Twitter progress report: Code of Practice on Disinformation

We are committed at Twitter to improving the collective health, openness and civility of the public conversation, and to hold ourselves publicly accountable to that mission. Twitter's conversational health will be built and measured by how we help encourage open democratic debate, healthy civic discourse, and critical thinking.

The fight against such things as spam, malicious automation activity and coordinated disinformation campaigns goes beyond any single election or event. We work every day to give everyone the power to create and share ideas and information instantly, without barriers. To preserve that experience, we are always working to ensure that we surface for our users the highest quality and most relevant content first. While Twitter's open and real-time environment is a powerful antidote to the abusive spreading of false information, we do not rely on user interaction alone.

In light of the upcoming European Parliament elections and in line with our commitments to the European Commission Code of Practice on Disinformation, we are taking active steps to stop malicious accounts and Tweets from spreading at scale, which involves trying to stay ahead of the constantly evolving tactics of bad actors. To try to achieve this goal, we have made a number of changes, including enhanced safety policies, better tools and resources for detecting and stopping malicious activity, tighter advertising standards, and increased transparency to promote better public understanding of all of these areas.

Scrutiny of ad placements

Safety in advertising on Twitter

In order to circumvent the possibility of ads and ad placement misleading users, we employ strict policies at Twitter around advertising, thereby ensuring quality and safety on the platform. Our advertising policy is organised around six key principles: keeping users safe, promoting honest content and targeting it responsibly, prohibiting the distribution of spam, harmful code, or other disruptive content, setting high editorial standards for the Twitter Ads content created, and being informed about the Twitter Ads processes that support these policies. Before an account can advertise on Twitter, it must meet certain criteria in order to be eligible, for example, a newly created account must meet quality requirements and will be held in review for a period before they can begin advertising with Twitter Ads. When advertisers on Twitter choose to promote their content with Twitter Ads, their content may become subject to a review process. The review process is designed to support the quality and safety of the Twitter Ads platform. This process helps Twitter check that advertisers are complying with our advertising policies, which are available at twitter.com/adspolicy.

Ad policies

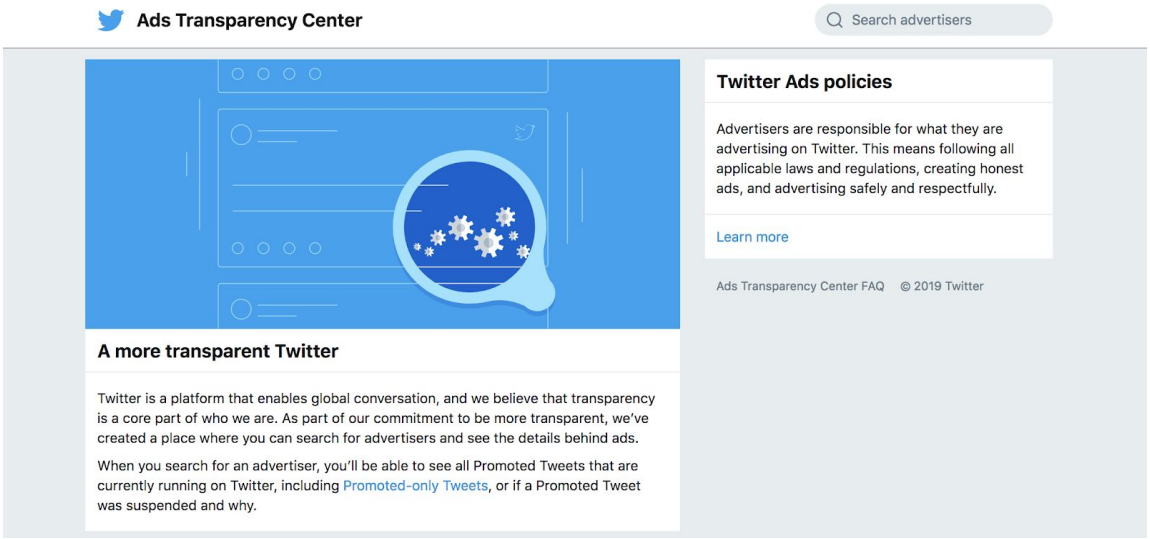
We provide policies for advertisers to ensure the quality of paid advertising products on our platform. Advertisers on Twitter are responsible for their Twitter Ads. This means following all applicable laws and regulations, creating honest ads, and advertising safely and respectfully. The clear policies we provide ensure that advertisements must meet specific criteria, a few examples include having a functioning URL in their bio, the ad should represent the brand and product being promoted, and **the ad cannot mislead users** to open content by including exaggerated or sensational language or misleading calls to action, among others. All our ads policies undergo regular revision and updates in order to be compliant with existing legislation and our product developments. We invest heavily in these areas to keep our product and processes updated.

Brand safety

An extension of our efforts to serve quality ads on Twitter, and to avoid exposure to any misleading or fraudulent information, is our work in brand safety. At Twitter, we take brand safety very seriously. We employ a combination of machine learning, people resources and placement controls to strive to ensure that brands' messages are served in a brand-safe environment. We continually invest in these areas to keep our product and processes updated.

Transparency of advertising

Transparency is a core value of the work we do at Twitter. Our bi-annual Twitter Transparency Report (launched in 2012) is one initiative to provide users more details on trends in legal requests, intellectual property-related requests, Twitter Rules enforcement, platform manipulation, and email privacy best practices on Twitter. In the summer of 2018, we had a key evolution of our transparency commitment, the launch of our new Ads Transparency Center. It enables anyone - whether they are a Twitter user or not - to view promoted Tweets that have been served on Twitter anywhere in the world. In light of the upcoming EU elections, and the growing concern over the role of coordinated disinformation campaigns in subverting the democratic process, our developments in this area include an expansion of the ATC across Europe, the details of which will be expanded upon in our first monthly report.



To ensure that users have a positive experience on Twitter, all ads are clearly identifiable and clearly marked with a “promoted” icon. You may see different kinds of ads on Twitter, such as:

Promoted Tweets

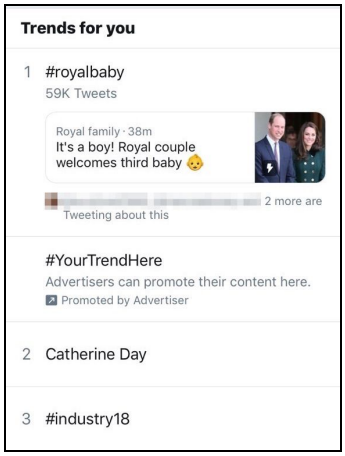
Promoted Tweets are ordinary that are clearly labeled as **Promoted** when an advertiser is paying for their placement on Twitter. In every other respect, Promoted Tweets act just like regular Tweets and can be retweeted, replied to, liked, and more.

Promoted accounts

Promoted Accounts are displayed in in multiple locations across Twitter, including your Home Timeline, Who to Follow section, and search results and is clearly labeled as **Promoted** to distinguish it from other recommended accounts.

Promoted Trends

Promoted Trends, which are not available for self-serve advertisers, show users time-, context-, and event-sensitive trends promoted by advertising partners. These paid Promoted Trends appear at the top of the Trending Topics list on Twitter and are clearly marked as “**Promoted**.”



Tackling malicious actors on Twitter (integrity of services)

Twitter is making significant progress in fighting malicious automation and spam, which can provide a vehicle for the spread of disinformation on platforms. Twitter fights spam and malicious automation strategically and at scale. Our focus is increasingly on proactively identifying problematic accounts and behaviour rather than waiting until we receive a report. Our primary goal on this front is to identify and challenge accounts engaging in spammy or manipulative behavior before users are exposed to misleading, inauthentic, or distracting content.

Below are some examples of how Twitter addresses spam and malicious automation through established policies and enforcement measures, tools for users, and transparency.

Policies and enforcement updates to address spam, malicious automation, and fake accounts

Throughout 2018, Twitter rolled out a number of initiatives to address malicious actors and spammy accounts that build on our policies around impersonation, inactive accounts, automation, and spam. Some key updates from 2018 include:

Tackling malicious automation

- We have made significant efforts to curb malicious automation and abuse originating via Twitter's APIs. **In 2018, we suspended more than 1,432,000 applications** in violation of our rules for serving low-quality, spammy Tweets. We are increasingly using automated detection methods to find misuses of our platform before they impact anyone's experience. More than half of the applications we suspended in the first quarter of 2018 were suspended within one week of registration, many within hours.

Auditing existing accounts for signs of automation

- We challenge millions of potentially spammy accounts every month, requesting additional details like email address and phone numbers to authenticate them. **From January to June, 2018, approximately 75% of accounts challenged ultimately did not pass those challenges and were suspended.**

Improving our sign-up process

- To make it harder to register spam accounts, new accounts are required to confirm either an email address or phone number when they sign up to Twitter. This is an important change to defend against people who try to take advantage of our openness. **The new protections we have developed as a result of this audit have already helped us prevent more than 50,000 spammy sign-ups per day.**

Updates to spam and fake account reporting

- To provide users more tools when they encounter potentially spammy or inauthentic accounts, we updated our spam reporting tools to allow users to report fake accounts.

New fake account and inauthentic activity policies

- As platform manipulation tactics continue to evolve, we updated and expanded our rules to better reflect how we identify fake accounts, and what types of inauthentic activity violate our policies. We now may remove fake accounts engaged in a variety of emergent, malicious behaviors. Some of the factors that we will take into account when determining whether an account is fake include: use of stock or stolen avatar photos, stolen or copied profile bios, or use of intentionally misleading profile information, including profile location.

Tweet source labelling (bots, automation)

- Tweets can be directly created by humans or, in some circumstances, automated by an application. In providing transparency around automation - and to help users understand the implication of automation and the intentions of accounts employing automation (which can be as simple scheduling a tweet with Tweetdeck, a social media dashboard application for management of Twitter accounts, commonly employed by accounts used in a professional context) - we have devised a number of measures, for example, Tweet source labels. These labels help users better understand how a Tweet was posted. This additional information provides context about the tweet and its author. If users do not recognise the source, they can use these labels to learn more. We have also published a partners page for a list of common third-party sources to provide more transparency and clarity for users.

Strengthening information security and actioning hacked material

- Our rules prohibit the distribution of hacked material. We have updated our policies and will now take action on accounts which claim responsibility for a hack, which includes threats and public incentives to hack specific people and accounts. Commentary about a hack or hacked materials, such as news articles discussing a hack, are generally not considered a violation of this policy.

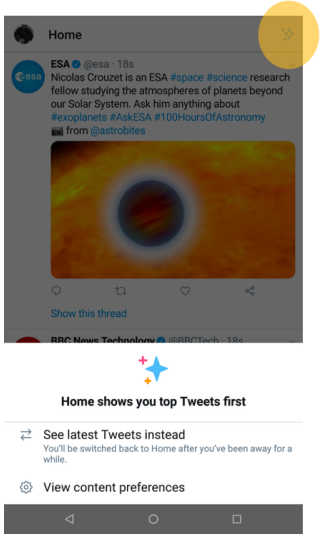
Transparency on actions to fight malicious automation and spam

In December 2018, we published the first Twitter Transparency Report with metrics pertaining to our actions to fight malicious automation and spam. The average number of reports we received through our reporting flow for spam related issues continues to drop — from an average of approximately 868,349 in January 2018 to approximately 504,259 in June 2018. This report indicates the effectiveness of our proprietary built technology in proactively identifying and challenging accounts at source and at scale.

Expanding digital literacy and empowering Consumers

Providing users more controls and a better understanding of advertisements on Twitter

We recognise that a healthy democracy needs well-informed citizens. The reality of the information age means that platforms like Twitter must do more to empower citizens with better media and information literacy skills, to help them critically analyse content they engage with and share online, and ensure it is credible.



In line with Commitment 9 of the Code of Practise on Disinformation, we have provided detailed insights in order to help users understand why they see certain advertisements on our platform, for examples, when a user follows, tweets, searches, views, or interact with tweets or Twitter accounts, we may use these actions to customise Twitter Ads. Furthermore we have provided tools to enable users to make meaningful privacy choices as regard ad personalisation and internet-based ads.

Furthermore, last year Twitter made it easier for users to switch their timeline from seeing the latest, chronologically posted tweets, and top tweets from their network with a simple click.

Building media literacy skills

We believe citizens should be empowered to develop the skills that help them critically analyse material for credibility and to ask the right questions of the information they are engaging with and sharing online.

For our part, we have already taken steps to help empower citizens become more media literate. This includes our dedicated educational resource, the Educator's Guide to Twitter, which raises awareness of media and information literacy among parents, educators, and academics. We have also supported the efforts of many non-profits across the world who work to promote media and information literacy.

As part of our ongoing efforts within Europe, last year we partnered with UNESCO to promote Global Media and Information Literacy Week 2018. Twitter has a long-standing, fruitful relationship with UNESCO, and this year we are particularly excited by their focus on developing a network of Media and Information Literate Cities, and will work closely with UNESCO to publish a co-branded educational resource entitled Teaching and Learning with Twitter, an easy-to-read manual for educators of all stripes who want to unlock the benefits of Twitter as a learning tool in the classroom and at home, while receiving guidance directly from UNESCO on media and information literacy best practices.

Other initiatives within Europe to support efforts aimed at improving critical thinking and digital media literacy include:

- Launching the new hashtag #ThinkBeforeSharing with an accompanying emoji as part of a global call to action which encourages people to pause and to be mindful of the content they choose to share online.
- Distributing Ads for Good grants to 10 separate NGOs in UNESCO's media and literacy network, to help raise awareness of their work.
- Partnership and ongoing work with the SELMA Education Task Force
- Membership of and participation in furthering the aims of the EU Media Literacy Expert Group
- Delivering a speech at MIL Week's feature conference at the Vytautas Magnus University in Kaunas, Lithuania, along with speakers from the Council of Europe, the European Commission, Harvard University, the London School of Economics, the Communications University of China, the Centre for Media and Information Literacy in Kenya, and the Brazilian Institute of Information in Science and Technology, amongst a truly global gathering of expert participants.

Empowering the research community

Enabling further research of Twitter

In line with our strong principles of transparency and with the goal of improving understanding of foreign influence and information campaigns, this year for the first time Twitter released all the accounts and related content associated with suspected information operations that we have found on our service since 2016, to enable independent academic research and investigation.

We made this data available with the goal of encouraging open research and investigation of these behaviors from researchers and academics around the world and include more than 10 million Tweets and more than 2 million images, GIFs, videos, and Periscope broadcasts, including the earliest on-Twitter activity from accounts connected with these campaigns, dating back to 2009.

Some examples of researchers using this data include Digital Forensic Research Lab, Atlantic Council and University of Washington.

Independent analysis of this activity by researchers is a key step toward promoting shared understanding of these threats. To support this effort, we have provided early access to a small group of researchers with specific expertise in these issues. Working with law enforcement and the authorities will always be our first priority, but we strongly believe that this level of transparency can enhance the health of the public conversation on the internet. This is our singular mission.

API access

Since 2006, Twitter's APIs have given developers the opportunity to tap into what's happening in the world. We are continually amazed by the innovative and helpful use cases developers discover for the Twitter platform; as well as the companies creating powerful tools to help businesses get the most out of Twitter.

The use of Twitter's API's are strictly governed and safeguarded by robust developer policies. Recognising the challenges facing Twitter and the public — from spam and malicious automation to surveillance and invasions of privacy — we announced the additional steps we planned to take to ensure that our developer platform works in service of the overall health of the conversation on Twitter.

We do not tolerate the use of our APIs to produce spam, manipulate conversations, or invade the privacy of people using Twitter. Indeed, between April and June 2018, we removed more than 143,000 apps which violated our policies, and we are continuing to invest in building out improved tools and processes to help us stop malicious apps faster and more efficiently.

Elections and civic engagement partnerships

Elections and civic engagement partnerships have emboldened our work in election integrity and we have funded academic research into exploring the use of manipulative techniques and disinformation with an initial focus on elections and civic engagement. We provide financial support and work closely with a wide range of organisations in this area, of particular benefit to European environment, for example:

Atlantic Council's DFRL Lab

- The Atlantic Council's Digital Forensic Research Lab (@DFRLab) is building a global hub of digital forensic analysts tracking events in governance, technology, security, and where each intersect as they occur.

EU DisinfoLab

- The EU DisinfoLab (@DisinfoEU) is a non-governmental organization based in Brussels with a mission to fight disinformation with innovative methodology and scientific support to the counter-disinformation community.

This year we launched request for proposals to help us develop a measurement framework. Partnering with outside experts who can provide thoughtful analysis, external perspective, and rigorous review is the best way to measure our work and stay accountable to those who use Twitter. Since March, we have reviewed more than 230 proposals from around the world that challenged us to think critically about how we can define and measure the health of public conversation on Twitter. The selected proposals will focus on:

Examining echo chambers and uncivil discourse

- Led by Dr. Rebekah Tromble, assistant professor of political science at Leiden University, along with Dr. Michael Meffert at Leiden, Dr. Patricia Rossini and Dr. Jennifer Stromer-Galley at Syracuse University, Dr. Nava Tintarev at Delft University of Technology, and Dr. Dirk Hovy at Bocconi University, this project will develop two sets of metrics: how communities form around political discussions on Twitter, and the challenges that may arise as those discussions develop.

Bridging gaps between communities on Twitter

- Professor Miles Hewstone and John Gallacher at the University of Oxford, in partnership with Dr. Marc Heerdink at the University of Amsterdam, will be studying how people use Twitter, and how exposure to a variety of perspectives and backgrounds can decrease prejudice and discrimination.