

January 2024
version 1

DISINFORMATION ON X: RESEARCH AND CONTENT MODERATION POLICIES



TABLE OF CONTENTS

INTRODUCTION	3
X'S ORGANISATION	4
INVESTIGATIONS ON X	7
HOW TO FLAG CONTENT ON TWITTER AND ITS ENFORCEMENT	8
RELEVANT CASES ON HOW TWITTER IS USED IN DISINFORMATION CAMPAIGNS	13

Authors: **Nicolas Hénin & Maria Giovanna Sessa**, EU DisinfoLab

Since its ownership change in 2022, the company operating Twitter has announced permanent updates to their policies and the staff in charge of implementing them. As a consequence, this document may already be outdated a few hours after publication, having been last updated as of 11 January 2024. However, the added value in compiling such a factsheet is providing a reference for past policies implemented by X (formerly Twitter) in order to understand future developments better.

INTRODUCTION

- Created in 2006, X (formerly Twitter) has quickly become hugely popular for its ease of spreading messages and following them. The preferred network of opinion makers such as the media and politicians in many countries, it has also become one of the preferred platforms for auditing public opinion (“[social listening](#)”), to the point where it is perhaps [exaggeratedly](#) considered to be a reliable reflection of [public debates](#).
- This position makes X a prime platform for [influence and disinformation operations](#). However, unlike its large and wealthy competitor, Meta, X has a fragile business model and has not made a profit, except [briefly in 2018 and 2019](#). This disproportion between X’s influence and its financial resources has led to an imbalance in the capacities it could deploy to counter harmful operations.
- In addition, Elon Musk’s takeover of X, announced in April 2022 and effective in October of the same year, has profoundly changed the company’s profile. Many employees were [dismissed or resigned](#), particularly in the [moderation and integrity](#) functions. For instance, it is unclear if the content curation team working on “[Topics, Trends descriptions, and Moments](#)” is still active. The new owner reinstated many [suspended accounts](#) and promoted [extremist conspiracy](#) accounts or others supporting [Russian narratives](#). [Disinformation indicators](#) spiked after this takeover.
- Finally, X informed the European Commission in May 2023 of its [withdrawal](#) from the EU voluntary Code of Practice against disinformation. While still part of it, the platform received criticism from the European Commission for being the only signatory to file an incomplete activity report and providing insufficient information on its counter-disinformation efforts.
- In July 2023, Musk announced that Twitter would be rebranded to X and that the bird logo would be retired. This brought many changes in how the platform addresses misinformation, in the direction of dismantling numerous counter-disinformation initiatives. However, references to some instruments and policies (e.g., [X Moments](#) or [pre-bunking](#)) are still available on the platform, causing confusion regarding what is still available.
- As a Very Large Online Platform (VLOP), X has to comply with the Digital Services Act (DSA)’s requirements. In the meantime, [rumour](#) has it that Elon Musk might consider removing the platform from Europe due to the legislation.
- In December 2023, the European Commission opened [formal proceedings](#) against X to assess whether the platform may have breached the DSA regarding risk management, content moderation, ad transparency and access to data for researchers.

X'S ORGANISATION

For an overview of how to use the platform, the Help Center offers a [guide](#).

- X works primarily on posting, sharing, and reading short messages, formerly known as 'tweets', but not simply called posts. Users sign up with an email address or phone number. They choose a username (also known as a handle) that begins with a "@". The username serves as a unique identifier on the platform and is attached to a unique user ID composed of digits. Users can change their username but not their user ID.

Once started, possible actions on X and its main features are:

- **Following and followers:** users can follow other users whose posts will appear on their algorithmically-defined feed.
- **Posting:** a post (formerly 'tweet') is a message or post on X. It can contain up to 280 characters, including text, hashtags, links, and media such as photos, videos, or GIFs. Users can write a post by clicking on the "What is happening" text box on the platform homepage or the plus symbol on the mobile app. The posts will become visible on the homepage of one's followers. If users opt for a public account, anyone can see their posts. Moreover, users can emphasise their content by clicking on the three dots next to a post and selecting either "pin to your profile" or "highlight on your profile" to see a selection of these on top of their posts section.

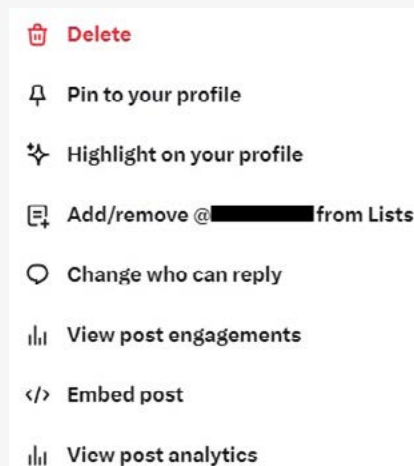


Figure 1. List of options that appear by clicking on the three dots next to a post

- **Reposting and liking:** X allows users to engage with posts in various ways. Users can share a post with their followers, allowing them to see the post and the user who posted it. They can repost (previously referred to as 'retweet') the message as it is or "quote" it, which allows them to add a comment. The platform has a feature inviting users to read an article before sharing it. Liking a post generally indicates either endorsement or simple acknowledgement. Reposts and likes help to amplify content and increase its visibility on the platform.



Figure 2. Alert suggesting to read an article before reposting it

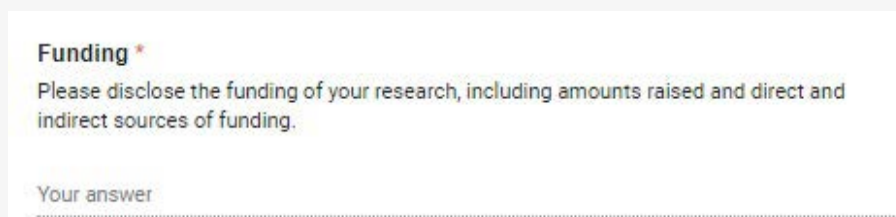
- **Replies and mentions:** users can reply to a post, allowing them to respond to the original post or participate in a conversation. When somebody replies to a post, the user who posted it receives a notification. Additionally, it is possible to mention another user in a post by including their username (e.g., "@username"). This notifies the mentioned user and helps to bring their attention to the post.
- **Hashtags and trending topics:** hashtags are keywords or phrases preceded by the “#” symbol. They help categorise posts and make them discoverable by others interested in similar topics. When clicking on a hashtag, other posts related to that topic appear. X also displays trending topics, which are popular hashtags or keywords being widely discussed at a given time in a certain geographic area.
- **Direct messages:** X allows users to send private messages called direct messages (DMs) to each other, allowing users to have private conversations with users that follow each other outside of the public sphere of posts and interactions.
- **Lists and bookmarks:** X provides features like lists and bookmarks to help users organise content. Lists allow users to create custom timelines with posts from specific users, making it easier to follow specific topics or groups of people. Bookmarks allow users to save posts for later reference.
- **Live:** Users can record [live videos](#), which will be available afterwards on their profiles. During live videos, it is possible to interact with the broadcaster and other viewers by commenting and clicking the heart icon on the bottom right corner of the video.
- **Spaces:** This is a [live audio conversation feature](#) that provides an interactive way for people to connect and engage in real-time discussions on X. Users are allowed to create and join live audio conversations where they can listen, participate, and engage with others on various topics.
- **Privacy settings and notifications:** The platform offers privacy settings that allow users to control who can see their posts and interact with them. Notification settings can be customised to manage what notifications one wishes to receive, such as mentions, likes, or reposts.
- **Explore and discover:** The ‘Explore’ tab helps users discover new content, including trending topics (from football to politics) that might be interesting based on their behaviour on the platform (e.g., engagement and past interactions). It provides a broader view of what is happening on social media outside of one’s bubble of followed accounts.
- **Communities:** These are “a dedicated place to connect, share, and get closer to the discussions they care about most”. Admins and moderators are users who manage the [Community](#), and those who accept invites to join it are members.

- **Subscriptions:** Recently, X launched an “opt-in, paid [subscription service](#) that offers additional features”, and it has three basic tiers: Basic, Premium, and Premium+. The latter two also include a [blue checkmark](#) and the opportunity to post long-form content. Moreover, businesses, governments and nonprofits can subscribe to Verified Organisations and receive a gold or grey checkmark, affiliation badges, premium support, impersonation defence and more features.
- **Analytics:** This business feature provides insights and metrics into how one’s account is performing through an [analytics dashboard](#)”.
- **Ads:** The [X Ads Help Center](#) explains that whether they are Promoted Ads, Follower Ads, or Trend Takeover, all ads on X are clearly labelled as such. Ads can be customised based on user’s behaviour and information shared with the app, and it is possible to interact with promoted content in the same way as one can do with organic content.

INVESTIGATIONS ON X

X is a valuable resource for conducting investigations or studying various topics despite [increasing difficulties](#) in researching the platform. Still, [directories](#) of popular [tools](#) usable for OSINT are available online. To investigate the platform, researchers can rely on the following:

- **Data collection:** Researchers could traditionally collect data from Twitter using the platform's Application Programming Interfaces (APIs) to access public posts and user data. The Help Center has a [dedicated page](#) on API.
- In February 2023, Twitter limited access to its API, introducing [new pricing](#): a free, write-only plan, a “basic” plan that allows one to get 10,000 posts per month for 100 USD, and the top-tier “Enterprise” plan that is tailored, in terms of price and capacities, to the customer's need. This new pricing [terminated](#) a large set of free-to-use, very popular web-based OSINT tools.
- **Advanced search:** X's [advanced search feature](#) enables researchers to refine their searches by specific [criteria](#) such as keywords, hashtags, dates, locations, and user mentions.
- **Data analysis:** Once researchers have collected the necessary data, they can analyse it using various methods. This could involve manual coding and categorisation of posts, sentiment analysis to understand the emotional tone, network analysis to examine relationships between users, or other quantitative and qualitative techniques. It is also possible to use data analysis tools and programming languages like Python or R to process and analyse large datasets.
- **Social network analysis:** Researchers can explore the network structure, identify influential users, and analyse information flow within the network. Network analysis tools like [Gephi](#) or [NodeXL](#) can be used to visualise and analyse these social connections.
- **Topic modelling:** Researchers may want to identify and analyse specific topics or themes in a large volume of posts. Topic modelling techniques, such as [Latent Dirichlet Allocation](#) (LDA), can be applied to uncover inherent topics and understand the distribution of different themes within the dataset.
- **Access to public data:** In compliance with DSA Article 40(12) on access to public data, requests must be made through the “[X DSA Researcher Application](#)”. Besides relevant questions (e.g., the description of the research or the organisation's data security and confidentiality capabilities), other questions regarding funding, “including amounts raised and direct and indirect sources of funding”, are overly detailed and sensitive information that might prevent many stakeholders from obtaining data access.



Funding *
Please disclose the funding of your research, including amounts raised and direct and indirect sources of funding.

Your answer

Figure 3. X DSA Researcher Application's question about funding

HOW TO FLAG CONTENT ON TWITTER AND ITS ENFORCEMENT

REPORTING CONTENT

Help Centre: [X Rules](#) aimed at ensuring “all people can participate in the conversation freely and safely” include safety, privacy, and authenticity. Users can report specific violations through the [Help Centre](#):

- Unauthorised trademark use,
- Unauthorised use of copyrighted material,
- Sale or promotion of counterfeit goods,
- Privacy policy towards children (under 13 years old),
- Child sexual exploitation,
- Pornography,
- Impersonation of an individual or brand,
- Private information posted on X,
- Abusive behaviour and violent threats,
- Spam and system abuse,
- Violation of X Ads policy.

The Help Centre still lists the content that can be reported in X Moment (i.e., violation of posting private information, abuse, hateful conduct, violent threats, and self-harm) despite the feature being discontinued.

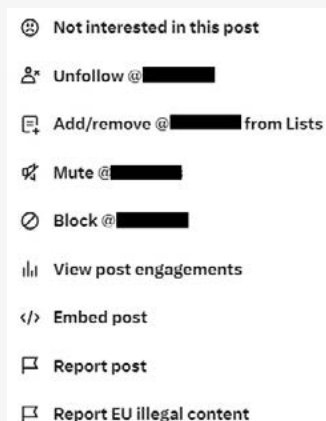


Figure 4. What appears when clicking on the three dots on the top right of a post

Reporting content in the app: The easiest way to report content is to click on the three dots on the top right corner of a post or ad and select “Report content”, which can be done for the following issues:

- Hate,
- Abuse & harassment,

- Violent speech,
- Child safety,
- Privacy,
- Spam,
- Suicide or self-harm,
- Sensitive or disturbing media,
- Deceptive identities,
- Violent & hateful entities.

Some of these categories can include forms of disinformation, e.g., “financial scams” and “fake engagement” in Spam, or “impersonation” in Deceptive identities.

Reporting EU illegal content in the app: Recently, X introduced another reporting venue in compliance with the DSA. When clicking on this opinion, the user is redirected to the Help Centre’s DSA dedicated page. The options are:

- Report illegal content in the EU,
- Appeal an illegal content decision,
- Information about out-of-court dispute settlement.
- Moreover, users need to fill out a template detailing information about the post or ad they wish to report, choosing among a series of legal reasons (see Figure below).

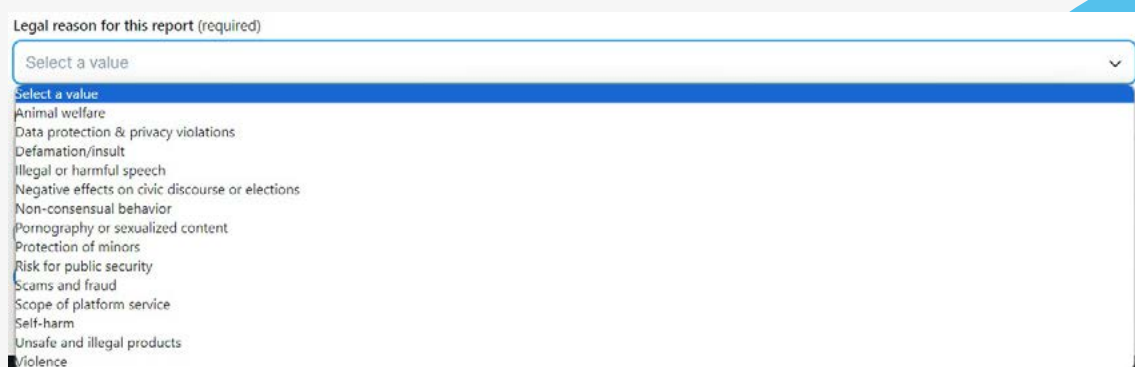


Figure 5. Options for which illegal content can be reported under the DSA

“HOW WE ADDRESS MISINFORMATION ON X”

- A dedicated page titled “[How we address misinformation on X](#)”, stated that X defines “misleading content (‘misinformation’) as claims that have been confirmed to be false by external, subject-matter experts or include information that is shared in a deceptive or confusing manner”.

- To identify this content, X uses “a combination of human review and technology, and through partnerships with global third-party experts”. However, at present, X does not have any ongoing partnerships with European fact-checking organisations.
- The rationale to act is “the risk of public harm”, and actions include reduced visibility, labelling, or removal of misleading content based on the level of potential harm. These actions are also detailed as “[enforcement options](#)” (e.g., regarding ads).
- According to the page, other activities include “X Moments”, where users can learn from trusted sources, and, in some markets, the Misleading Info Reporting Flow. However, these functions seem to be no longer available as X has recently removed the option that allowed users to report [misleading information](#), including [misleading information about politics](#), directly.

Although the abovementioned criteria are still available online at the time of our writing, it is unclear if they are still enforced. The only regularly communicated counter-misinformation tool provided by X is the following:

- [Community Notes](#), formerly known as Birdwatch, is X’s crowdsourced fact-checking system. The feature allows users to add context to posts and takes an open-source approach toward debunking misinformation. In addition, [Helpful Community Notes](#) collects Community Notes that contributors rated as helpful.



Figure 6. A screenshot of the Community Notes account.

POLICIES AGAINST MISINFORMATION

- Navigating X's website is rather complex, as its policies have undergone various edits. Besides the general X [Rules and policies](#), other policies tackle "Platform integrity and authenticity", "Safety and cybercrime", and "Intellectual property". Furthermore, the platform provides platform use and [law enforcement](#) guidelines.
- It is worth noting that some aspects of the [X Ads policies](#) are relevant for disinformation, such as the prohibition of "[Deceptive & fraudulent content](#)". It is worth noting that this and other policies are being continuously updated as [inappropriate content](#) included, until recently, categories such as "Misrepresentative content", "Misleading synthetic or manipulated content", and "Content engaged in coordinated harmful activity".

X implements three specific policies, namely:

Crisis misinformation policy

The [policy](#) aims to act against the use of the platform "to share false or misleading information that could bring harm to crisis-affected populations". Crises consist of "situations of armed conflict, public health emergencies, and large-scale natural disasters". The public safety and serious harm rationale is maintained as crises are considered to entail "a widespread threat to life, physical safety, health, or basic subsistence".

While this policy was initially outlined in the framework of the pandemic, X is no longer enforcing its [COVID-19 misleading information policy](#).

It is unclear if the policy has been implemented since the ownership change.

Synthetic and manipulated media policy

This [policy](#) addresses "synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm".

EU DisinfoLab compiled a more comprehensive factsheet on platforms' policies on [AI-manipulated and generated misinformation](#).

It is unclear if this policy is implemented, as we have no knowledge of reports or communications that have been made on such cases.

Civic integrity policy

X's first election integrity policy was published in [April 2019](#). The [current version](#), renamed "civic integrity", aims to avoid using the platform "for the purpose of manipulating or interfering in elections or other civic processes" (e.g., political elections, censuses, and major referenda and ballot initiatives).

While some country-specific policies around electoral misinformation as still available, for instance, to safeguard [French elections](#), it is [not possible anymore](#) to report "false information about voting or registering to vote", leaving some doubts and uncertainty regarding what can be done currently and in the future.

More considerations on electoral disinformation policies adopted by the platform are available in a [fact-sheet](#) compiled by EU DisinfoLab.

It is unclear if this policy is enforced and by which team.

COMPLIANCE WITH THE DIGITAL SERVICES ACT

As a VLOP, X is required to comply with DSA requirements that include [transparency reporting](#), for which it has set a specific section.

The description of content moderation practices details that human moderation also includes scaled human investigation, and automated moderation also combines machine learning and heuristic models.

At the time of our writing, “Enforcement Activity Summary Data” is reported for the period between August 28 and October 20, 2023. It should be noted that although the DSA focuses on illegal content exclusively, some categories delve into disinformation, such as “misleading and deceptive identities”, “synthetic and manipulated media”, or “platform manipulation”.

RELEVANT CASES ON HOW TWITTER IS USED IN DISINFORMATION CAMPAIGNS

Far from being exclusive, this section lists several recent studies exploring disinformation campaigns on X.

- In October 2023, the EU opened an [investigation](#) into Elon Musk's X over the possible spread of terrorist and violent content and hate speech after Hamas' attack on Israel.
- For starters, a September 2023 [comparative analysis](#) of the prevalence and sources of disinformation across major social media platforms in Poland, Slovakia, and Spain found that X had the highest amount of disinformation content, engagement, and actors among the data analysed.
- According to the [Climate Action Against Disinformation](#), X ranks worst compared to Meta, Pinterest, YouTube, and TikTok in preventing climate change misinformation. It scored 1 out of 21 due to its lack of relevant policy measures.
- The Center for Countering Digital Hate (CCDH) published eight papers, including one evidencing that X did not take any action against 99 of the 100 accounts identified for posting hateful content. In June 2023, X sent CCDH a [letter](#), accusing it of negatively affecting the social media through its research on hate speech, labelled as "false, misleading or both".
- In April 2023, [NewsGuard](#) found a network of seemingly inauthentic Chinese-language X accounts spreading disinformation to discredit two well-known Chinese activists and dissidents, including one identified in a previous Axios investigation.
- In the context of the 2023 Nigerian elections, [research](#) sheds light on the role of X Spaces in spreading disinformation, as the live audio conversation feature is "riddled with fake and unsubstantiated claims and inaccurate and exaggerated figures" that remain largely unregulated.
- An ongoing Russian-based influence operation network labelled [Doppelganger](#), operating in Europe since at least May 2022, used multiple clones of authentic media to spread pro-Kremlin disinformation and propaganda about the Ukraine war. Various social media accounts, including on X, have actively amplified this operation.
- The [@TwitterSafety](#) blog page reported only four takedowns by the platform: in August 2019, December 2019, June 2020 and February 2021. Only a few others were documented.