviour, continues to evolve, including by deepening our understanding of the social contexts these problems exist within. Currently, there is not conclusive evidence that requiring the display of names and identities will reduce social problems, and many studies have documented the problems it actually creates, like posing real threats to vulnerable communities.¹⁹

Additionally, the potential consequences that digital identification policies might have on participation, access, and the widening of the ‘digital divide’ should be considered. Personal identification can pose risks to vulnerable groups who are not able to safely use services under their real name, such as those seeking information or support for domestic violence, whistleblowers, or LGBTQIA+ people. The requirement for digital identification verification could also cause inadvertent repercussions in people accessing services on the internet. According to the World Bank, an estimated 1 billion people worldwide do not have an official form of identification.²⁰ It is often the marginalised, vulnerable, and impoverished who lack government-issued IDs, leading to larger inequalities amongst people being able to access online services.

Twitter’s Trust & Safety Council, partnerships, and mental health

As we work to improve the health of the public conversation, we’re committed to reaching beyond Twitter’s virtual walls to integrate diverse perspectives that make our service better for everyone.²¹ That’s why we regularly collaborate with trusted partners, including on our [Trust & Safety Council](#), to develop products and programs, and to improve the Twitter Rules.

We know the best version of Twitter is the one that people who use it help build. Over the past year, we’ve engaged with the Trust and Safety Council on thirteen projects early in the development process. We distilled and put to use their feedback on ways we can offer a better and safer experience for people using Twitter. Their feedback directly informed our approach on several products, including:

- **Communities:** We incorporated feedback on the need to manage expectations on the role that moderators play by limiting the number of responsibilities and building tools to help them manage potential harassment.²²

¹⁶ <https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jul-dec>

¹⁷ <https://blog.twitter.com/common-thread/en/topics/stories/2021/whats-in-a-name-the-case-for-inclusivity-through-anonymity>

¹⁸ <https://jilliancyork.com/2021/01/14/everything-old-is-new-part-2-why-online-anonymity-matters/>

¹⁹ <https://coralproject.net/blog/the-real-name-fallacy/>

²⁰ <https://id4d.worldbank.org/global-dataset>

²¹ https://blog.twitter.com/en_us/topics/company/2021/our-continued-collaboration-with-trusted-partners

²² https://blog.twitter.com/en_us/topics/product/2021/testing-communities



- **Tips:** We incorporated feedback on the need to emphasize that the people on our service are responsible for transactions in a user-friendly way by asking people to agree to terms of service when enabling the feature.²³
- **Safety Mode:** To mitigate the effect on limiting counter-speech, a concern raised by the council particularly for people in positions of power, we decided to automatically time out interventions for seven days.²⁴
- **Conversation settings:** We started testing a notification that reminds people they can change who can reply to their Tweets to increase awareness and adoption as a direct recommendation from the council.²⁵
- **Parental resources:** Working with partners, we developed a Digital Safety Playbook to help parents learn about the tools available to help them feel safer, be in control, and manage digital footprints.²⁶
- **Education assets:** Working with UNESCO, we developed digital assets that teach how Twitter can be used in the classroom and help people get UNESCO's guidance on the best practices for media and information literacy.²⁷
- **UN Envoy on Youth:** In partnership with the UN, we developed an accessible booklet about digital safety and online protection for young people with a checklist produced in collaboration with the Office of the UN Secretary-General's Envoy on Youth.²⁸
- **Digital Safety Playbook:** In December 2021, we launched a consolidated Twitter safety guide for NGOs and others to reference and share with their community members - especially women journalists and youth. We developed a visually inspired, one-stop resource covering select Twitter tools that were designed to help people on Twitter feel safer, more in control, and empowered to manage their digital footprint.

We also worked with a number of partners to build on our mental health resources and support. As we continue to grapple with the weight and broad reaching effects of an unprecedented public health crisis through the pandemic, it is our job to ensure that Twitter remains a safe space for anyone interested in mental health tips and resources, or opening up about their individual mental health concerns.

Recently for World Suicide Prevention Day, we worked with various mental health partners across the globe to raise awareness and encourage honest conversation around the emotional challenges experienced amid the unprecedented COVID-19 crisis. We've consistently and continuously expanded our work with NGOs focused on mental health. In particular, we've continued to engage suicide prevention organisations and counseling services to ensure that people on Twitter feel safe and have access to support when they need it most.²⁹

In Australia, our longstanding partnerships with Beyond Blue and Lifeline Australia have helped us provide support for people that may be at risk or experiencing harmful thoughts. These partners are featured in our #ThereIsHelp search prompt initiative – a notification service that provides valuable mental health information and resources via Twitter and email.³⁰ 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 and a button that can quickly and easily connect them to information from support services.

Consistent with our commitment to provide information from authoritative sources and support services when people on Twitter need it most, in 2020, we also launched a #KnowtheFacts search notification prompt for COVID-19 to combat misinformation and provide credible resources when people are looking for information related to the pandemic.³¹ Also, building on our #ThereIsHelp notification for mental health and suicide prevention, we launched a dedicated domestic violence search prompt in partnership with 1800RESPECT to

²³ https://blog.twitter.com/en_us/topics/product/2021/bringing-tips-to-everyone

²⁴ https://blog.twitter.com/en_us/topics/product/2021/introducing-safety-mode

²⁵ https://blog.twitter.com/en_us/topics/product/2020/new-conversation-settings-coming-to-a-tweet-near-you

²⁶ <https://about.twitter.com/content/dam/about-twitter/en/tfg/download/twitter-digital-safety-playbook.pdf>

²⁷ <https://about.twitter.com/content/dam/about-twitter/en/tfg/download/teaching-learning-with-twitter-unesco.pdf>

²⁸ <https://about.twitter.com/content/dam/about-twitter/en/tfg/download/staying-safe-with-twitter-youth-activist-checklist.pdf>

²⁹ https://blog.twitter.com/en_us/topics/company/2020/amplifying-suicideprevention-resources-on-twitter

³⁰ https://blog.twitter.com/en_au/topics/company/2019/Together-for-a-Better-Internet-Twitter-supports-SaferInternetDay-2019

³¹ https://blog.twitter.com/en_au/topics/company/2020/helping-you-find-reliable-public-health-information-on-twitter



provide information to those in need to Australia's national support service for sexual assault, domestic, and family violence.³²

In partnership with Lifeline Australia, our #BeALifeline Twitter Direct Message (DM) Chatbot helps family and friends better support any loved ones who are going through a tough time.³³ The chatbot makes Lifeline's resources easily discoverable and acts as a touch point for primary caregivers to discreetly request support in times of need.

Building off these partnerships and feedback, we've also recently begun testing an overhauled reporting process that will make it easier for people to alert Twitter of harmful behaviour.³⁴ The new approach, which is currently being tested with a small group in the US, simplifies the reporting process and lifts the burden from the individual to interpret what violation of the Twitter Rules has taken place. Instead it asks them to describe what happened, adopting a symptoms-first approach. Twitter first asks the person what's occurred as a starting point to find out what's happening instead of asking the individual to immediately diagnose the issue.

By refocusing on the experience of the person reporting the Tweet, Twitter hopes to improve the quality of the reports we receive. The more first-hand information we can gather about how people are experiencing certain content, the more precise Twitter can be when it comes to addressing it or, ultimately, removing it. Thus, even if the Tweets in question don't technically violate any rules, it provides valuable input and context about conversational trends that can be used to improve peoples' experience on the service.

In 2022, as the new process begins to roll out to a wider audience, we will be working on improving its communication process, ensuring that those who are taking the time to report are notified of the outcome. The Twitter Rules should help keep everyone safe while balancing freedom of speech and promoting public conversation. We also want to ensure that if there are new issues emerging — ones that we may not have rules for yet — there is a globally applicable method for us to learn about them. The intention and goals of the reporting flow updates are to empower the people that use our service, give Twitter actionable information that can be used to improve product and experiences, and also to improve the overall trust and safety process.

Twitter's Developer Platform and Commitment to Transparency

Transparency is core to Twitter's approach. We are committed to transparency surrounding our efforts to tackle some of the biggest challenges we all face online.

For the last ten years since 2012, we have provided biannual Transparency Reports, which shine a light on our own practices, including enforcement of the Twitter Rules and our ongoing work to disrupt global state-backed information operations.³⁵ We want the general public and policy makers to be better informed about our actions, which is why our original report has evolved into a more comprehensive Twitter Transparency Center covering a broader array of our transparency efforts. We now include sections covering information requests, removal requests, copyright notices, trademark notices, email security, Twitter Rules enforcement, platform manipulation, and state-backed information operations.

Through initiatives, such as Twitter's open application programming interfaces (APIs), our developer platform, our information operations archive, and our disclosures in the Twitter Transparency Center and Lumen, we continue to support third-party research about what's happening on Twitter.³⁶

We'll continue to build on these efforts and inform the public as we improve Twitter in the open. In the past year, we have:

³² https://blog.twitter.com/en_au/topics/company/2019/our-work-to-combat-domestic-violence

³³ https://blog.twitter.com/en_au/topics/company/2018/Lifeline-launches-Twitter-DM-chatbot-to-help-BeALifeline

³⁴ <https://blog.twitter.com/common-thread/en/topics/stories/2021/twitters-new-reporting-process-centers-on-a-human-first-design>

³⁵ <https://transparency.twitter.com/>

³⁶ <https://transparency.twitter.com/>



- **Twitter API for Academic Research:** In early 2021, we launched a dedicated Academic Research product track on the new Twitter API giving qualified researchers access to the entire history of public conversation and elevated access to real-time data for free.³⁷
- **Algorithmic bias bounty challenge:** When we introduced our commitment to responsible machine learning, we also said, “the journey to responsible, responsive, and community-driven machine learning systems is a collaborative one.” That’s why we introduced the industry’s first algorithmic bias bounty competition to draw on the global ethical AI community’s knowledge of the unintended harms of saliency algorithms to expand our own understanding and to reward the people doing work in this field.³⁸
- **Twitter Moderation Research Consortium (TMRC):** We announced the creation of a new global expert group of academics, members of civil societies and NGOs, and journalists to study platform governance issues. We look forward to deeper analysis from this range of global experts and expect the collaboration to result in expanded disclosures beyond information operations to include sharing data in areas like misinformation, coordinated harmful activity, and safety.³⁹
- **Launch of an API curriculum:** “Getting started with the Twitter API for Academic Research” is now being used at universities, enabling students and teachers to learn how to use Twitter data for academic research. It is currently starred by over 200 academics on Github.⁴⁰
- **Developer Platform Academic Research advisory board:** This group of 12 scholars began work with our team this year to better understand how we can enhance the use of the Twitter API for academic research, while increasing meaningful dialogue between the Twitter Academic program and the academic community.⁴¹
- **Developer research highlights:** We published and continued to spotlight key research areas Twitter teams are working on today in an effort to inspire even more researchers to pursue these topics.⁴²

As we continue to invite trusted partners and the public to share feedback on ways to make Twitter safe, it’s important to be transparent about how we develop and enforce the Twitter Rules. Our newly formed Content Governance Initiative (CGI) aims to do this by developing a governance framework that provides a consistent and principled approach to the development, enforcement, and assessment of our global rules and policies. To build our governance framework, we’re engaging external stakeholders and have created an additional advisory group on our Trust and Safety Council.⁴³

We’ll continue collaborating with this group and cross-functional teams across Twitter to establish standardised guidelines on policy development, enforcement, and appeals that help drive a common understanding of Twitter’s approach to content moderation. The framework’s principles and guidelines will aim to fulfill the following objectives:

- Build legitimacy and trust through transparency and accountability.
- Deepen our commitment to good governance and human rights.
- Provide additional clarity on Twitter’s content moderation processes.
- Affirm our commitment to serving a diverse and inclusive global community.

We recognise that achieving these objectives will not be easy. Content moderation at scale is a highly complex and challenging process. This initiative reflects our ongoing commitment to working systematically — in partnership with external stakeholders around the world — to improve the transparency and consistency of our content moderation processes.

³⁷ https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api

³⁸ https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge

³⁹ https://blog.twitter.com/en_us/topics/company/2021/-expanding-access-beyond-information-operations-

⁴⁰ <https://github.com/twitterdev/getting-started-with-the-twitter-api-v2-for-academic-research>

⁴¹ https://blog.twitter.com/developer/en_us/topics/community/2021/introducing-the-developer-platform-academic-research-advisory-board

⁴² <https://developer.twitter.com/en/use-cases/do-research/academic-research/research-areas>

⁴³ <https://about.twitter.com/en/our-priorities/healthy-conversations/trust-and-safety-council>



Content Authenticity Initiative

The Coalition for Content Provenance and Authenticity (C2PA) has developed a draft technical specification for providing content provenance and authenticity for online content.⁴⁴ The C2PA Steering Committee members include Adobe, Intel, Microsoft, BBC and Twitter.

This specification is designed to enable global, opt-in, adoption of digital provenance techniques through the creation of a rich ecosystem of digital provenance enabled applications, for a wide range of individuals and organisations, while meeting appropriate security requirements. The specification can also be leveraged by regulatory bodies and governmental agencies to establish standards for digital provenance.

The C2PA draft specification is now open for public feedback.⁴⁵ Prior to developing this specification, the C2PA created its Guiding Principles that enabled it to remain focused on ensuring that the specification can be used in ways that respect privacy and personal control of data, with a critical eye toward potential abuse and misuse.⁴⁶ For example, that creators and publishers of the media assets always have control over whether provenance data is included, as well as what specific pieces of data are included.

Current online safety landscape in Australia

The *Online Safety Act 2021* (Cth) (**OSA**), which will come into force on 23 January 2022, includes five online content schemes that empower the eSafety Commissioner, and her office, to take action on online content and behaviours using a robust suite of tools, including, but not limited to, directing removal requests to service providers to take down content within 24 hours, utilise information and investigative powers, levy civil penalties, and issue end-user notices. The OSA also enables the enactment of an updated Restricted Access System (RAS), the formation of industry codes of conduct, the development of an Age Verification Roadmap by the eSafety Commissioner, and the creation of the Basic Online Safety Expectations (BOSE) enacted by the Minister for Communications, Urban Infrastructure, Cities and the Arts of Australia. Additionally, section 474.17(1) of the *Criminal Code 1995* (Cth) creates an offence of using a carriage service to menace, harass or offend another person.

Additionally, in February 2021, DIGI launched the Australian Code of Practice on Disinformation and Misinformation (**ACPD**), Twitter is both a member of DIGI and a signatory to the Code.⁴⁷ Signatories to the Code have agreed to safeguards to protect Australians from harmful misinformation online, including publishing and implementing policies on their approach, providing a way for users to report content that may violate those policies, implementing a range of scalable measures that reduce its spread and visibility online, and releasing annual transparency reports about those safeguards in order to improve public understanding of these challenges over time.

Given the legislative and regulatory landscape that is coming to shape with current and new legislation, we would like to recommend that the Committee conduct a review of the online safety space in Australia one-year from its initial report. This would allow for the Committee's recommendations to fully encompass the quantifiable impacts of the OSA on online harms and review the advancement of industry measures consistent with the OSA industry codes and the ACPDM. This would also allow the Committee to fully consider the digital landscape and tailor its recommendations to advance online safety for Australians in the context of the OSA.

⁴⁴ <https://contentauthenticity.org/blog/announcing-the-c2pa-draft-specification>

⁴⁵ <https://c2pa.org/public-draft/>

⁴⁶ <https://c2pa.org/principles/>

⁴⁷ <https://digi.org.au/>



Conclusion

Twitter is engaged in open dialogue with governments around the world as we seek to foster collaborative partnerships and continue to drive forward online safety solutions while protecting vital public expression. Across all areas, the investments Twitter has made to protect the health of the public conversation are now generating clear and tangible safety benefits for people who use our service.⁴⁸

The issues raised in the Committee's terms of reference are broad and complex, and cannot be fully explored in the time allotted for this inquiry. There's a desire to deal with the companies and issues that are most commonly in the headlines today, without sufficient consideration of how this will impact the future of the Internet or where different policy objectives might be creating contradictions. As the control of digital infrastructure is increasingly a focus of geopolitical action, these issues cannot be viewed in isolation. It is essential that there is a coordinated, multi-stakeholder strategy to respond to these threats and defend the free, secure, and global Open Internet.

With the range of factors that need to be considered to holistically advance online safety, we therefore ask for the timeline be extended for the Select Committee Inquiry into Social Media and Online Safety to allow for the effective introduction and implementation of the *Online Safety Act 2021* (Cth) and to ensure meaningful consultation with the community.

We also believe that wider efforts on promoting safe use of online services that are focused on bolstering the voices of non-governmental organisations and not-for-profits would facilitate the desired cooperation in the private sector. Many of these not-for-profit and non-governmental groups do critical work, and policy makers should continue to find ways to broaden support for community-led efforts and initiatives that promote best practices concerning the safe use of services.

Our work will never be complete as the threats we face constantly evolve. Going forward, we look forward to continuing to work collaboratively and in good faith with the Government, civil society, not-for-profits, academia, and industry to address online safety and work to create lasting global solutions to build a safer and Open Internet.

⁴⁸ Twitter Blog, 2021 [online] [blog.twitter.com](https://blog.twitter.com/en_us/topics/company/2019/health-update.html). Available at: <https://blog.twitter.com/en_us/topics/company/2019/health-update.html> [Accessed 7 December 2021].



Appendix

Twitter's product and policies

In addition to our guiding principles and work to combat negative activities platform-wide, Twitter has also invested in a suite of product and policy-focused solutions aimed at ensuring a safe experience for everyone who uses Twitter.

The Twitter Rules

Our Rules are in place to ensure all people can participate in the public conversation freely and safely. These policies are enforced for all people who use Twitter, and set the standard for content and behavior not permitted on the platform. These policies address: Violence, Terrorism/violent extremism, Child sexual exploitation, Abuse/harassment, Hateful conduct, Suicide or self-harm, Sensitive media, Illegal or regulated goods & services, Private information, Non-consensual nudity, Platform manipulation and spam, Civic integrity, Impersonation, Synthetic and manipulated media, and Copyright & trademark.⁴⁹

Product Features and Controls

We are continually investing in new ways to give people additional control over the conversations they start on Twitter, and have several features actively in use today.

- **Hidden Replies:** All people on Twitter have the capability to hide any replies to any of their Tweets that they deem abusive or irrelevant to the conversation. Learn more here. In August 2020 we released an API endpoint for this capability to allow our API Partners to build more automated ways to employ this feature.
- **Conversation Settings:** In August of 2020, we made new conversation settings available to everyone on Twitter, allowing people to have more control over the conversations they start. These conversation settings let everyone choose who can reply to their Tweets with three options: 1) everyone (standard Twitter, and the default setting), 2) only people they follow, or 3) only people they mention. As of March 2021, over 11 million people had applied conversation settings to 70 million conversations.

Beginning in March 2021, we made these capabilities available to our advertisers when they compose Tweets directly through Tweet Composer or through our Ads API. This update extends the ability to apply conversation settings to Promoted-only Tweets and to those that use our most popular ad formats, in addition to organic Tweets.

Private Conversations

Twitter is a public platform, and we work to ensure this open forum remains healthy through our policies and platform capabilities. Direct Messages, while private between the sender and recipients (up to a max of 50), are subject to the Twitter Rules, as are all individuals and content on Twitter. In a Direct Message conversation, when a participant reports another person, we will stop the violator from sending messages to the person who reported them. The conversation will also be removed from the reporter's inbox. We will review reports and action appropriately.

Transparency, Measurement, and Independent Accreditation

First published in July 2012, our biannual Twitter Transparency Report highlights trends in legal requests, intellectual property-related requests, Twitter Rules enforcement, platform manipulation, and email privacy best

⁴⁹ <https://help.twitter.com/en/rules-and-policies/enforcement-options>



practices. The report also provides insight into whether or not we take action on these requests. First published in July 2012, our biannual Twitter Transparency Report highlights trends in legal requests, intellectual property-related requests, Twitter Rules enforcement, platform manipulation, and email privacy best practices. The report also provides insight into whether or not we take action on these requests. In August 2020, we completely revamped these reports and consolidated them into a comprehensive Transparency Center. See our latest update [here](#).

Transparency in advertising is also a core belief for us. In December 2020, as part of our efforts to provide increased transparency to our partners, we made two announcements:

First, we committed to undergo the accreditation process across all four of the MRC's offered Accreditation Services: Viewability, Sophisticated Invalid Traffic Filtration, Audience Measurement and Brand Safety. We will prioritize the Brand Safety accreditation but believe that all four are critical in demonstrating our enduring commitment to transparency.

We also announced that following an extensive 5 month vetting process, we selected DoubleVerify and Integral Ad Science to be our preferred partners for providing independent reporting on the context in which ads appear on Twitter. This is an opportunity to build solutions that will give advertisers a better understanding of the types of content that appear adjacent to their ads, helping them make informed decisions to reach their marketing goals.

Twitter provides additional transparency into campaign performance through measurement solutions and third-party studies based on your objectives. Our goal is to empower advertisers with measurement solutions to help you understand how your campaigns help achieve your broader marketing and business goals.

In March 2021, Twitter successfully earned the Trustworthy Accountability Group (TAG) Brand Safety Certified Seal, which covers Twitter's global operations and was attained via an independent audit. Learn more about this certification [here](#).

Our Commitment to Health Over Time

Health has always been and will remain a top priority for Twitter, and our work is ever-evolving. Twitter has made significant improvements around online safety over the past several years. Below are a few notable highlights of changes we've made in the last few years:

2021⁵⁰

- December
 - In an effort to better support people using Twitter in getting the help and support they need, we began testing a new reporting flow.⁵¹ This updated process is aimed at ensuring that everyone feels safe and heard and at making it easier for people to report unhealthy or unwanted content.
 - In efforts to educate people about our safety tools, we started testing a notification to remind people that they can change who can reply after tweeting.⁵² Some people will see the notification when their Tweet is getting unexpected or unwanted attention on Android & iOS.⁵³
 - We published a Safety Playbook, available in French, Hindi, Japanese, Portuguese, Spanish and will continue to find ways to help people know about and use our safety tools when they need it most.⁵⁴
 - We made updates to our Hateful Conduct dehumanization policy to encompass all protected categories, which means that Twitter's prohibits language that dehumanizes others on the basis

⁵⁰ <https://business.twitter.com/en/help/ads-policies/brand-safety.html>

⁵¹ <https://twitter.com/TwitterSafety/status/1468287536787243010?s=20>

⁵² <https://twitter.com/TwitterSafety/status/1415025551773892608>

⁵³ <https://twitter.com/TwitterSupport/status/1466514564627980291>

⁵⁴ <https://twitter.com/TwitterForGood/status/1469370026218164225?s=20>



- of religion, caste, age, disability, disease, race, ethnicity, national origin, gender, gender identity, or sexual orientation. The content must display abusive intent to be in scope for the policy, defined as content posted to shame, harass or degrade someone.⁵⁵
- We began testing a new way for Tweet authors to indicate that one of their Tweets includes sensitive media. This functionality builds upon the ways in which people on Twitter or Twitter's enforcement teams can already place sensitive media warnings to Tweets. Twitter proactively prevents ad placement adjacent to Tweets that have been labeled as "Sensitive Media".
 - We disclosed an additional 3,465 accounts to our archive of accounts linked to state-linked information operations. We have been periodically making these disclosures since October 2018 and, this year, have shared relevant data about these operations with key independent research partners. We announced that we will be updating our approach for future disclosures, with the introduction of the Twitter Moderation Research Consortium. The TMRC, set to launch in early 2022, will bring together a global group of experts from across academia, civil society, NGOs, and journalism to study platform governance issues.
 - November
 - Beginning in early 2020, Twitter introduced labels to alert people to Tweets including potentially misleading information around synthetic and manipulated media, civic integrity and COVID-19 vaccine misinformation. In November, we introduced a new design for these labels, which has resulted in more people clicking into the labels to learn more, and fewer people Retweeting or liking potentially misleading Tweets with these labels.⁵⁶ We began testing new designs in July 2021 and learned that better labels meant fewer impressions on misinformation.⁵⁷
 - October
 - To give people more control over their followers and how they interact with others on Twitter, we launched a test that allows people to remove a follower without blocking them.
 - September
 - We began testing a feature that allows automated accounts to identify themselves to give people more context about who they're interacting with on Twitter.
 - In an effort to ensure that people are able to engage with the public conversation in safe and healthy ways, we began a public test of a new feature called Safety Mode. When someone on Twitter activates this feature, it autoblocks accounts for 7 days that may use harmful language or send repetitive, uninvited replies or mentions.
 - August
 - We began testing a new reporting flow in the United States, South Korea and Australia which allows people to report Tweets that seem misleading. The intention of this pilot is to better understand whether this is an effective approach to address misinformation on the platform. We plan to iterate on this workflow as we learn from our test.
 - To promote credible information about vaccines, we served a COVID-19 PSA at the top of people's Timelines in 14 global markets. These prompts push people to local information covering a wide range of topics relevant to that country including topics like vaccine safety, effectiveness, availability and distribution plans.
 - Stemming from growing concerns around the impact of certain types of ads on physical, mental health, and body image, particularly for minors, we updated our global advertising policies to include restrictions on weight loss content, particularly prohibiting the targeting of minors.
 - Twitter condemns racism in all its forms - our aim is to become the world's most diverse, inclusive, and accessible tech company, and lead the industry in stopping such abhorrent views being shared on our platform. We published a blog post detailing our analysis of the conversation around the Euro 2020 final and laying out the steps we put in place to quickly identify and remove racist, abusive Tweets targeting the England team, the wider Euros conversation and the football conversation in general.

⁵⁵ <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

⁵⁶ <https://twitter.com/TwitterSafety/status/1460716542182805504>

⁵⁷ <https://twitter.com/nickpickles/status/1460984540709675017>



- We announced new partnerships with @AP and @Reuters as one part of our ongoing efforts to help people understand the conversation happening on Twitter. People experience a range of public conversations on our service every day, and we're committed to continuing our work to elevate credible information and context.
- July
 - As part of our ongoing effort to improve Twitter's accessibility, we introduced captions for voice Tweets, allowing more people to join the conversation.
 - We announced that we signed an agreement with the Media Ratings Council (MRC) for the Brand Safety pre-assessment. This represents a milestone in our progress towards our commitment to earning all four of the MRC's accreditations in Viewability, Sophisticated Invalid Traffic Filtration, Audience Measurement and Brand Safety.
 - We released our latest update to the Twitter Transparency Center, inclusive of data from July 1 to December 31, 2020. As part of this release, we shared a new metric for the first time - impressions - which is the number of views violative Tweets received prior to removal. We found that impressions on violative Tweets accounted for less than 0.1% of all impressions of all Tweets during the reporting time frame and that 77% of these Tweets received fewer than 100 impressions prior to removal.
 - In an update to the conversation settings we introduced in August of 2020, we made it possible for people on Twitter to change who can reply to a Tweet after it has been Tweeted out. This tweak to the product is designed to give people more control over their conversations in overwhelming moments when their Tweets may be getting more attention than they previously anticipated.
 - Abuse and harassment disproportionately affect women and underrepresented communities online and our top priority is keeping everyone who uses Twitter safe and free from abuse. Following a year-long consultative process working alongside partner NGOs, Twitter committed to the Web Foundation's framework to end online gender-based violence as part of the @UN_Women #GenerationEquality initiative.
- June
 - In collaboration with key industry partners, Twitter released an open letter in response to the Digital Services Act, calling on the EU commission to protect the Digital Single Market, fair competition, and the Open Internet.
 - We updated the Twitter Help Center to more clearly articulate when we will take enforcement action moving forward on our Hateful Conduct & Abusive Behavior policies which prohibit abuse and harassment of protected categories, & cover a wide range of behaviors. Specifically, we do not permit the denial of violent events, including abusive references to specific events where protected categories were the primary victims. This policy now covers targeted and non-targeted content.
- May
 - Twitter engaged OpenSlate to provide third-party verification of the safety and suitability of the content in our Twitter Amplify offering. The study found that of the over 455,000 monetized videos analyzed, 100% fell above the industry-standard GARM Brand Safety Floor. They also found that 99.9% of analyzed videos were considered low risk, based on OpenSlate's proprietary video content categorization and the GARM Brand Suitability Framework.
 - For people on Twitter with English-language settings enabled, we introduced prompts that encourage people to pause and reconsider a potentially harmful or offensive reply before they hit send. We know that people come to Twitter to find, read about and discuss their interests and that sometimes when things get heated, people say mean things they might regret. In an effort to make Twitter a better place, when we detect potentially harmful or offensive Tweet replies, we'll prompt people and ask them to review their replies before Tweeting. This change comes after multiple tests resulting in people sending fewer potentially offensive replies across the service, and improved behavior on Twitter.
- April
 - Twitter testified before the United States Senate Judiciary Committee regarding our approach to responsible machine learning technology, focused on taking responsibility for our algorithmic



decisions, equity and fairness of outcomes, transparency about our decisions and enabling agency and algorithmic choice.

- We introduced an interstitial addressing COVID-19 vaccines at the top of people's timelines in 16 markets around the world as part of World Immunization Week. The prompts directed customers to market-specific information on vaccine safety, effectiveness and availability, ensuring access to credible sources and combatting public health misinformation.
- We introduced Twitter's first Global Impact Report, a cohesive representation of our work across corporate responsibility, sustainability, and philanthropy. We consider this report to be a big step in our commitment to sharing more about the work we know is important to the people we serve.
- March
 - We officially launched new Curated Categories within our Twitter Amplify offering in the US, the UK, Brazil and MENA. These categories are Twitter-curated sets of publishers that are bundled together around specific themes and they are designed to help Advertisers reach their audiences by aligning with brand safe, feel-good content.
 - We put out a call for responses to a public survey to help inform the future of our policy approach to world leaders. Politicians and government officials are constantly evolving how they use our service, and we look to our community to help us ensure that our policies remain relevant to the ever-changing nature of political discourse on Twitter and protect the health of the public conversation.
 - Twitter successfully earned the Trustworthy Accountability Group (TAG) Brand Safety Certified Seal, which covers Twitter's global operations and was attained via independent audit.
 - Following the launch of conversation settings for everyone on Twitter in August 2020, we made it possible for our advertisers to use conversation settings when they compose Tweets in our Ads Manager. This update extends the ability to apply conversation settings to Promoted-only Tweets and to those that use our most popular ad formats, in addition to organic Tweets.
 - We announced that moving forward we will apply labels to Tweets that may contain misleading information about COVID-19 vaccines, in addition to our continued efforts to remove the most harmful COVID-19 misleading information from the service. These changes are made in accordance with our COVID-19 policy which we expanded in December of 2020.
- February
 - We disclosed the removal of 373 accounts related to independent, state-affiliated information operations for violations to our platform manipulation policies. These operations were attributed to Armenia, Russia and a previously disclosed network from Iran.
- January
 - We further expanded our Hateful Conduct policy to prohibit inciting behavior that targets individuals or groups of people belonging to protected categories. This includes incitement of fear or spreading fearful stereotypes, incitement of harassment on or off platform, and incitement to deny economic support.
 - We launched a pilot for a community-driven approach to address misinformation on Twitter, which we're calling Birdwatch. In this pilot, we will allow a select group of participants in the United States identify Tweets they believe are misleading, write public notes to add context, and rate the quality of other participants' notes.
 - We updated the Twitter Transparency Center with data reflecting the timeframe of January 1, 2020 - June 30, 2020. We released a blog post highlighting the trends and insights surfaced in this latest disclosure, including the impact of COVID-19 during the reporting timeframe.
 - In the wake of the events at the US Capitol on January 6, we took unprecedented action to enforce our policies against Glorification of Violence. In light of these events, we took additional action to protect the conversation on our service from attempts to incite violence, organize attacks, and share deliberately misleading information about the election outcome.

2020⁵⁸

- December

⁵⁸ <https://business.twitter.com/en/help/ads-policies/brand-safety.html>



- As the world continues to fight the COVID-19 pandemic and prepare for the global distribution of vaccines, we announced that we will be expanding our COVID-19 policy. Moving forward, we may require people to remove Tweets which advance harmful false or misleading narratives about COVID-19 vaccinations and beginning in early 2021, we may label or place a warning on Tweets that advance unsubstantiated rumors, disputed claims, as well as incomplete or out-of-context information about vaccines.
- We announced that we've selected Integral Ad Science (IAS) and Double Verify (DV) to be Twitter's preferred partners for providing independent reporting on the context in which ads appear on Twitter.
- We have announced that we have committed to working with the Media Ratings Council (MRC) to begin the accreditation process across all four of their offered Accreditation Services: Viewability, Sophisticated Invalid Traffic Filtration, Audience Measurement, and Brand Safety.
- We expanded our hateful conduct policy to extend to Tweets which seek to dehumanize people on the basis of race, ethnicity and national origin.
- November
 - In the week following the 2020 US Elections, we shared some key statistics about the labels, warnings, and additional restrictions we applied to Tweets that included potentially misleading information about the US Election from October 27 to November 11:
 - Approximately 300,000 Tweets were labeled under our Civic Integrity Policy for content that was disputed and potentially misleading. These represent 0.2% of all US election-related Tweets sent during this time period.
 - 456 of those Tweets were also covered by a warning message and had engagement features limited (Tweets could be Quote Tweeted but not Retweeted, replied to or liked).
 - Approximately 74% of the people who viewed those Tweets saw them after we applied a label or warning message.
 - We saw an estimated 29% decrease in Quote Tweets of these labeled Tweets due in part to a prompt that warned people prior to sharing.
- October
 - Ahead of the 2020 US Elections, we implemented a slate of additional, significant product and enforcement updates aimed at increasing context and encouraging more thoughtful consideration before Tweets are amplified. These updates included:
 - In accordance with our expanded civic integrity policy, we announced that people on Twitter, including candidates for office, may not claim an election win before it is authoritatively called. Tweets which include premature claims will be labeled and will direct people to our official US election page. Additionally, Tweets meant to incite interference with the election process or with the implementation of election results, such as through violent action, will be subject to removal.
 - We introduced enhanced prompts and warnings on Tweets that feature misleading information including a prompt which provides credible information for people before they are able to amplify misleading messages. We also added additional warnings and restrictions on Tweets with a misleading information label from US political figures and US-based accounts with more than 100,000 followers, or that obtain significant engagement.
 - To encourage more thoughtful amplification of information on the platform, we implemented some temporary changes over the period surrounding the election. These changes include encouraging people to add their own commentary prior to amplifying content by prompting them to Quote Tweet instead of Retweet and only surfacing Trends in the "For You" tab in the United States that include additional context.
- September
 - We launched a new feature to prompt people to read news articles before they amplify them. This has resulted in people opening articles 40% more often after seeing the prompt and a 33% increase in people opening articles before they Retweet.
 - We expanded our Civic Integrity Policy to help us more effectively address attempts to abuse Twitter in a manner that could lead to suppression of voting and other harms to civic processes.



- We will now label or remove false or misleading information intended to undermine voter turnout and/or erode public confidence in an election or other civic process.
- Twitter is part of the inaugural group of companies to hold the Brand Safety Certified Seal from TAG (the Trustworthy Accountability Group) as part of their new TAG Brand Safety Certified Program. This indicates that Twitter meets all of the requirements of upholding an industry regulated framework for Brand Safety in the UK.
 - August
 - We introduced the Twitter Transparency Center which highlights our efforts across a broader array of topics than had previously been shared in our Twitter Transparency Reports. We now include intuitive, interactive sections covering information requests, removal requests, copyright notices, trademark notices, email security, Twitter Rules enforcement, platform manipulation, and state-backed information operations. We have also newly introduced reporting on actions broken out by both content type and geographic location.
 - In a step towards our goal of improving conversation health and ensuring that Twitter is a safe space for all people, we released new Tweet Settings designed to give people more control over their conversations by letting them choose who can reply to their Tweets.
 - We began labeling accounts belonging to state-affiliated media entities and official representatives of the US, UK, France, Russia, and China. We will also no longer amplify state-affiliated media accounts through our recommendation systems including on the home timeline, notifications, and search.
 - July
 - We expanded our policy to address links to websites that feature hateful conduct or violence. Our goal is to block links in a way that's consistent with how we remove Tweets that violate our rules, and reduce the amount of harmful content on Twitter from outside sources.
 - June
 - We made our latest disclosure of information on more than 30,000 accounts in our archive of state-linked information operations, the only one of its kind in the industry, regarding three distinct operations that we attributed to the People's Republic of China (PRC), Russia, and Turkey.
 - May
 - We began testing new settings that let you choose who can reply to your Tweet and join your conversation.
 - We introduced new labels and warning messages that provide additional context and information on some Tweets containing disputed or misleading information.
 - April
 - Twitter UK was certified against the IAB's Gold Standard v1.1. This certification reinforces our commitment to reduce ad fraud, improve the digital advertising experience, and increase brand safety within the UK market.
 - March
 - We further expanded our rules against dehumanizing speech to prohibiting language that dehumanizes on the basis of age, disability or disease.
 - We broadened our definition of harm to address content that goes directly against guidance on COVID-19 from authoritative sources of global and local public health information.
 - February
 - Informed by public feedback, we launched our policy on synthetic information and manipulated media, outlining how we'll treat this content when we identify it.
 - January
 - We launched a dedicated search prompt intended to protect the public conversation and help people find authoritative health information around COVID-19. This work is constantly evolving, and the latest information can be found here.

2019⁵⁹

⁵⁹ <https://business.twitter.com/en/help/ads-policies/brand-safety.html>



- December
 - We launched the Twitter Privacy Center to provide more clarity around what we're doing to protect the information people share with us. We believe companies should be accountable to the people that trust them with their personal information, and responsible not only to protect that information but to explain how they do it.
- November
 - We made the decision to globally prohibit the promotion of political content. We made this decision based on our belief that political message reach should be earned, not bought.
 - We launched the option to hide replies to Tweets to everyone globally.
 - Twitter became certified against the DTSG Good Practice Principles from JICWEBS.
 - We asked the public for feedback on a new rule to address synthetic and manipulated media.
- October
 - We clarified our principles & approach to reviewing reported Tweets from world leaders.
 - We published our most recent Transparency Report covering H1 2019.
 - We launched 24/7 internal monitoring of trending topics to promote brand safety on search results.
- August
 - We updated our advertising policies to reflect that we would no longer accept advertising from state-controlled news media entities.
- July
 - Informed by public feedback, we launched our policy prohibiting dehumanizing speech on the basis of religion.
- June
 - We joined the Global Alliance for Responsible Media at Cannes.
 - We refreshed our Rules with simple, clear language, paring down from 2,500 words to under 600.
 - We clarified our criteria for allowing certain Tweets that violate our rules to remain on Twitter because they are in the public's interest.
- April
 - We shared an update on our progress towards improving the health of the public conversation, one year after declaring it a top company priority.

2018⁶⁰

- October
 - We released all of the accounts and related content associated with potential information operations that we found on our service since 2016. This was our first of many disclosures we've since made for our public archive of state-backed information operations.
- September
 - We asked the public for feedback on an upcoming policy expansion around dehumanizing speech, and took this feedback into consideration to update our rules.
- May
 - We made the decision to exclude accounts we suspect may be automated from monetizable audiences, meaning we do not serve ads to these accounts. Learn more about how we identify automated accounts here.
- March
 - We launched 24/7 human review of all monetized publisher content for Amplify Pre-Roll, along with an all-new Brand Safety policy for the program.
 - Co-founder and former CEO Jack Dorsey publicly announced our commitment and approach to making Twitter a safer place.

Online Safety Policies and Enforcement Actions

⁶⁰ <https://business.twitter.com/en/help/ads-policies/brand-safety.html>



Most people don't break our rules, but if they do, we have a range of ways to take action under the Twitter Rules. People are at the core of everything we do. People tell us what they care about, and what matters to them most. They use their voices to impact our businesses and the world around us, and as the public conversation continues to grow exponentially, we need to find more ways to support our community.

The social media landscape can be a safer place by serving the public conversation and creating an atmosphere in which everyone feels comfortable, secure, and confident enough to share their voice.

And while Twitter's current mission already recognizes that focusing on people is our priority, we still have work to do. Through our policies, products, and partnerships, we're already seeing progress happen, and we'll continue to invest in brand safety to make Twitter a safer place for all. Here's how we're doing it.

Prohibition against threats or glorification of violence

You may not threaten violence against an individual or a group of people. We also prohibit the glorification of violence. Healthy conversation is only possible when people feel safe from abuse and don't resort to using violent language. For this reason, we have a policy against threatening violence on Twitter. We define violent threats as statements of an intent to kill or inflict serious physical harm on a specific person or group of people.⁶¹

We recognize that some people use violent language as part of hyperbolic speech or between friends, so we also allow some forms of violent speech where it's clear that there is no abusive or violent intent, e.g., "I'll kill you for sending me that plot spoiler!". This policy is enforced in tandem with our policies on abusive behavior⁶² and hateful conduct.⁶³ Statements that express a wish or hope that someone experiences physical harm, making vague or indirect threats, or threatening actions that are unlikely to cause serious or lasting injury are not actionable under this policy, but may be reviewed and actioned under those policies.

Prohibition against terrorism or violent extremism

There is no place on Twitter for violent organizations, including terrorist organizations, violent extremist groups, or individuals who affiliate with and promote their illicit activities.⁶⁴ The violence that these groups engage in and/or promote jeopardizes the physical safety and well-being of those targeted. Our assessments under this policy are informed by national and international terrorism designations, as well as our violent extremist group and violent organizations criteria. We will immediately and permanently suspend any account that we determine to be in violation of this policy.

Under this policy, you can't affiliate with and promote the illicit activities of a terrorist organization or violent extremist group. Examples of the types of content that violate this policy include, but are not limited to:

- engaging in or promoting acts on behalf of a violent organization;
- recruiting for a violent organization;
- providing or distributing services (e.g., financial, media/propaganda) to further a violent organization's stated goals; and
- using the insignia or symbol of violent organizations to promote them or indicate affiliation or support.

We may make limited exceptions for groups that have reformed or are currently engaging in a peaceful resolution process, as well as groups with representatives who have been elected to public office through democratic elections. We may also make exceptions related to the discussion of terrorism or extremism for clearly educational or documentary purposes. This policy also doesn't apply to state or governmental organizations.

⁶¹ <https://help.twitter.com/en/rules-and-policies/violent-threats-glorification>

⁶² <https://help.twitter.com/rules-and-policies/abusive-behavior>

⁶³ <https://help.twitter.com/rules-and-policies/hateful-conduct-policy>

⁶⁴ <https://help.twitter.com/en/rules-and-policies/violent-groups>



Zero tolerance for child sexual exploitation

We have zero tolerance for child sexual exploitation on Twitter. Twitter has zero tolerance towards any material that features or promotes child sexual exploitation, one of the most serious violations of the Twitter Rules.⁶⁵ This may include media, text, illustrated, or computer-generated images. Regardless of the intent, viewing, sharing, or linking to child sexual exploitation material contributes to the re-victimization of the depicted children. This also applies to content that may further contribute to victimization of children through the promotion or glorification of child sexual exploitation. For the purposes of this policy, a minor is any person under the age of 18.

Twitter also has a long-standing collaboration with the National Center for Missing and Exploited Children (NCMEC). We are active members of several coalitions, such as the Technology Coalition, the ICT Coalition, the WeProtect Global Alliance, INHOPE and the Fair Play Alliance, that bring companies and NGOs together to develop solutions that disrupt the exchange of child sexual abuse materials online and prevent the sexual exploitation of children.

Prohibition against abuse/harassment

People may not engage in the targeted harassment of someone, or incite other people to do so, on Twitter. We consider abusive behavior an attempt to harass, intimidate, or silence someone else's voice.

We believe people should feel safe expressing their unique point of view. We believe in freedom of expression and open dialogue, but that means little as an underlying philosophy if voices are silenced because people are afraid to speak up. In order to facilitate healthy dialogue on the platform, and empower individuals to express diverse opinions and beliefs, we prohibit behavior that harasses or intimidates, or is otherwise intended to shame or degrade others. In addition to posing risks to people's safety, abusive behavior may also lead to physical and emotional hardship for those affected.

Some Tweets may seem to be abusive when viewed in isolation, but may not be when viewed in the context of a larger conversation. When we review this type of content, it may not be clear whether it is intended to harass an individual, or if it is part of a consensual conversation. To help our teams understand the context of a conversation, we may need to hear directly from the person being targeted, to ensure that we have the information needed prior to taking any enforcement action.

We will review and take action against reports of accounts targeting an individual or group of people with any of the following behavior within Tweets or Direct Messages. For accounts engaging in abusive behavior on their profile, please refer to our abusive profile policy. For behavior targeting people based on their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease, this may be in violation of our hateful conduct policy.

Violent threats

We prohibit content that makes violent threats against an identifiable target. Violent threats are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm, where an individual could die or be significantly injured, e.g., "I will kill you." We have a zero tolerance policy against violent threats. Those deemed to be sharing violent threats will face immediate and permanent suspension of their account.

Wishing, hoping, or calling for serious harm on a person or group of people

We do not tolerate content that wishes, hopes, promotes, incites, or expresses a desire for death, serious bodily harm or serious disease against an individual or group of people. This includes, but is not limited to:

- Hoping that someone dies as a result of a serious disease e.g., "I hope you get cancer and die."

⁶⁵ <https://help.twitter.com/en/rules-and-policies/sexual-exploitation-policy>



- Wishing for someone to fall victim to a serious accident e.g., “I wish that you would get run over by a car next time you run your mouth.” Saying that a group of individuals deserves serious physical injury e.g., “If this group of protesters don’t shut up, they deserve to be shot.”

About wishes of harm exceptions on Twitter

We recognize that conversations regarding certain individuals credibly accused of severe violence may prompt outrage and associated wishes of harm. In these limited cases, we will request the user to delete the Tweet without any risk of account penalty, strike, or suspension. Examples are, but not limited to “I wish all rapists to die,” or “Child abusers should be hanged.”

Unwanted sexual advances

While some consensual nudity and adult content is permitted on Twitter, we prohibit unwanted sexual advances and content that sexually objectifies an individual without their consent. This includes, but is not limited to:

- sending someone unsolicited and/or unwanted adult media, including images, videos, and GIFs;
- unwanted sexual discussion of someone’s body;
- solicitation of sexual acts; and
- any other content that otherwise sexualizes an individual without their consent.

Using insults, profanity, or slurs with the purpose of harassing or intimidating others

We take action against the use of insults, profanity, or slurs to target others. In some cases, such as (but not limited to) severe, repetitive usage of insults or slurs where the primary intent is to harass or intimidate others, we may require Tweet removal. In other cases, such as (but not limited to) moderate, isolated usage of insults and profanity where the primary intent is to harass or intimidate others, we may limit Tweet visibility as further described below. Please also note that while some individuals may find certain terms to be offensive, we will not take action against every instance where insulting terms are used.

Encouraging or calling for others to harass an individual or group of people

We prohibit behavior that encourages others to harass or target specific individuals or groups with abusive behavior. This includes, but is not limited to; calls to target people with abuse or harassment online and behavior that urges offline action such as physical harassment.

Denying mass casualty events took place

We prohibit content that denies that mass murder or other mass casualty events took place, where we can verify that the event occurred, and when the content is shared with abusive intent. This may include references to such an event as a “hoax” or claims that victims or survivors are fake or “actors.” It includes, but is not limited to, events like the Holocaust, school shootings, terrorist attacks, and natural disasters.

Consequences

When determining the penalty for violating any of these policies, we consider a number of factors including, but not limited to, the severity of the violation and an individual’s previous record of rule violations. The following is a list of potential enforcement options for content that violates this policy:

- Downranking Tweets in replies, except when the user follows the Tweet author.
- Making Tweets ineligible for amplification in Top search results and/or on timelines for users who don’t follow the Tweet author.
- Excluding Tweets and/or accounts in email or in-product recommendations.
- Requiring Tweet removal.



For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent suspension. Suspending accounts whose primary use we've determined is to engage in abusive behavior as defined in this policy, or who have shared violent threats. If someone believes their account was suspended in error, they can submit an appeal.

Hateful conduct policy

Twitter's mission is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers. Free expression is a human right – we believe that everyone has a voice, and the right to use it. Our role is to serve the public conversation, which requires representation of a diverse range of perspectives.

We recognize that if people experience abuse on Twitter, it can jeopardize their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. This includes; women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities. For those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature and more harmful.

We are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalised. For this reason, we prohibit behavior that targets individuals or groups with abuse based on their perceived membership in a protected category.

Under this policy, people may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.⁶⁶ We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

People may also not use hateful images or symbols in their profile image or profile header. Users also may not use their username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.

Suicide or self-harm policy

At Twitter, we recognize that suicide and self-harm are significant social and public health challenges that require collaboration between all stakeholders – public, private, and civil society – and that we have a role and responsibility to help people access and receive support when they need it.

When developing this policy, we consulted extensively with experts to ensure that people who have engaged in self-harm or experienced suicidal thoughts can share their personal experiences. We also recognized the need to protect people from the potential harm caused by exposure to content that could promote or encourage self-harm – intentionally or inadvertently. That's why our policy prohibits content that promotes or encourages self-harming behaviors and provides support to those undergoing experiences with self-harm or suicidal thoughts.

Under this policy, you can't promote, or otherwise encourage, suicide or self-harm.⁶⁷ We define promotion and encouragement to include statements such as "the most effective", "the easiest", "the best", "the most successful", "you should", "why don't you". Violations of this policy can occur via Tweets, images or videos, including live video.

⁶⁶ <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.html>

⁶⁷ <https://help.twitter.com/en/rules-and-policies/glorifying-self-harm>



We define suicide to be the act of taking one's own life. We define self-harm to include:

- self-inflicted physical injuries e.g., cutting; and
- eating disorders e.g., bulimia, anorexia.

Violations of this policy include, but are not limited to:

- encouraging someone to physically harm or kill themselves;
- asking others for encouragement to engage in self-harm or suicide, including seeking partners for group suicides or suicide games; and
- sharing information, strategies, methods or instructions that would assist people to engage in self-harm and suicide.

Some examples of behavior that are not considered a violation of this policy include:

- sharing personal stories and experiences related to self-harm or suicide;
- sharing coping mechanisms and resources for addressing self-harm or suicidal thoughts; and
- discussions that are focused on research, advocacy, and education related to self-harm or suicide prevention.

It's important to note that while our policy does allow for people to share their personal experiences, but should avoid sharing detailed information about specific strategies or methods related to self-harm, as this could inadvertently encourage this behavior.

Sensitive media policy

People use Twitter to show what's happening in the world, often sharing images and videos as part of the conversation. Sometimes, this media can depict sensitive topics, including violent and adult content. We recognize that some people may not want to be exposed to sensitive content, which is why we balance allowing people to share this type of media with helping people who want to avoid it to do so.

For this reason, you can't include violent, hateful, or adult content within areas that are highly visible on Twitter, including in live video, profile, header, or List banner images.⁶⁸ If you share this content on Twitter, you need to mark your account as sensitive. Doing so places images and videos behind an interstitial (or warning message), that needs to be acknowledged before your media can be viewed. Using this feature means that people who don't want to see sensitive media can avoid it, or make an informed decision before they choose to view it.

Under this policy, there are also some types of sensitive media content that we don't allow at all, because they have the potential to normalize violence and cause distress to those who view them.

Anyone can report potential violations of this policy via our dedicated reporting flows. We group sensitive media content into the following categories: (1) graphic violence; (2) adult content; (3) violent sexual content; (4) gratuitous gore; and (5) hateful imagery.

(1) Graphic violence

Graphic violence is any media that depicts death, violence, medical procedures, or serious physical injury in graphic detail. Some examples include, but are not limited to, depictions of:

- violent crimes or accidents;
- physical fights;
- physical child abuse;
- bodily fluids including blood, feces, semen etc.;
- serious physical harm, including visible wounds; and
- severely injured or mutilated animals.

⁶⁸ <https://help.twitter.com/en/rules-and-policies/media-policy>



Some exceptions may be made for documentary or educational content.

(2) Adult content

Adult content is any consensually produced and distributed media that is pornographic or intended to cause sexual arousal. Some examples include, but are not limited to, depictions of:

- full or partial nudity, including close-ups of genitals, buttocks, or breasts (excluding content related to breastfeeding);
- simulated sexual acts; and
- sexual intercourse or other sexual acts – this also applies to cartoons, hentai, or anime involving humans or depictions of animals with human-like features.

Some exceptions may be made for artistic, medical, health, or educational content.

Pornography and other forms of consensually produced adult content are allowed on Twitter, provided that this media is marked as sensitive.⁶⁹ Doing so provides people who may not want to see this type of content with a warning that they will need to acknowledge before viewing your media.

For content that was created or distributed without the consent of those featured, this would violate our non-consensual nudity policy.⁷⁰ Under this policy, people can't post or share explicit images or videos that were taken, appear to have been taken or that were shared without the consent of the people involved. Examples of the types of content that violate this policy include, but are not limited to:

- hidden camera content featuring nudity, partial nudity, and/or sexual acts;
- creepshots or upskirts - images or videos taken of people's buttocks, up an individual's skirt/dress or other clothes that allows people to see the person's genitals, buttocks, or breasts;
- images or videos that superimpose or otherwise digitally manipulate an individual's face onto another person's nude body;
- images or videos that are taken in an intimate setting and not intended for public distribution; and
- offering a bounty or financial reward in exchange for intimate images or videos.

We will immediately and permanently suspend any account that we identify as the original poster of intimate media that was created or shared without consent. We will do the same with any account that posts only this type of content as well.

(3) Violent sexual conduct

Violent sexual conduct is any media that depicts violence, whether real or simulated, in association with sexual acts. Some examples include, but are not limited to, depictions of:

- rape and other forms of violent sexual assault, or sexual acts that occur without the consent of participants, including a simulated lack of consent; and
- sexualized violence – inflicting physical harm on an individual within an intimate setting, where it is not immediately obvious if those involved have consented to take part.

(4) Gratuitous gore

Gratuitous gore is any media that depicts excessively graphic or gruesome content related to death, violence or severe physical harm, or violent content that is shared for sadistic purposes. Some examples include, but are not limited to, depictions of:

- dismembered or mutilated humans;
- charred or burned human remains;
- exposed internal organs or bones; and

⁶⁹ https://twitter.com/settings/privacy_and_safety

⁷⁰ <https://help.twitter.com/en/rules-and-policies/intimate-media>



- animal torture or killing. Note: exceptions may be made for religious sacrifice, food preparation or processing, and hunting.

(5) Hateful imagery

Hateful imagery is any logo, symbol, or image that has the intention to promote hostility against people on the basis of race, religious affiliation, disability, sexual orientation, gender/gender identity or ethnicity/national origin. Some examples of hateful imagery include, but are not limited to:

- symbols historically associated with hate groups, e.g., the Nazi swastika;
- images depicting others as less than human, or altered to include hateful symbols, e.g., altering images of individuals to include animalistic features; or
- images altered to include hateful symbols or references to a mass murder that targeted a protected category, e.g., manipulating images of individuals to include yellow Star of David badges, in reference to the Holocaust.

Our aim is to limit exposure to sensitive images and videos and to prevent the sharing of potentially disturbing types of sensitive media. For this reason, under our sensitive media policy, we differentiate our enforcement approach depending on the type of media that has been shared and where it has been shared.

Graphic violence, adult content, and hateful imagery:

- you can't target people with unsolicited images or videos that contain graphic violence, adult content, or hateful imagery; and
- you can't include graphic violence, adult content, or hateful imagery within live video, profile, header, or List banner images.

Violent sexual conduct and gratuitous gore:

- We prohibit violent sexual conduct to prevent the normalization of sexual assault and non-consensual violence associated with sexual acts.
- We prohibit gratuitous gore content because research has shown that repeated exposure to violent content online may negatively impact an individual's wellbeing.
- For these reasons, you can't share images or videos that depict violent sexual conduct or gratuitous gore on Twitter. However, very limited exceptions may be made for gory media associated with newsworthy events.

You can share graphic violence and consensually produced adult content within your Tweets, provided that you mark this media as sensitive. We may also allow limited sharing of hateful imagery, provided that it is not used to promote a terrorist or violent extremist group, that you mark this content as sensitive and don't target it at an individual (via mentioning someone or including an identifiable individual within such images).

To mark your media as sensitive, navigate to your safety settings and select the Mark media you Tweet as containing material that may be sensitive option. If you don't mark your media as sensitive, we will do so manually if your content is reported for review.

Illegal or certain regulated goods or services

Twitter takes the safety of our users seriously. In an effort to address the potential for real-world harm, we do not allow the use of Twitter for any unlawful behavior or to further illegal activities.⁷¹ This includes selling, buying, or facilitating transactions in illegal goods or services, as well as certain types of regulated goods or services. In some cases, we may ask you to contact a law enforcement agency and have them contact us via our law enforcement request page to ensure we have enough context to enforce this policy. In addition to reports received, we proactively surface activity that may violate this policy for human review.

⁷¹ <https://help.twitter.com/en/rules-and-policies/regulated-goods-services>



Goods or services covered under this policy include, but are not limited to:

- counterfeit goods and services;⁷²
- drugs and controlled substances;
- human trafficking;
- products made from endangered or protected species;
- sexual services;
- stolen goods; and
- weapons, including firearms, ammunition, and explosives, and instructions on making weapons (e.g. bombs, 3D printed guns, etc.)

The consequences for violating this policy depends on the severity of the violation and the account's previous history of violations.

If you violate this policy more than once and/or if your account is dedicated to the sale of illegal or regulated goods and/or services, your account may be suspended permanently.

Accounts that appear to be using misleading account information in order to engage in spamming, abusive, or disruptive behavior to promote the sale of illegal and regulated goods and/or services may be subject to suspension under our platform manipulation and spam policy.⁷³

Privacy and private information

Sharing someone's private information online without their permission, sometimes called doxxing, is a breach of their privacy and of the Twitter Rules.⁷⁴ Sharing private information can pose serious safety and security risks for those affected and can lead to physical, emotional, and financial hardship.

When reviewing reports under this policy, we consider a number of things, including:

- What type of information is being shared?
 - We take this into consideration because certain types of private information carry higher risks than others, if they're shared without permission. Our primary aim is to protect individuals from potential physical harm as a result of their information being shared, so we consider information such as physical location and phone numbers to be a higher risk than other types of information.
- Who is sharing the information?
 - We also consider who is sharing the reported information and whether or not they have the consent of the person it belongs to. We do this because we know that there are times when people may want some forms of their personal information to be shared publicly. For example, sharing a personal phone number or email for professional networking or to coordinate social events or publicly sharing someone's home addresses to seek help after a natural disaster.
- Is the information available elsewhere online?
 - If the reported information was shared somewhere else before it was shared on Twitter, e.g., someone sharing their personal phone number on their own publicly accessible website, we may not treat this information as private, as the owner has made it publicly available. Note: we may take action against home addresses being shared, even if they are publicly available, due to the potential for physical harm.
- Why is the information being shared?
 - We also factor in the intent of the person sharing the information. For example, if we believe that someone is sharing information with an abusive intent, or to harass or encourage others to harass another person, we will take action. On the other hand, if someone is sharing information

⁷² <https://help.twitter.com/rules-and-policies/counterfeit-goods-policy>

⁷³ <https://help.twitter.com/rules-and-policies/platform-manipulation>

⁷⁴ <https://help.twitter.com/en/rules-and-policies/personal-information>



in an effort to help someone involved in a crisis situation like in the aftermath of a violent event, we may not take action.

Under this policy, you can't share the following types of private information, without the permission of the person who it belongs to:

- home address or physical location information, including street addresses, GPS coordinates or other identifying information related to locations that are considered private;
- identity documents, including government-issued IDs and social security or other national identity numbers – note: we may make limited exceptions in regions where this information is not considered to be private;
- contact information, including non-public personal phone numbers or email addresses;
- financial account information, including bank account and credit card details; and
- other private information, including biometric data or medical records.
- media of private individuals without the permission of the person(s) depicted.
- The following behaviors are also not permitted:
 - threatening to publicly expose someone's private information;
 - sharing information that would enable individuals to hack or gain access to someone's private information without their consent, e.g., sharing sign-in credentials for online banking services;
 - asking for or offering a bounty or financial reward in exchange for posting someone's private information;
 - asking for a bounty or financial reward in exchange for not posting someone's private information, sometimes referred to as blackmail.

With respect to private media, where individuals have a reasonable expectation of privacy in an individual piece of media, we believe they should be able to determine whether or not it is shared. Sharing private media could potentially violate users' privacy and may lead to emotional or physical harm. When we are notified by individuals depicted, or their authorized representative, that they did not consent to having media shared, we will remove the media. This policy is not applicable to public figures.

Anyone can report private information that has been shared in a clearly abusive way (whether they have a Twitter account or not). In cases where the information hasn't been shared with a clearly abusive intent, we need to hear directly from the owner of this information (or an authorized representative, such as a lawyer) before taking enforcement action.

When reporting private media, we need a first person or authorized representative report in order to make the determination that the image or video has been shared without their permission.

Authenticity

Platform manipulation and spam

We want Twitter to be a place where people can make human connections, find reliable information, and express themselves freely and safely. To make that possible, we do not allow spam or other types of platform manipulation. We define platform manipulation as using Twitter to engage in bulk, aggressive, or deceptive activity that misleads others and/or disrupts their experience.⁷⁵

Platform manipulation can take many forms and our rules are intended to address a wide range of prohibited behavior, including:

- commercially-motivated spam, that typically aims to drive traffic or attention from a conversation on Twitter to accounts, websites, products, services, or initiatives;
- inauthentic engagements, that attempt to make accounts or content appear more popular or active than they are;

⁷⁵ <https://help.twitter.com/en/rules-and-policies/platform-manipulation>



- coordinated activity, that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting; and
- coordinated harmful activity that encourages or promotes behavior which violates the Twitter Rules.

Under this policy we prohibit a range of behaviors in the following areas: (1) accounts and identity; (2) engagement and metrics; (3) misuse of Twitter product features; and (4)

(1) Accounts and identity

You can't mislead others on Twitter by operating fake accounts. This includes using misleading account information to engage in spamming, abusive, or disruptive behavior. Some of the factors that we take into consideration include:

- use of stock or stolen profile photos, particularly those depicting other people;
- use of stolen or copied profile bios; and
- use of intentionally misleading profile information, including profile location.

You can't artificially amplify or disrupt conversations through the use of multiple accounts or by coordinating with others to violate the Twitter Rules. This includes:

- overlapping accounts – operating multiple accounts with overlapping use cases, such as identical or similar personas or substantially similar content;
- mutually interacting accounts – operating multiple accounts that interact with one another in order to inflate or manipulate the prominence of specific Tweets or accounts; and
- coordination – creating multiple accounts to post duplicative content or create fake engagement, including:
 - posting identical or substantially similar Tweets or hashtags from multiple accounts you operate;
 - engaging (Retweets, Likes, mentions, Twitter Poll votes) repeatedly with the same Tweets or accounts from multiple accounts that you operate;
 - coordinating with or compensating others to engage in artificial engagement or amplification, even if the people involved use only one account; and
 - coordinating with others to engage in or promote violations of the Twitter Rules, including violations of our abusive behavior policy.

(2) Engagement and metrics

You can't artificially inflate your own or others' followers or engagement. This includes:

- selling/purchasing Tweet or account metric inflation – selling or purchasing followers or engagements (Retweets, Likes, mentions, Twitter Poll votes);
- apps – using or promoting third-party services or apps that claim to add followers or add engagements to Tweets;
- reciprocal inflation – trading or coordinating to exchange follows or Tweet engagements (including but not limited to participation in “follow trains,” “decks,” and “Retweet for Retweet” behavior); and
- account transfers or sales – selling, purchasing, trading, or offering the sale, purchase, or trade of Twitter accounts, usernames, or temporary access to Twitter accounts.

(3) Misuse of Twitter product features

You can't misuse Twitter product features to disrupt others' experience. This includes:

- Tweets and Direct Messages
 - sending bulk, aggressive, high-volume unsolicited replies, mentions, or Direct Messages;
 - posting and deleting the same content repeatedly;
 - repeatedly posting identical or nearly identical Tweets, or repeatedly sending identical Direct Messages; and



- repeatedly posting Tweets or sending Direct Messages consisting of links shared without commentary, so that this comprises the bulk of your Tweet/Direct Message activity.
- Following
 - “follow churn” – following and then unfollowing large numbers of accounts in an effort to inflate one’s own follower count;
 - indiscriminate following – following and/or unfollowing a large number of unrelated accounts in a short time period, particularly by automated means; and
 - duplicating another account’s followers, particularly using automation.
- Engagement
 - aggressively or automatically engaging with Tweets to drive traffic or attention to accounts, websites, products, services, or initiatives.
 - aggressively adding users to Lists or Moments.
- Hashtags
 - using a trending or popular hashtag with an intent to subvert or manipulate a conversation or to drive traffic or attention to accounts, websites, products, services, or initiatives; and
 - Tweeting with excessive, unrelated hashtags in a single Tweet or across multiple Tweets.
- URLs
 - publishing or linking to malicious content intended to damage or disrupt another person’s browser (malware) or computer or to compromise a person’s privacy (phishing); and
 - posting misleading or deceptive links; e.g., affiliate links and clickjacking links.

It is not a violation of this policy to use Twitter pseudonymously or as a parody, commentary, or fan account.

Additionally, posting links without commentary occasionally; coordinating with others to express ideas, viewpoints, support, or opposition towards a cause, provided such behavior does not result in violations of the Twitter Rules; and operating multiple accounts with distinct identities, purposes, or use cases does not violate this policy. These accounts may interact with one another, provided they don’t violate other rules.

Civic Integrity policy

The public conversation occurring on Twitter is never more important than during elections and other civic events. Any attempts to undermine the integrity of our service is antithetical to our fundamental rights and undermines the core tenets of freedom of expression, the value upon which our company is based.

We believe we have a responsibility to protect the integrity of those conversations from interference and manipulation. Therefore, we prohibit attempts to use our services to manipulate or disrupt civic processes, including through the distribution of false or misleading information about the procedures or circumstances around participation in a civic process.⁷⁶ In instances where misleading information does not seek to directly manipulate or disrupt civic processes, but leads to confusion on our service, we may label the Tweets to give additional context. Given the significant risks of confusion about key election information, we may take these actions even if Tweets contain (or attempt to contain) satirical or humorous elements.

Twitter considers civic processes to be events or procedures mandated, organized, and conducted by the governing and/or electoral body of a country, state, region, district, or municipality to address a matter of common concern through public participation. Some examples of civic processes may include: (1) Political elections; (2) Censuses; and (3) Major referenda and ballot initiatives.

This policy addresses 4 categories of misleading behavior and content: (1) Misleading information about how to participate; (2) Suppression and intimidation; (3) Misleading information about outcomes; and (4) False or misleading affiliation.

(1) Misleading information about how to participate

⁷⁶ <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>



We will label or remove false or misleading information about how to participate in an election or other civic process. This includes but is not limited to:

- misleading information about procedures to participate in a civic process (for example, that you can vote by Tweet, text message, email, or phone call in jurisdictions where these are not a possibility);
- misleading information about requirements for participation, including identification or citizenship requirements;
- misleading claims that cause confusion about the established laws, regulations, procedures, and methods of a civic process, or about the actions of officials or entities executing those civic processes; and
- misleading statements or information about the official, announced date or time of a civic process.

(2) Suppression and intimidation

We will label or remove false or misleading information intended to intimidate or dissuade people from participating in an election or other civic process. This includes but is not limited to:

- misleading claims that polling places are closed, that polling has ended, or other misleading information relating to votes not being counted;
- misleading claims about police or law enforcement activity related to voting in an election, polling places, or collecting census information;
- misleading claims about long lines, equipment problems, or other disruptions at voting locations during election periods;
- misleading claims about process procedures or techniques which could dissuade people from participating; and
- threats regarding voting locations or other key places or events (note that our violent threats policy may also be relevant for threats not covered by this policy).

(3) Misleading information about outcomes

We will label or remove false or misleading information intended to undermine public confidence in an election or other civic process. This includes but is not limited to:

- disputed claims that could undermine faith in the process itself, such as unverified information about election rigging, ballot tampering, vote tallying, or certification of election results; and
- misleading claims about the results or outcome of a civic process which calls for or could lead to interference with the implementation of the results of the process, e.g. claiming victory before election results have been certified, inciting unlawful conduct to prevent the procedural or practical implementation of election results (note that our violent threats policy may also be relevant for threats not covered by this policy).

(4) False or misleading affiliation

You can't create fake accounts which misrepresent their affiliation, or share content that falsely represents its affiliation, to a candidate, elected official, political party, electoral authority, or government entity. Read more about our parody, commentary, and fan account policy.

Not all false or untrue information about politics or civic processes constitutes manipulation or interference. In the absence of other policy violations, the following are generally not in violation of this policy:

- inaccurate statements about an elected or appointed official, candidate, or political party;
- organic content that is polarizing, biased, hyperpartisan, or contains controversial viewpoints expressed about elections or politics;
- discussion of public polling information;
- voting and audience participation for competitions, game shows, or other entertainment purposes; and



- using Twitter pseudonymously or as a parody, commentary, or fan account to discuss elections or politics.

Accurate reporting of suspected violations of this policy requires information and knowledge specific to an election or civic process. Therefore, we enable reporting of false or misleading information about civic processes in advance of major events, for people located in the relevant countries and locations. We also work with select government and civil society partners in these countries to provide additional channels for reporting and expedited review.

Twitter has worked extensively with the Australian Electoral Commission (AEC) and other government agencies through the Election Integrity Assurance Taskforce (EIAT) on past elections. For more detailed information regarding how these policies have been applied in practice in Australia, we have provided comprehensive submissions and evidence at past hearings held by the Australian Senate Select Committee on Foreign Interference through Social Media and the Australian Parliamentary Joint Standing Committee on Electoral Matters.

Impersonation policy

Impersonation is a violation of the Twitter Rules.⁷⁷ Twitter accounts that pose as another person, brand, or organization in a confusing or deceptive manner may be permanently suspended under Twitter's impersonation policy. Accounts with similar usernames or that are similar in appearance (e.g., the same profile image) are not automatically in violation of the impersonation policy. In order to violate our impersonation policy, the account must portray another entity in a misleading or deceptive manner. An account will not be removed if:

- The user shares your name but has no other commonalities, or
- The profile clearly states it is not affiliated with or connected to any similarly-named individuals or brands.

Twitter users are allowed to create parody, newsfeed, commentary, or fan accounts. Please refer to Twitter's parody, newsfeed, commentary, and fan account policy for more information about these types of accounts.⁷⁸

Synthetic and manipulated media policy

Users 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.⁷⁹

In order for content with misleading media (including images, videos, audios, gifs, and URLs hosting relevant content) to be labeled or removed under this policy, it must:

- Include media that is significantly and deceptively altered, manipulated, or fabricated, or
- Include media that is shared in a deceptive manner or with false context, and
- Include media likely to result in widespread confusion on public issues, impact public safety, or cause serious harm

We use the following criteria as we consider Tweets and media for labeling or removal under this policy as part of our ongoing work to enforce our rules and ensure healthy and safe conversations on Twitter:

1. Is the content significantly and deceptively altered, manipulated, or fabricated?

⁷⁷ <https://help.twitter.com/en/rules-and-policies/twitter-impersonation-policy>

⁷⁸ <https://help.twitter.com/en/rules-and-policies/parody-account-policy>

⁷⁹ <https://help.twitter.com/en/rules-and-policies/manipulated-media>



In order for content to be labeled or removed under this policy, we must have reason to believe that media are significantly and deceptively altered, manipulated, or fabricated. Synthetic and manipulated media take many different forms and people can employ a wide range of technologies to produce these media. Some of the factors we consider include:

- whether media have been substantially edited or post-processed in a manner that fundamentally alters their composition, sequence, timing, or framing and distorts their meaning;
- whether there are any visual or auditory information (such as new video frames, overdubbed audio, or modified subtitles) that has been added, edited, or removed that fundamentally changes the understanding, meaning, or context of the media;
- whether media have been created, edited, or post-processed with enhancements or use of filters that fundamentally changes the understanding, meaning, or context of the content; and
- whether media depicting a real person have been fabricated or simulated, especially through use of artificial intelligence algorithms
- We will not take action to label or remove media that have been edited in ways that do not fundamentally alter their meaning, such as retouched photos or color-corrected videos.

In order to determine if media have been significantly and deceptively altered or fabricated, we may use our own technology or receive reports through partnerships with third parties. In situations where we are unable to reliably determine if media have been altered or fabricated, we may not take action to label or remove them.

2. Is the content shared in a deceptive manner or with false context?

We also consider whether the context in which media are shared could result in confusion or suggests a deliberate intent to deceive people about the nature or origin of the content, for example, by falsely claiming that it depicts reality. We assess the context provided alongside media to see whether it provides true and factual information. Some of the types of context we assess in order to make this determination include:

- whether inauthentic, fictional, or produced media are presented or being endorsed as fact or reality, including produced or staged works, reenactments, or exhibitions portrayed as actual events;
- whether media are presented with false or misleading context surrounding the source, location, time, or authenticity of the media;
- whether media are presented with false or misleading context surrounding the identity of the individuals visually depicted in the media

We will not take action to label or remove media that have been shared with commentary or opinions that do not advance or present a misleading claim on the context of the media as listed above.

In order to determine if media have been shared in a deceptive manner or with false context, we may use our own technology or receive reports through partnerships with third parties. In situations where we are unable to reliably determine if media have been shared with false context, we will not label or remove the content.

3. Is the content likely to result in widespread confusion on public issues, impact public safety, or cause serious harm?

Tweets that share misleading media are subject to removal under this policy if they are likely to cause serious harm. Some specific harms we consider include:

- Threats to physical safety of a person or group
- Incitement of abusive behavior to a person or group
- Risk of mass violence or widespread civil unrest
- Risk of impeding or complicating provision of public services, protection efforts, or emergency response
- Threats to the privacy or to the ability of a person or group to freely express themselves or participate in civic events, such as:
- Stalking or unwanted and obsessive attention



- Targeted content that includes tropes, epithets, or material that aims to silence someone
- Voter suppression or intimidation
- We also consider the time frame within which the content may be likely to impact public safety or cause serious harm, and are more likely to remove content under this policy if immediate harm is likely to result.

Tweets with misleading media that are not likely to result in immediate harm but still have a potential to impact public safety, result in harm, or cause widespread confusion towards a public issue (health, environment, safety, civil rights and equality, immigration, and social and political stability) may be labeled to reduce their spread and to provide additional context.

While we have other rules also intended to address these forms of harm, including our policies on violent threats, civic integrity, COVID-19 misleading information, and hateful conduct, we will err toward removal in borderline cases that might otherwise not violate existing rules for Tweets that include misleading media.

In the absence of other policy violations, the following are generally not in violation of this policy:

- Memes or satire, provided these do not cause significant confusion about the authenticity of the media;
- Animations, illustrations, and cartoons, provided these do not cause significant confusion about the authenticity of the media.
- Commentary, reviews, opinions, and/or reactions. Sharing media with edits that only add commentary, reviews, opinions, or reactions allows for further debate and discourse relating to various issues and are not in violation of this policy.
- Counterspeech. We allow for direct responses to misleading information which seek to undermine its impact by correcting the record, amplifying credible information, and educating the wider community about the prevalence and dynamics of misleading information.
- Doctored or fake Tweets, social media posts, or chat messages. Due to the challenges associated with conclusively verifying whether an alleged Tweet, post, or message existed, we do not enforce on doctored or fake Tweets, social media posts, or chat messages under this policy.

We enforce this policy in close coordination with trusted partners, including our partnership with AP and Reuters, other news agencies, public health authorities, and governments.⁸⁰ Our team has open lines of communication with various partners to consult and get various media and claims reviewed.

In Australia, South Korea, and the US, Twitter has begun testing a new reporting feature that will allow users to report Tweets that seem misleading. As part of the experiment, the phrase “It’s misleading” will appear as an option when you select “Report an issue.”

Copyright and trademark policies

People may not violate others’ intellectual property rights, including copyright and trademark.

(1) Trademark policy

A trademark is a word, logo, phrase, or device that distinguishes a trademark holder’s good or service in the marketplace. Trademark law may prevent others from using a trademark in an unauthorized or confusing manner. Using another’s trademark in a way that may mislead or confuse people about your affiliation may be a violation of our trademark policy.⁸¹

⁸⁰ https://blog.twitter.com/en_us/topics/company/2021/bringing-more-reliable-context-to-conversations-on-twitter

⁸¹ <https://help.twitter.com/en/rules-and-policies/twitter-trademark-policy>



Twitter only investigates requests that are submitted by the trademark holder or their authorized representative e.g., a legal representative or other representative for a brand. You can submit a trademark report through our trademark report form.⁸²

If we determine that you violated our trademark policy, we may suspend your account. Depending on the type of violation, we may give you an opportunity to comply with our policies. In other instances, an account may be permanently suspended upon first review. If you believe that your account was suspended in error, you can submit an appeal.

(2) Copyright policy

Twitter responds to copyright complaints submitted under the Digital Millennium Copyright Act (“DMCA”). Section 512 of the DMCA outlines the statutory requirements necessary for formally reporting copyright infringement, as well as providing instructions on how an affected party can appeal a removal by submitting a compliant counter-notice.⁸³

Twitter will respond to reports of alleged copyright infringement, such as allegations concerning the unauthorized use of a copyrighted image as a profile or header photo, allegations concerning the unauthorized use of a copyrighted video or image uploaded through our media hosting services, or Tweets containing links to allegedly infringing materials. Note that not all unauthorized uses of copyrighted materials are infringements (see our fair use article for more information⁸⁴).

You can report alleged copyright infringement by visiting Twitter’s Help Center and filing a copyright complaint. If you are logged in to twitter.com, you can visit the Twitter Help Center directly from your Twitter account by clicking the ‘Help’ link located in the sidebar.

Filing a DMCA complaint is the start of a pre-defined legal process. Your complaint will be reviewed for accuracy, validity, and completeness. If your complaint has satisfied these requirements, we will take action on your request - which includes forwarding a full copy of your notice (including your name, address, phone and email address) to the user(s) who posted the allegedly infringing material in question.

If you are concerned about your contact information being forwarded, you may wish to use an agent to report for you.

Please be aware that under 17 U.S.C. § 512(f), you may be liable for any damages, including costs and attorneys’ fees incurred by us or our users, if you knowingly materially misrepresent that material or activity is infringing. If you are unsure whether the material you are reporting is in fact infringing, you may wish to contact an attorney before filing a copyright complaint.

How are claims processed?

We process copyright complaints in the order in which they are received. Once you've submitted your ticket, we will email you a ticket confirmation. If you do not receive a ticket confirmation that means we did not receive your complaint and you should re-submit your complaint. However, please note, submitting duplicate copyright complaints will result in a delay in processing.

If we decide to remove or disable access to the material, we will notify the affected user(s) and provide them with a full copy of the reporter’s complaint (including the provided contact information) along with instructions on how to file a counter-notice. We will also forward a redacted copy of the complaint to Lumen, with your personal information removed.

⁸² <https://help.twitter.com/forms/trademark>

⁸³ <https://help.twitter.com/en/rules-and-policies/copyright-policy>

⁸⁴ <https://help.twitter.com/en/rules-and-policies/fair-use-policy.html>



What information gets forwarded to the reported user(s)?

If we remove or disable access to the materials reported in a copyright complaint, the reported user(s) will receive a copy of the complaint, including the reporter's full name, email, street address, and any other information included in the complaint.

If you are uncomfortable sharing your contact information with the reported user(s), you may wish to consider appointing an agent to submit your DMCA notice on your behalf. Your agent will be required to submit the DMCA notice with valid contact information, and identify you as the content owner that they are representing.

What happens next?

Twitter's response to copyright complaints may include the removal or restriction of access to allegedly infringing material. If we remove or restrict access to user content in response to a copyright complaint, Twitter will make a good faith effort to contact the affected account holder with information concerning the removal or restriction of access, including a full copy of the complaint, along with instructions for filing a counter-notice.

If you've not yet received a copy of the copyright complaint regarding the content removed from your account, please respond to the support ticket we sent you.

In an effort to be as transparent as possible regarding the removal or restriction of access to user-posted content, we clearly mark withheld Tweets and media to indicate to viewers when content has been withheld (examples below). We also send a redacted copy of each copyright complaint and counter-notice that we process to Lumen, where they are posted to a public-facing website (with your personal information removed).

What if I want to contest the takedown?

If you believe that the materials reported in the copyright complaints were misidentified or removed in error, you may send us a counter-notification(s) through our Help Center. A counter-notice is a request for Twitter to reinstate the removed material, and it has legal consequences. Alternatively, you may be able to seek a retraction of the copyright complaint from the reporter.

How do I seek a retraction?

The DMCA complaint you received includes the contact information of the reporter. You may want to reach out and ask them to retract their notice. The reporter can send retractions to copyright@twitter.com, and should include: (1) identification of the material that was disabled, and (2) a statement that the reporter would like to retract their DMCA notice. This is the fastest and most efficient means of resolving an unresolved copyright complaint. A retraction is at the sole discretion of the original reporter.

When should I file a counter-notice?

A counter-notice is a request for Twitter to reinstate the removed material, and is the start of a legal process that has legal consequences. For example, submitting a counter notice indicates that you consent to the jurisdiction of a U.S. Federal court and that you consent to the disclosure of your personal information to the reporter and Lumen website.

With these considerations in mind, you may file a counter-notice if you believe that this material was misidentified, or you have a good faith belief that the material should not have been removed. If you're unsure whether or not you should file a counter-notice, you may want to consult with an attorney.

What happens after I submit a counter-notice?

Upon receipt of a valid counter-notice, we will promptly forward a copy to the person who filed the original notice. This means that the contact information that is submitted in your counter-notice will be shared to the person who filed the original notice.

If the copyright owner disagrees that the content was removed in error or misidentification, they may pursue legal action against you. If we do not receive notice within 10 business days that the original reporter is seeking



a court order to prevent further infringement of the material at issue, we may replace or cease disabling access to the material that was removed.

What happens if my account receives multiple copyright complaints?

If multiple copyright complaints are received Twitter may lock accounts or take other actions to warn repeat violators. These warnings may vary across Twitter's services. Under appropriate circumstances we may suspend user accounts under our repeat infringer policy. However, we may take retractions and counter-notices into account when applying our repeat infringer policy.

Enforcement and Appeals

Twitter is reflective of real conversations happening in the world and that sometimes includes perspectives that may be offensive, controversial, and/or bigoted to others. While we welcome everyone to express themselves on our service, we will not tolerate behavior that harasses, threatens, or uses fear to silence the voices of others.

We have the Twitter Rules in place to help ensure everyone feels safe expressing their beliefs and we strive to enforce them with uniform consistency.⁸⁵

Our policy development process

Creating a new policy or making a policy change requires in-depth research around trends in online behavior, developing clear external language that sets expectations around what's allowed, and creating enforcement guidance for reviewers that can be scaled across millions of Tweets.

While drafting policy language, we gather feedback from a variety of internal teams as well as our Trust & Safety Council. This is vital to ensure we are considering global perspectives around the changing nature of online speech, including how our rules are applied and interpreted in different cultural and social contexts. Finally, we train our global review teams, update the Twitter Rules, and start enforcing the new policy.

Our enforcement philosophy

We empower people to understand different sides of an issue and encourage dissenting opinions and viewpoints to be discussed openly. This approach allows many forms of speech to exist on our platform and, in particular, promotes counterspeech: speech that presents facts to correct misstatements or misperceptions, points out hypocrisy or contradictions, warns of offline or online consequences, denounces hateful or dangerous speech, or helps change minds and disarm.

Thus, context matters. When determining whether to take enforcement action, we may consider a number of factors, including (but not limited to) whether:

- the behavior is directed at an individual, group, or protected category of people;
- the report has been filed by the target of the abuse or a bystander;
- the user has a history of violating our policies;
- the severity of the violation;
- the content may be a topic of legitimate public interest.

Is the behavior directed at an individual or group of people?

To strike a balance between allowing different opinions to be expressed on the platform, and protecting our users, we enforce policies when someone reports abusive behavior that targets a specific person or group of people. This targeting can happen in a number of ways (for example, @mentions, tagging a photo, mentioning them by name, and more).

⁸⁵ <https://help.twitter.com/en/rules-and-policies/enforcement-philosophy>



Has the report been filed by the target of the potential abuse or a bystander?

Some Tweets may seem to be abusive when viewed in isolation, but may not be when viewed in the context of a larger conversation or historical relationship between people on the platform. For example, friendly banter between friends could appear offensive to bystanders, and certain remarks that are acceptable in one culture or country may not be acceptable in another. To help prevent our teams from making a mistake and removing consensual interactions, in certain scenarios we require a report from the actual target (or their authorized representative) prior to taking any enforcement action.

Does the user have a history of violating our policies?

We start from a position of assuming that people do not intend to violate our Rules. Unless a violation is so egregious that we must immediately suspend an account, we first try to educate people about our Rules and give them a chance to correct their behavior. We show the violator the offending Tweet(s), explain which Rule was broken, and require them to remove the content before they can Tweet again. If someone repeatedly violates our Rules then our enforcement actions become stronger. This includes requiring violators to remove the Tweet(s) and taking additional actions like verifying account ownership and/or temporarily limiting their ability to Tweet for a set period of time. If someone continues to violate Rules beyond that point then their account may be permanently suspended.

What is the severity of the violation?

Certain types of behavior may pose serious safety and security risks and/or result in physical, emotional, and financial hardship for the people involved. These egregious violations of the Twitter Rules — such as posting violent threats, non-consensual intimate media, or content that sexually exploits children — result in the immediate and permanent suspension of an account. Other violations could lead to a range of different steps, like requiring someone to remove the offending Tweet(s) and/or temporarily limiting their ability to post new Tweet(s).

Is the behavior newsworthy and in the legitimate public interest?

Twitter moves at the speed of public consciousness and people come to the service to stay informed about what matters. Exposure to different viewpoints can help people learn from one another, become more tolerant, and make decisions about the type of society we want to live in.

To help ensure people have an opportunity to see every side of an issue, there may be the rare occasion when we allow controversial content or behavior which may otherwise violate our Rules to remain on our service because we believe there is a legitimate public interest in its availability. Each situation is evaluated on a case by case basis and ultimately decided upon by a cross-functional team.

Some of the factors that help inform our decision-making about content are the impact it may have on the public, the source of the content, and the availability of alternative coverage of an event.

- **Public impact of the content:** A topic of legitimate public interest is different from a topic in which the public may be curious. We will consider what the impact is to citizens if they do not know about this content. If the Tweet does have the potential to impact the lives of large numbers of people, the running of a country, and/or it speaks to an important societal issue then we may allow the the content to remain on the service. Likewise, if the impact on the public is minimal we will most likely remove content in violation of our policies.
- **Source of the content:** Some people, groups, organizations and the content they post on Twitter may be considered a topic of legitimate public interest by virtue of their being in the public consciousness. This does not mean that their Tweets will always remain on the service. Rather, we will consider if there is a legitimate public interest for a particular Tweet to remain up so it can be openly discussed.



- **Availability of coverage:** Everyday people play a crucial role in providing firsthand accounts of what's happening in the world, counterpoints to establishment views, and, in some cases, exposing the abuse of power by someone in a position of authority. As a situation unfolds, removing access to certain information could inadvertently hide context and/or prevent people from seeing every side of the issue. Thus, before actioning a potentially violating Tweet, we will take into account the role it plays in showing the larger story and whether that content can be found elsewhere.

<<<>>>