‘A THOUSAND CUTS’

TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE AGAINST POLAND’S LGBTI COMMUNITY ON X

AMNESTY INTERNATIONAL

Amnesty International is a movement of 10 million people which mobilizes the humanity in everyone and campaigns for change so we can all enjoy our human rights. Our vision is of a world where those in power keep their promises, respect international law and are held to account. We are independent of any government, political ideology, economic interest or religion and are funded mainly by our membership and individual donations. We believe that acting in solidarity and compassion with people everywhere can change our societies for the better.

© Amnesty International 2025
Except where otherwise noted, content in this document is licensed under a Creative Commons (attribution, non-commercial, no derivatives, international 4.0) licence.
https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode
For more information please visit the permissions page on our website: www.amnesty.org
Where material is attributed to a copyright owner other than Amnesty International this material is not subject to the Creative Commons licence.
First published in 2025
by Amnesty International Ltd
Peter Benenson House, 1 Easton Street
London WC1X 0DW, UK

Index: EUR 37/0098/2025
Original language: English

amnesty.org

Cover illustration: *Illustration of a distressed person looking at their phone, which has an LGBTI sticker on it, surrounded by eyes glaring out of the darkness and the X logo behind them. Highlights the adverse impacts of TfGBV on X on LGBTI people. © Aleksandra Herzyk*

AMNESTY INTERNATIONAL

1. EXECUTIVE SUMMARY

CONTENT WARNING

This report covers sensitive issues including technology-facilitated gender based-violence (TfGBV) and contains examples of content which include graphic calls for violence and discrimination, which may be distressing for some readers.

"Twitter is basically a never-ending stream of deadnaming, misgendering, insults and death wishes", Maja Heban told Amnesty International, describing her experience as an openly trans woman on X (formerly known as Twitter). This description of a platform awash with content targeting the LGBTI community was repeated by all the LGBTI activists interviewed by Amnesty International for this report.

For decades, Poland's LGBTI community has struggled with systemic discrimination. This discrimination was made more acute between 2015 and 2023 under the government led by the Law and Justice party (Prawo i Sprawiedliwość, PiS), during which Polish authorities took actions that shrank space for civil society, by undermining the rule of law and attacking the rights of women and LGBTI people and creating an increasingly inhospitable environment for LGBTI people and their allies.

Hostile and stigmatizing rhetoric against LGBTI people, including by high-level politicians, became commonplace. In 2022, Amnesty International found compelling evidence of how this rhetoric translated into violence against the community, with a marked increase in attacks on LGBTI people at peaceful gatherings, such as Equality Marches and protests.

A prominent example of this is the 2019 Bialystok Equality March, where attendees were attacked with bottles, paving stones and firecrackers, and subjected to hateful slurs from counter-protestors. A few months later at the Lublin Equality March, police arrested dozens of counter-protesters who came to attack the peaceful march. It was later revealed that two of the counter-protesters had brought home-made explosives to the march.

In 2020, hostility towards LGBTI people in Poland was so high that around one-third of regions in the country had passed symbolic resolutions against "LGBT ideology".

Against this backdrop, X became awash with content advocating hatred that constituted incitement to violence, hostility or discrimination against LGBTI people, amounting to technology-facilitated gender-based violence (TfGBV) and entailing a range of human rights harms. This content was particularly prominent on the X accounts of politicians, many of whom posted content that advocated hatred and dehumanized LGBTI people, suggesting that their identity was a political "ideology" and that they presented a threat to children's safety. The proliferation of these posts on the platform enabled an environment in which advocating hatred towards LGBTI people became increasingly normalized and socially acceptable.

The presence of content constituting TfGBV on X was exacerbated by the company's poor content-moderation practices, which deteriorated further because of drastic staff cuts after Elon Musk's takeover of the platform in October 2022. A week after Elon Musk's takeover, individuals promoting anti-rights narratives appeared to begin testing X's limits on anti-LGBTI speech. Former Ultimate Fighting Championship fighter Jake Shields (who has 34,000 followers on X), posted a photo of a drag queen with the caption: "This is a

groomer". He went on to say, "I was suspended for this exact tweet a month ago so we will see if Twitter is now free."

Shortly after taking ownership of X, Elon Musk disbanded the Trust and Safety Council, an advisory group comprising 100 civil society, human rights and other organizations that sought to address child exploitation, suicide, self-harm and hate speech on the platform. It is estimated that Elon Musk also fired 80% of the engineers dedicated to trust and safety. In late 2022, it was reported that he planned to rely heavily on automation to moderate content, a method known to be error-prone, removing certain manual reviews. In 2023, X introduced Community Notes, essentially outsourcing some content moderation functions to randomly selected platform users who sign up as contributors and meet certain eligibility criteria.

X's policies on harmful content, including content which may constitute TfGBV, have also shifted during Elon Musk's tenure. For example, in April 2023, X removed a policy against the "targeted misgendering and deadnaming of transgender individuals". This policy was reinstated in 2024.

Elon Musk had previously said that he would relax the rules about what content was allowed on the platform, suggesting that X should permit all posts that stop short of violating the domestic law of the countries in which it operates.

It seems that he has made good on his word. Before 30 October 2023, X's Community Guidelines stated, "we have a zero tolerance policy towards violent speech in order to ensure the safety of our users and prevent the normalization of violent actions." (Emphasis added.) After the update, the policy now reads, "we may remove or reduce the visibility of violent speech in order to ensure the safety of our users and prevent the normalization of violent actions." (Emphasis added.)

LGBTI community members in Poland told Amnesty International that, by being visible on the platform, they faced a tide of hatred based on their real and/or perceived gender, sexual orientation, gender identity and/or expression. Many interviewees explained that the online rhetoric had an adverse effect on their well-being. For example, Jolanta Prochowicz, a lesbian woman based in the city of Lublin, told Amnesty International: "We should recognize social media as part of our social life, if we say something on the internet, it hurts like it's real... It's harmful, it's painful and it can be very powerful. Social media does not *affect* our normal life, it *is* our normal life, and it has influence on us."

Aleksandra Herzyk, an asexual woman living in the city of Krakow, told Amnesty International that she was targeted on X after speaking about her asexuality on the platform. Aleksandra also experienced being targeted with content constituting TfGBV on X after writing about her decision to have breast reduction surgery, which led some platform users to incorrectly identify her as a trans woman. Aleksandra told Amnesty International: "You know, the things that you read about yourself – they're not true but somehow, they stay in your head. It's like death by a thousand cuts".

Aleksandra told Amnesty International that, after experiencing hate on X, she no longer uses the platform, logging out permanently in early 2024. In 2018, in a report named "Toxic Twitter", Amnesty International found that X (then known as Twitter) was failing to respect women's rights online by not appropriately mitigating online abuse, with women of colour, women from ethnic and religious minorities, lesbian, bisexual and transgender women, non-binary individuals, and women with disabilities being exposed to the most abuse on the platform. In 2020, Amnesty International found that, although X had made some progress on addressing TfGBV since 2018, the company continued to fall short of its human rights responsibilities.

It now seems that little has improved since 2020 – at least in the context of Poland. In 2024 a Polish NGO called the Never Again Association published a report documenting 343 examples of "hate" which it reported to X between August 2023 and August 2024. Never Again Association is registered as a Trusted Flagger by an online monitoring project financially supported by the EU's Citizens, Equality, Rights and Values programme. In most of the documented cases, X either refused to remove the posts or ignored the reports. The posts contained content which could be considered as inciting violence and discrimination against marginalized communities, including the LGBTI community. Several of the posts reported by Never Again Association – including posts portraying LGBTI people as deviants, using slurs and calling for discrimination against, or even the elimination of, the LGBTI community – remain visible on the platform. This report outlines how X – through its poor content moderation practices and lack of human rights due diligence – has failed to prevent and adequately mitigate TfGBV targeting Poland's LGBTI community on its platform and has therefore contributed to human rights abuses perpetrated against the community. It details how under-resourcing of content moderation was an issue at the company even prior to Elon Musk's takeover in 2022 and how the company has failed to adequately engage with LGBTI civil society organizations in Poland to mitigate risks to the community on the platform. These failures, combined with the company's unjustifiable removal of safeguards to protect platform users from harmful speech – in alignment with Elon Musk's self-declared policy of "free speech absolutism" – has led to X becoming awash with

content constituting TfGBV, including advocacy of hatred that constitutes incitement to violence, hostility or discrimination against LGBTI people.

As part of this research, Amnesty International conducted quantitative research on X in partnership with the National Conference on Citizenship's Algorithmic Transparency Institute (ATI), using 32 research accounts which collected 163,048 tweets between 1 March and 31 March 2025.[1] This quantitative research found that anti-LGBTI content is highly prevalent on the platform. Analysis of the sample 1,387 tweets suggests that homophobic and transphobic content is highly prevalent on X, particularly for accounts that follow politicians who do not support the rights of LGBTI people. Amnesty International found that almost 4% of tweets collected by research accounts from the accounts of politicians who do not support the rights of LGBTI people were homophobic or transphobic and, more than 25% of all LGBTI-related content seen by these accounts was homophobic or transphobic. Additionally, Amnesty International found that a high amount of the content related to LGBTI issues contained homophobic and transphobic content (whether in posts or in replies to posts) and that the research accounts following politicians supportive of the rights of LGBTI people were more exposed to these replies.

From 2015 to 2023, Poland was ruled by the Law and Justice (PiS) party, which was overtly anti-LGBTI. Despite the change in government in Poland after the 2023 election, the years of targeting the LGBTI community have resulted in what activists have described as "top-down polarization", reflected in the pervasive nature of anti-LGBTI content on X. This prevalence is made more concerning by the fact that X's business model relies on recommending content that users will find engaging, regardless of its potential impact.

In this report, Amnesty International has for the first time undertaken a comprehensive human rights-based analysis of X's business model and found that it operates a surveillance-based business model, as we have found for other technology companies including Meta, Google and TikTok. Similar to other companies operating a surveillance-based business model, the collection of user data is central to X as a platform, not only because it allows the platform to better predict what content will interest its users, but also because the value of the data determines the value of the company to potential advertisers. This appeal to potential advertisers is crucial because of X's reliance on targeted advertising.

Since 2013, almost all of X's revenue came from targeted advertising on the site. In order to maintain and optimise the collection of user data, X's algorithms prioritize maximizing 'user engagement' above all else, by surfacing content users are most likely to interact with (in the case of X, inferred through comments, retweets and liking content). X also offers premium subscriptions, allowing users to pay for additional features such as longer posts, and enhanced algorithmic amplification, which includes "reply prioritization", meaning that replies by premium users are more visible underneath posts.

As Amnesty International has previously documented, surveillance-based business models risk fuelling the spread of harmful content in the quest for ever-more engagement and user data. This business model, combined with poor content moderation policies and practices, puts Poland's LGBTI community at great risk of the compounding harms of being targeted with large amounts of content constituting TfGBV.

To look at a typical example of content targeting the LGBTI community and circulating on the platform, in July 2023 the Polish political party Konfederacja posted a clip of one of its then MPs, Grzegorz Braun, speaking about the LGBTI community in parliament. In the clip, he says: "We don't want deviants, promoters of deviance and ostentatious professional sodomites teaching our children tolerance." As of May 2025, the post remains visible on X. It has been viewed more than 99,000 times.

LGBTI people told Amnesty International that they regularly see posts on the platform dehumanizing them or even calling for their extermination. One interviewee described posts stating that: "LGBTQ people will be in gas chambers, or they talk like we are trash, and they think that we have to be cleansed". Another said that they have seen posts claiming that "[LGBTI] people are not normal, they are against Polish families, they are destroying Polish families, they are not people, they are [an] ideology."

X's wholly inadequate investment in content moderation in general, and specifically in Poland, is a significant factor in the company's failure to remove content constituting TfGBV targeting the LGBTI community. According to its own transparency reports, X has just two Polish-speaking content moderators – one of whom has Polish as their second language – responsible for covering a population of 37.45 million people and 5.33 million X users. This is indicative of the company's lack of investment in content moderation resources, also demonstrated by X's introduction of Community Notes, which effectively outsources content moderation to

[1] Research accounts are online fictitious identities. They can be used for multiple purposes. In this research, Amnesty International and ATI used them to better understand the prevalence and amplification of anti-LGBTI content on X in Poland.

platform users. The combination of poor resourcing, policy and practice has contributed to X becoming a platform awash with hateful content targeting the LGBTI community.

All companies have a responsibility to respect human rights wherever in the world they operate and throughout their operations. To meet this responsibility, companies must engage in ongoing and proactive human rights due diligence processes to identify, prevent, mitigate and account for how they address their impacts on human rights. For technology companies such as X, due diligence should also include addressing situations in which their business model, operations, design decisions and content moderation practices create or exacerbate human rights risks.

Under the EU's Digital Services Act (DSA) regulation, so-called Very Large Online Platforms (VLOPs) such as X, are obligated to assess and mitigate systemic risks and must produce yearly risk assessments. In X's most recent publicly available risk assessment from 2024, the platform acknowledges that individuals and groups might be targeted with hateful content or abuse on the platform, and that this could create a sense of fear and intimidation and lead to self-censorship. X listed several mitigation measures for this, including downranking content (reducing the visibility of certain content), transparency about rules and processes, and quality controls and process reviews of policies. However, the risk assessment makes no specific mention of risks to the LGBTI community. The DSA-mandated independent audit of X's risk assessment covering the year to 23 August 2024 found that the platform's risk assessment process was not sufficiently rigorous and that the current mitigation measures it outlined were ineffective in reducing systemic risks and highlighted a lack of mitigation measures relating to algorithmic systems, among other failings.

This report finds that X has failed to conduct appropriate human rights due diligence in respect of its operations in Poland, even after being mandated to conduct risk assessments by the DSA. It therefore has failed to take adequate measures to prevent or mitigate any risks or harm that its products, services and operations could create. This analysis makes clear that X has facilitated the spread of content constituting TfGBV on its platform and has contributed to human rights abuses against Poland's LGBTI community.

On 22 August 2024, Amnesty International wrote to X, posing questions regarding the company's actions in relation to its business activities in Poland between 2019 and 2024. X did not respond.

As detailed throughout this report, X's failure to uphold its human rights responsibilities, as outlined in the UN Guiding Principles on Business and Human Rights (UN Guiding Principles), as well as its legal obligations contained in the DSA, has contributed to significant harm for Poland's LGBTI community. X's grossly inadequate mitigation measures and cavalier attitude to hateful content, combined with a business model that exacerbates human rights risks, heightens the possibility of repetition of harm in the future. Urgent, wide-ranging reforms are needed to ensure that X does not continue to contribute to these human rights harms – including, crucially, adequate resourcing of content moderation and a change to its surveillance-based business model.

X's repeated failures in Poland demonstrate that the company is still failing to address its systemic risks to human rights. The DSA provides an important route for accountability and remedy and must be robustly and meaningfully enforced.

Unfortunately, the Polish government has not yet fully implemented the legislation nationally, does not have a fully designated or empowered national Digital Services Coordinator (DSC), as mandated by the DSA, and has not laid down the rules for DSA penalties. It is vital that the Polish government addresses the lack of a DSC as a matter of urgency and ensures that the role is effectively resourced in terms of expertise, capacity and funding. Without a DSC, users of X in Poland are unable to fully exercise their rights under the DSA. In May 2025 the European Commission referred Poland – alongside Czechia, Spain, Cyprus and Portugal – to the Court of Justice of the European Union due to their respective failures to effectively implement the DSA domestically.

Meanwhile, the European Commission can launch an investigation into X immediately, and further scrutinize the platform's mitigation of systemic risks stemming from both its business model and its content moderation practices. This is of particular importance due to the continuing negative effects on Poland's LGBTI community of TfGBV on X – including adverse effects on individuals' rights to freedom of expression and non-discrimination.

The EU has the tools to meet its obligation to protect human rights – including the right to live free from gender-based violence (GBV). It must not hesitate to use them.

7. THE BUSINESS OF HATE: HOW X'S BUSINESS MODEL FUELS HUMAN RIGHTS RISKS AND HARMS

> **"It's not a very friendly place and it's very frustrating when we are reading things there. It's very hard to maintain your psychological health, [using] Twitter."[323]**

7.1 A SURVEILLANCE-BASED BUSINESS MODEL

Amnesty International has previously found that the technology companies Meta (Facebook's parent company) and Google operate a surveillance-based business model which relies on constant data collection from their users in order to better target them with advertising on the platform. This is inherently incompatible with the right to privacy and poses a threat to a range of human rights including freedom of opinion and expression, freedom of thought, and the right to equality and non-discrimination.[324]

The US's Federal Trade Commission (FTC) has similarly found that major social media and video streaming services – including X – are engaged in vast surveillance of consumers to monetize their personal information while failing to adequately protect users online.[325]

[323] Amnesty International interview with Mateusz Kaczmareck, 28 July 2024.
[324] Amnesty International, *Surveillance Giants: How the Business Model of Google and Facebook Threatens Human Rights*, (Index: POL 30/1404/2019), 21 November 2019, https://www.amnesty.org/en/documents/pol30/1404/2019/en/
[325] FTC, "FTC staff report finds large social media and video streaming companies have engaged in vast surveillance of users with lax privacy controls and inadequate safeguards for kids and teens", 19 September 2024, https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-staff-report-finds-large-social-media-video-streaming-companies-have-engaged-vast-surveillance

This section will outline the features of X's surveillance-based business model and how it presents a systemic risk to human rights.

7.1.1 RELIANCE ON USER DATA

X's business model relies on the ubiquitous collection of user data, in a manner that cannot be considered compatible with the company's responsibility to respect the right to privacy.[326] User data is central to X, as it helps the platform predict content users will engage with, and its quality largely decides how valuable X is to advertisers[327]

X's Privacy policy stipulates that some level of information must be provided to the company in order to open an account, making data collection a key requirement for accessing the platform's products and services.[328] Personal accounts require a display name, username, password, email address and phone number, date of birth, display language and third-party single sign-in information.[329] Platform users can also opt to share their location in their profile and posts, and to upload their address book to find people they know.[330]

X's Privacy policy outlines that data on preference settings is also collected, as well as other information about how users engage with the platform: "When you use our services, we collect information about how you use our products and services. We use that information to provide you with products and services, to help keep X more secure and respectful for everyone, and more relevant to you".[331] The focus on relevance speaks to the centrality of engagement in X's business model; the company collects data on users to recommend content which will keep them on the platform for longer, allowing X to collect more data on them. This ubiquitous corporate surveillance is at odds with the right to privacy and can have adverse consequences on the rights to freedom of thought, freedom of expression and non-discrimination.

The policy also makes clear that data will be collected specifically for making job and advertising recommendations, stating that X will collect and use personal information (such as employment history, educational history, employment preferences, skills and abilities, job search activity and engagement "and so on") to recommend potential jobs, enable employers to find potential candidates, and to show more relevant targeted advertising.[332]

In 2022 the FTC took action against X for deceptively using account security data for targeted advertising, resulting in a US$150 million penalty and a permanent injunction from profiting from the deceptively collected data.[333]

7.1.2 TARGETED ADVERTISING

Since 2013, almost all of X's revenue has come from targeted advertising on its site.[334] In 2021, advertising accounted for more than 90% of the company's US$5.1 billion revenue.[335]

As recently as 2023, it was clear that advertising remained a key source of income for X. The social media platform was hit by a 40% drop in revenue after more than 500 advertising clients paused their spending over concerns around the changes being made to X's policies.[336] X's Terms of Service, which were last updated on 15 November 2024, make clear the centrality of advertising on the platform: "You will see

[326] Cornelius Puschmann and Jean Burgess, "The politics of Twitter data", 23 January 2013, HIIG Discussion Paper Series, No. 2013-01, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2206225
[327] Cornelius Puschmann and Jean Burgess, "The politics of Twitter data" (previously cited).
[328] X, "Privacy policy" https://x.com/en/privacy#update. Accessed on 5 July 2025
[329] X, "Privacy policy" (previously cited). Accessed on 5 July 2025
[330] X, "Privacy policy" (previously cited). Accessed on 5 July 2025
[331] X, "Privacy policy" (previously cited). Accessed on 5 July 2025 (Emphasis added).
[332] X, "Privacy policy" (previously cited). Accessed 5 July 2025
[333] FTC, "A look behind the screens: examining the data practices of social media and video streaming services", September 2024, https://www.ftc.gov/system/files/ftc_gov/pdf/Social-Media-6b-Report-9-11-2024.pdf
[334] BBC News, "How does Twitter make money?", 7 November 2013, https://www.bbc.co.uk/news/business-24397472
[335] Bloomberg UK, "Documents show how Musk's X plans to become the next Venmo", 18 June 2024, https://www.bloomberg.com/news/articles/2024-06-18/documents-show-how-musk-s-x-plans-to-become-the-next-venmo; The Guardian, "Twitter hit by 40% revenue drop amid ad squeeze", 18 January 2023, https://www.theguardian.com/technology/2023/jan/18/twitter-revenue-drop-advertising-squeeze-elon-musk; Mashable, "Elon Musk's X revenue has officially plummeted", 18 June 2024, https://mashable.com/article/twitter-x-revenue-falls-x-payments-plans
[336] The Guardian, "Twitter hit by 40% revenue drop amid ad squeeze" (previously cited); Mashable, "Elon Musk's X revenue has officially plummeted" (previously cited).

advertising on the platform: In exchange for accessing the Services, X and our third-party providers and advertisers may display advertising to you."[337]

There are three main ways to advertise on X – promoting a tweet that will appear in people's timelines, promoting a whole account, or promoting a trending topic.[338] Like many social media companies, X tends to charge advertisers according to the amount of interaction their content generates, and advertisers pay per click or per retweet,[339] incentivising the platform to gather as much user data as possible to target advertisements as accurately as possible, ensuring a high number of clicks or retweets. X also has a "bidding system" in which advertisers compete to have their content appear in a particular space on the platform.[340]

At the time of writing, X is no longer publicly traded, making it difficult to obtain up-to-date information on the company's sources of revenue.[341] Most of the reports on revenue, including revenue issues, have come from internal leaks, rather than official sources.[342] It has been reported that, in the first six months of 2023, X's revenue fell by nearly 40% from the same period in 2022, and the company lost US$456 million in the first quarter of 2023.[343]

7.1.3 ALTERNATIVE SOURCES OF REVENUE

Since taking over the company in 2022, Elon Musk has made changes to the business model to create streams of revenue which are not dependent on advertising. This has included the X Premium subscription plan and a subscription service for creators.[344] However, neither service has yet been able to close the revenue gap left by the advertiser exodus.[345]

X has also sought to obtain a licence to become a money transmitter, in order to create an X Payment service as part of Musk's ambitions to expand the platform into an "everything app".[346] However, according to internal documents, X plans to use the payments service mainly to achieve "increased participation and engagement" on the social media platform and the intention is that X Payments does not plan to charge fees for most of its services,[347] suggesting that, despite seemingly significant changes to the business model, X will remain focused on generating engagement.

X PREMIUM

X Premium is an opt-in, paid subscription that offers additional features to users that "improve your experience" of the platform by elevating "quality conversations", according to X.[348] There are three tiers available as part of X Premium: basic, premium and premium+.[349] Each tier allows users to access greater algorithmic amplification, such as by allocating "reply prioritization", meaning their replies are more visible on the platform, as well as additional tools for content creation.[350]

The basic tier allows additional features including post editing, longer posts and longer video uploads, reply prioritization, text formatting, bookmark folders and custom app icons.[351]

The premium tiers allow all of the above as well as a "blue tick" checkmark (previously used as a symbol of verification), reduced ads, access to apply to ads for revenue sharing and creator subscriptions, larger reply prioritization, ID verification, access to a media studio and access to Grok, a generative AI chatbot developed by xAI.[352]

---

[337] X, "Terms of Service", 15 November 2024, https://x.com/en/tos
[338] BBC News, "How does Twitter make money?" (previously cited).
[339] BBC News, "How does Twitter make money?" (previously cited).
[340] BBC News, "How does Twitter make money?" (previously cited).
[341] Mashable, "Elon Musk's X revenue has officially plummeted" (previously cited).
[342] Mashable, "Elon Musk's X revenue has officially plummeted" (previously cited).
[343] Mashable, "Elon Musk's X revenue has officially plummeted" (previously cited).
[344] Mashable, "Elon Musk's X revenue has officially plummeted" (previously cited).
[345] Mashable, "Elon Musk's X revenue has officially plummeted" (previously cited).
[346] Bloomberg UK, "Documents show how Musk's X plans to become the next Venmo" (previously cited); Mashable, "Elon Musk's X revenue has officially plummeted" (previously cited).
[347] Mashable, "Elon Musk's X revenue has officially plummeted" (previously cited).
[348] X, "About X Premium", https://help.x.com/en/using-x/x-premium (accessed on 2 July 2025).
[349] X, "About X Premium" (previously cited).
[350] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441 (previously cited).
[351] X, "About X Premium" (previously cited).
[352] X, "About X Premium" (previously cited).

Premium+ includes all the premium features as well as additional benefits such as no ads anywhere on the platform and the largest reply prioritization.[353]

Maja Heban, a trans woman and LGBTI activist based in Warsaw, outlined to Amnesty International her view that the Premium feature had made X less safe:

"The way monetization works nowadays, where you can pay money to become a verified account and then be paid for creating engagement means that... people are encouraged to create engagement, even if it means making stuff up, fear mongering, spreading fake news, harassing people... As long as people reply to you and say that you are lying, you are gaining something, so they incentivize spreading misinformation in a way and spreading hate speech."[354]

7.2 ENGAGEMENT-BASED ALGORITHMS AND THE ARCHITECTURE OF X'S RECOMMENDER SYSTEM

This section will examine how X's recommender system works, outlining weightings given to various interactions that users may have on the platform, to explore how the platform increases engagement and personalization. This recommender system analysis shows that thoughtfully engineered safeguards, reinforced by genuine community engagement processes, could have substantially mitigated the system's potential harms. Instead, X appears to have prioritized engagement metrics, leaving these protections either weakly implemented or altogether absent.

Surveillance-based business models tend to prioritize maximizing 'user engagement' above all else; the longer someone stays on a platform, the more data can be gathered about them, and the more precisely they can be targeted with advertising.[355] Amnesty International has previously found that this business model can lead to recommender algorithms boosting content which is inflammatory, discriminatory and divisive, because such content is often what engages platform users the most.[356]

This algorithmic boosting is in part a result of personalized recommendations. On X, personalized recommendations are made for tweets, events, topics, hashtags and users.[357]

As a platform, X features two timelines – "Following" and "For You". The platform's recommendation algorithm's key focus is the For You timeline, which is designed to show users new content from accounts they do not already follow, as well as content from accounts they do follow, and is considered the platform's main feed.[358] The For You timeline was unveiled in January 2023 as part of a redesign of the site.[359] X's feeds originally showed tweets from the accounts a user followed chronologically, later showing posts liked by or replied to by a followed account.[360] Before 2022, X had begun showing recommendations of posts "You might Like", and the For You page leans into this model of engagement, moving away from the chronological feed.[361] X now defaults to the For You timeline.[362]

The foundation of X's algorithmic recommender system is a set of core models and features that extract latent information from tweet, user and engagement data.[363]

In a publicly available blog post from 2023, X describes this model as trying to answer questions such as "What is the probability you will interact with another user in the future?" or "What are the communities on Twitter and what are the trending tweets within them?"[364] The detail and analyses of X's recommender

[353] X, "About X Premium" (previously cited).
[354] Amnesty International interview with Maja Heban, 30 July 2024.
[355] Amnesty International, *Surveillance Giants* (previously cited).
[356] Amnesty International, *: The Social Atrocity: Meta and the Right to Remedy for the Rohingya* (Index: ASA 16/5933/2022), 28 September 2022, https://www.amnesty.org/en/documents/asa16/5933/2022/en/; Amnesty International, *"A Death Sentence for My Father": Meta's Contribution to Human Rights Abuses in Northern Ethiopia* (Index: AFR 25/7292/2023), 31 October 2023, https://www.amnesty.org/en/documents/afr25/7292/2023/en/
[357] Kayla Duskin and others, "Echo chambers in the age of algorithms: an audit of Twitter's friend recommender system", May 2024, WEBSCI24: Proceedings of the 16th ACM Web Science Conference, https://dl.acm.org/doi/abs/10.1145/3614419.3643996
[358] X, "Twitter's recommendation algorithm", 31 March 2023, https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm
[359] Washington Post, "Elon Musk's Twitter pushes hate speech, extremist content into 'For You' pages", 30 March 2023, https://www.washingtonpost.com/technology/2023/03/30/elon-musk-twitter-hate-speech/
[360] Washington Post, "Elon Musk's Twitter pushes hate speech, extremist content into 'For You' pages" (previously cited).
[361] Washington Post, "Elon Musk's Twitter pushes hate speech, extremist content into 'For You' pages" (previously cited).
[362] Washington Post, "Elon Musk's Twitter pushes hate speech, extremist content into 'For You' pages" (previously cited).
[363] X, "Twitter's recommendation algorithm" (previously cited).
[364] X, "Twitter's recommendation algorithm" (previously cited).

system's architecture are drawn from this blog post, and from Amnesty International's own analysis of the elements of the source code that were made publicly available in 2023 by X.[365]

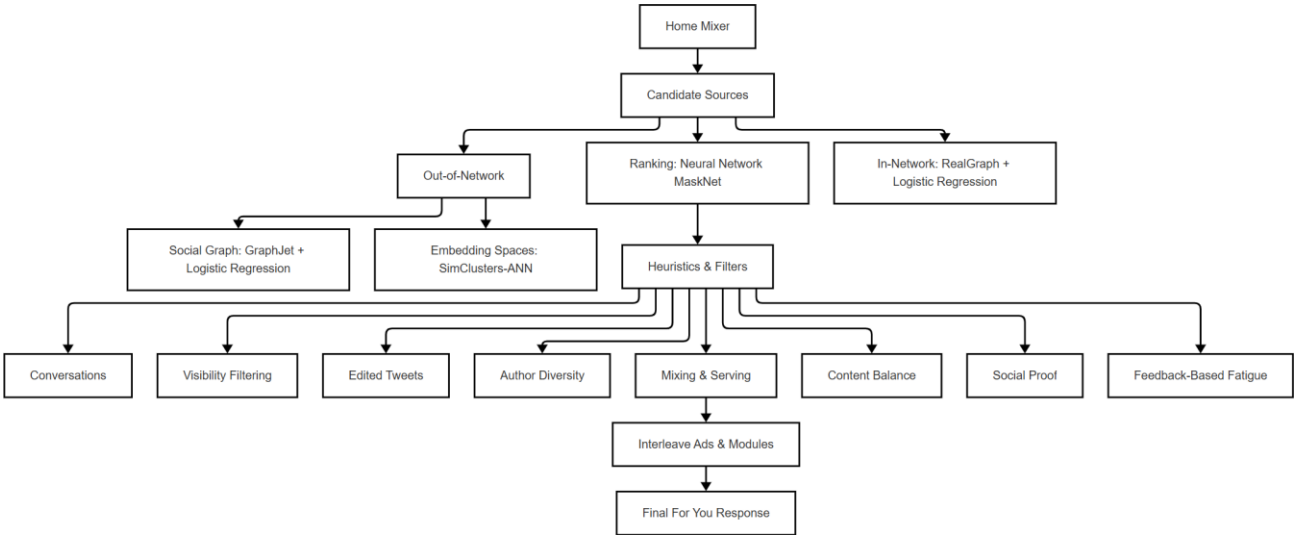
The recommendation pipeline has three main features:[366]

- Candidate sourcing (this fetches the most engaging tweets from different recommendation sources)..
- Ranking each candidate tweet to assign a probability score of the user engaging with the piece of content. The model predicts the likelihood of a range of interactions including whether the user will like the tweet, retweet it, reply, click on it, or even flag it as inappropriate.[367]
- Applying heuristics and filters, for example filtering out tweets from blocked users, 'not safe for work' content, and tweets that have already been seen.

The For You timeline is shaped by integrating these three features of the pipeline together and then applying boosting logic (amplifying specific tweets). Together, this service is known as the Home Mixer.[368] The Home Mixer pipeline runs approximately 5 billion times each day and completes in under 1.5 seconds on average, resulting in 150 billion tweets served to people's devices every single day.[369]

This is graphically visualized, with all technical detail and approaches employed, in Figure 5.

FIGURE 5: GRAPHICAL REPRESENTATION OF THE HOME MIXER PIPELINE THAT GENERATES USERS' FOR YOU FEED



To generate a personalized feed, the recommendation system must first retrieve a pool of "candidate tweets" that are potentially relevant to the user. X employs a candidate selection process that draws from two primary areas: in-network content (tweets from accounts the user follows) and out-of-network content (tweets from other accounts).[370] On average, the system pulls about 1,500 candidate tweets per user request[371], roughly half from each category.[372] This ensures a mix of familiar and new content in the For You timeline.

After a set of candidate tweets is assembled, X's recommendation system employs a set of machine-learning ranking algorithms[373] to score these candidates for the user. This stage is the heart of the personalization engine where a large-scale neural network model predicts how each user will react to each tweet and assigns a relevance score accordingly.

---

[365] See, https://github.com/twitter/the-algorithm
[366] X, "Twitter's recommendation algorithm" (previously cited).
[367] This is done using a deep neural network and this model is not available as open source. See, Kevin Feng and others, "Probing the ethical boundaries of personalization: a case study of Twitter's recommendation algorithm", 2024, CSE 581 - Computing Ethics, https://homes.cs.washington.edu/~micibr/assets/pdf/ethical_personalization_paper.pdf
[368] X, "Twitter's recommendation algorithm" (previously cited).
[369] X, "Twitter's recommendation algorithm" (previously cited).
[370] Aneesh Sharma and others, "GraphJet: real-time content recommendations at Twitter", 2016, Proceedings of the VLDB Endowment, Volume 9, Issue 13, https://www.vldb.org/pvldb/vol9/p1281-sharma.pdf
[371] This refers to each requested post – each piece of content that comes up on a user's "For You" feed
[372] X, "Twitter's recommendation algorithm" (previously cited).
[373] This model is not available in open-source code.

As of 2023, this ranking was performed by a deep neural network with around 48 million parameters. This model is continuously trained on the platform's enormous interaction logs, meaning it learns from the collective behaviour of X's users in near-real-time. Every time users either engage with (or ignore) tweets, it provides training data about what content tends to succeed for which audiences.

The model uses thousands of input features encompassing all aspects of the user, the tweet, and their interaction.[374] User features include demographics, inferred interests and past activity. Tweet features include text embeddings, author, engagement stats and community indicators. User-tweet features include whether the user follows the author, how often the user has interacted with similar tweets, and whether the tweet was recommended by a friend. Given all these inputs, the neural network produces a set of predicted probabilities for different engagement outcomes; for example, the probability that the user will like the tweet, retweet it, reply, click on it, or even flag it as inappropriate.[375]

According to X, the model is a multi-task learner that produces around 10 prediction scores per tweet (each corresponding to a specific user action of interest).[376] To convert these predictions into a score, X's system applies a hard-coded weighted formula that prioritizes certain actions more than others.

While a version of the ranking model is open-sourced (including its architecture and hyperparameters and a dummy training pipeline), the real model weights used in production were not provided. X cited privacy reasons for this; the released model might be re-trained on public data or partially randomized.[377] However, we can still draw inferences from the publicly available weights, which are detailed in Figure 6 below.

FIGURE 6: WEIGHTINGS FOR EACH PREDICTED PROBABILITY

FEATURE	WEIGHT	DESCRIPTION
FAVOURITE (LIKE)	0.5	Predicted probability of the user "favouriting" (liking) a tweet: very low influence on the final ranking score.
RETWEET	1.0	Predicted probability of the user retweeting: a light signal, only marginally more than a like.
REPLY	13.5	Predicted probability of the user replying: strongly boosts tweets that spark direct conversation.
GOOD PROFILE CLICK	12.0	Probability the user clicks into the author's profile and then likes/replies: valued nearly as much as a reply.
VIDEO PLAYBACK >50%	0.005	Probability the user watches more than 50% of a video: effectively zero impact on ranking.
REPLY ENGAGED BY AUTHOR	75.0	Probability the user replies, and the author subsequently engages: highest reward for sustained back-and-forth.
GOOD CLICK (CONVERSATION OPEN)	11.0	Probability the user opens the conversation view and then likes/replies: signals deep conversational interest.
GOOD CLICK V2 (2-MIN CONVERSATION VIEW)	10.0	Probability the user stays more than two minutes in the conversation view: strong indicator of engagement depth.
NEGATIVE FEEDBACK	–74.0	Probability of negative feedback (for example, "show less," block or mute): heavily penalizes disliked or unwanted content.
REPORT	–369.0	Probability the user reports the tweet: significantly demotes content deemed offensive or problematic.

[374] Anthony Alford, "Twitter open-sources recommendation algorithm", 11 April 2023, https://www.infoq.com/news/2023/04/twitter-algorithm/
[375] Kevin Feng and others, "Probing the ethical boundaries of personalization: a case study of Twitter's recommendation algorithm" (previously cited).
[376] X, "Twitter's recommendation algorithm" (previously cited).
[377] See, https://raw.githubusercontent.com/twitter/the-algorithm-ml/main/projects/home/recap/README.md#:~:text=contributes%20a%20near,you%20can%20run%20the%20model

As shown in Figure 6, not all forms of engagement are treated equally. The platform tends to value involved interactions (like replies or lengthy dwell time) more heavily than passive ones (such as a quick "like"). For instance, if the model believes a user is very likely to reply to a particular tweet, that tweet will be ranked higher in the feed, since replying is seen as a strong indicator of engagement. On the other hand, if the model detects a high probability that the user would give negative feedback on a tweet, such as muting the author or reporting the tweet, that content will be downranked or filtered out aggressively.[378] It is important to note that this ranking is specific to the user in question and their personalized feed, meaning that any downranking that may be applied on the basis of predicted negative feedback is not universal, and does not serve as an adequate mitigation measure to countering, and not amplifying, harmful or hateful content on the platform.

This machine learning-driven ranking is what tailors the timeline to each user. Two users with identical candidate pools will receive different ranked feeds if their past behaviour differs, because the model has learned different preference profiles for them. Importantly, the ranking model is periodically retrained and updated (and possibly fine-tuned online) to adapt to evolving trends and user tastes. X's blog also notes that the model is continuously refined on fresh interaction data to keep recommendations up to date with "what's happening now" on the platform.[379]

Overall, the example weightings indicate the main priority for the ranking is to generate conversation and engagement quality as they are heavily incentivized, while negative user reactions are harshly penalized. The single highest weighted action is Reply Engaged by Author which, at +75, is much higher than all the others. This indicates that the model promotes tweets that spark a response from the author.

After the ranking of the tweets, the Home Mixer applies a series of heuristics and business rules to filter and refine the content shown to each user. These aim to ensure there is sufficient diversity in each user's feed or remove content which violates X's content or policy rules. For example, the visibility and safety filters eliminate tweets from accounts a user has blocked or muted, while another filter implements "feedback-based fatigue" which lowers the score of certain tweets if the viewer has provided negative feedback – such as clicking "show less" - pertaining to them.[380]

7.3 RISKS OF ENGAGEMENT-BASED ALGORITHMS

As detailed above, X's recommender system architecture is built around maximizing user engagement, measured by actions such as likes, retweets, replies and time spent on the platform. Despite including select mitigation measures in the form of the "layered heuristics" (such as social safety filters), these lightweight interventions face technical trade-offs and remain secondary to the primary engagement-first objective embedded within the recommender system's design. As a result, the ability of these mitigation measures to curb the human rights risks of the engagement-based business model is limited by the overriding aim of boosting engagement.

By prioritizing engagement, the algorithm is incentivized to show users content that will generate interaction. Even with safeguards, many of which have recently been removed or significantly reduced, there are significant human rights risks inherent to the business model. Most notably, the recommender system risks leading to the amplification of harmful content that prompts strong reactions to retain a cycle of engagement.[381] Most studies into algorithmic amplification on social media platforms have shown that, if users begin to interact with harmful content, they are subsequently shown more of it by recommender algorithms.[382] For example, a 2023 Washington Post investigation found that accounts that followed "extremists" were subjected to a mix of other racist and incendiary speech.[383] Many of the users amplified in the For You timeline were previously suspended by X and then reinstated by Elon Musk following his takeover. Elon Musk pledged to dampen the spread of hate speech on the site, saying: "New Twitter policy is

[378] Stacey McLachlan, "The X (Twitter) algorithm explained: 2024 guide", 7 October 2024, https://blog.hootsuite.com/twitter-algorithm/
[379] X, "Twitter's recommendation algorithm" (previously cited).
[380] X, "Twitter's recommendation algorithm" (previously cited).
[381] Faculty of Public Health, "Response to 'Social media, misinformation and harmful algorithms', inquiry call for evidence", n.d., https://www.fph.org.uk/media/hoejpp0s/social-media-consultation-fph-response.pdf; Joe Whittaker and others, "What are the links between social media algorithms, generative AI and the spread of harmful content online?" Written evidence to the UK Parliament Science, Innovation and Technology Committee (SMH0018), 17 December 2024, https://committees.parliament.uk/writtenevidence/132875/pdf/
[382] Institute for Strategic Dialogue, "ISD written evidence to the Science, Innovation and Technology Inquiry on Social Media, Misinformation and Harmful Algorithms", 2025, https://www.isdglobal.org/wp-content/uploads/2025/01/ISD-Written-Evidence-to-the-Science-Innovation-and-Technology-Committee-Inquiry-on-Social-Media-Misinformation-and-Harmful-Algorithms.pdf; Joe Whittaker and others, "What are the links between social media algorithms, generative AI and the spread of harmful content online?" (previously cited).
[383] Washington Post, "Twitter pushes hate speech, extremist content into 'For You' pages" (previously cited).

freedom of speech, but not freedom of reach. You won't find the tweet unless you specifically seek it out, which is no different than the rest of the internet."[384] Jakub Szymik, a gay man based in Warsaw, told Amnesty International that he believes that X's focus on engagement has an adverse effect on platform users:

"Twitter's architecture of short snappy comments and polarizing algorithm impacts how people communicate online and offline and there are real world impacts of those actions. I think there is a very deep connection, and we could take this into consideration when thinking about all the platforms."[385]

X acknowledges the role that amplification plays in its recommendations: "Recommendations may amplify content, so it's important they are surfaced responsibly".[386] The company also stipulates that promoting healthy conversations is one of X's core principles and, as such, "freedom of speech is a fundamental human right – but freedom to have that speech amplified on X is not".[387] However, content which cannot be recommended (and therefore amplified) due to X's platform rules will still be available on X to people who follow the post author and on the post author's profile.[388] Content ineligible for recommendations includes content that violates any of X's rules but has been left on the platform because of the public-interest exception, which may include content that is deemed to be marginally abusive, harmful or misleading.[389] As well as individual pieces of content, accounts can also become ineligible for recommendations for the same reasons.[390]

X also allows for a limited amount of user control over recommendations on the For You and Following timelines. Users can mute and lock notifications on the Home timeline, or flag that they are not interested in a post or topic.[391]

While X has claimed to be transparent about its recommender algorithm, releasing the code in 2023, the DSA-mandated independent audit of X's risk assessment found that the company's terms of service do not adequately represent or explain the main parameters used in its recommender systems. Though some information is available in its Rules and Policies pages, it is not comprehensive enough.[392] The audit recommended that X include in its terms of service clear and understandable explanations of the parameters used within the recommender systems, as well as providing specific details about the criteria used and relative importance of each parameter.[393]

7.4 ECHO CHAMBERS

Some academic research into X has noted that the way in which the platform recommends content may lend itself to the creation of echo chambers[394] or 'filter bubbles', which expose users to ideologically homogenous content which is usually in line with their existing beliefs.[395] The phenomenon of echo chambers has been observed across many social media platforms and is not exclusive to X.[396]

A key tenet of echo chambers is interaction between two users with similar opinions – to achieve a high level of engagement on the platform.[397] Users in echo chambers can be understood as "users who share a common discourse, are exposed to the same news sources, and are exposed to the same opinions", often retweeting each other.[398]

---

[384] Washington Post, "Twitter pushes hate speech, extremist content into 'For You' pages" (previously cited).
[385] Amnesty International video call with Jakub Szymik, 5 August 2024.
[386] X, "About our approach to recommendations", n.d., https://help.x.com/en/rules-and-policies/recommendations#:~:text=We%20recommend%20posts%20to%20you,by%20those%20in%20your%20network.
[387] X, "About our approach to recommendations" (previously cited).
[388] X, "About our approach to recommendations" (previously cited).
[389] X, "About our approach to recommendations" (previously cited).
[390] X, "About our approach to recommendations" (previously cited).
[391] X, "About our approach to recommendations" (previously cited).
[392] FTI Consulting, "X Independent Audit" (previously cited).
[393] FTI Consulting, "X Independent Audit" (previously cited).
[394] In the context of social media, an echo chamber or "filter bubble" is the phenomenon in which a group of users primarily interact with and consume information from others who share similar beliefs, opinions and viewpoints. This can lead to the reinforcement of pre-existing beliefs and a reduction in exposure to diverse perspectives.
[395] Kayla Duskin and others, "Echo chambers in the age of algorithms: an audit of Twitter's friend recommender system" (previously cited); Manuel Pratelli and others, "Entropy-based detection of Twitter echo chambers", May 2024, PNAS Nexus, Volume 3, Issue 5, https://academic.oup.com/pnasnexus/article/3/5/pgae177/7658380
[396] Manuel Pratelli and others, "Entropy-based detection of Twitter echo chambers" (previously cited).
[397] Manuel Pratelli and others, "Entropy-based detection of Twitter echo chambers" (previously cited).
[398] Manuel Pratelli and others, "Entropy-based detection of Twitter echo chambers" (previously cited).

Interviewees told Amnesty International that they often had the impression that X was creating echo chambers:

"I do wonder if on Twitter if you see a tweet, it [the algorithm], might propose some similar accounts and its effect is a rabbit hole."[399]

A 2024 case study into echo chambers on X found that "users in echo chambers, while representing a small minority, strongly contribute to the debate, often disseminating misinformation."[400]

The study found that the results of the phenomenon can be long-lasting. After two years, the users trapped in echo chambers observed by the researchers held the same opinions *and* had become more extreme.[401] The researchers observed that the extreme views held by these users were not limited only to the initial topic that the researchers tracked (Covid-19 vaccination conspiracy theories), but that, after two years, the users held "extreme views on current controversial issues such as the war in Ukraine, migrants, and LGBT issues".[402]

To gain an understanding of the extent to which the research accounts in Amnesty International's quantitative research were subject to personalization, researchers analysed the type of accounts that were recommended to each sub-group under "Who to Follow", and subsequently the political partisanship of the accounts that were present on their For You feed.

Figure 7 below presents the political partisanship of the accounts that were recommended to each sub-group of research accounts. To interpret the table, we observe that across all research accounts in the 'Civil' group,(see methodology section) 527 of the accounts they were recommended to follow were also politicians belonging to, or accounts that are aligned with, the political parties that support civil rights (outlined in the methodology section).

As shown below, the evidence of personalization within the "Who to Follow" recommendations is strong, with the recommendations clearly aligning with the political partisanship of the research accounts. This demonstrates how echo chambers can be easily created for users, with the recommendations of "Who to Follow" closely aligning with their existing follower list.

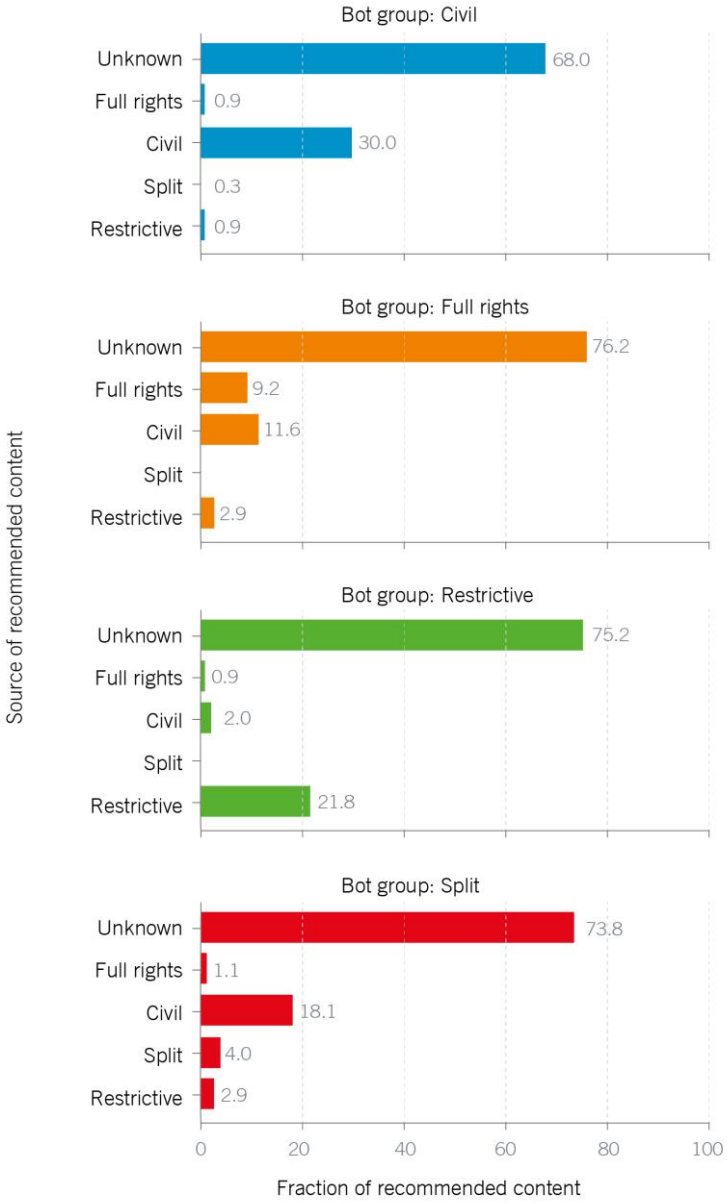
[399] Amnesty International interview with Aleksy (pseudonym), 28 July 2024.
[400] Manuel Pratelli and others, "Entropy-based detection of Twitter echo chambers" (previously cited).
[401] Manuel Pratelli and others, "Entropy-based detection of Twitter echo chambers" (previously cited).
[402] Manuel Pratelli and others, "Entropy-based detection of Twitter echo chambers" (previously cited), p. 5.

FIGURE 7: POLITICAL PARTISANSHIP OF ACCOUNTS PRESENTED IN THE "WHO TO FOLLOW" RECOMMENDATIONS

Group	Number of accounts presented in "Who to Follow" recommendations				
	Civil	Full Rights	Restrictive	Split	Unknown
Civil	527	38	4	-	462
Full Rights	26	712	1	5	297
Restrictive	2	735	-	-	277
Split	48	2	6	350	624

To assess the risk of echo chambers being created on the For You feeds, Amnesty International researchers analysed the partisanship of the accounts present on each sub-group's algorithmic timelines. Figure 8, below, details the findings from this analysis. It suggests that, outside of tweets posted directly by Elon Musk, there is further evidence of personalization on the research accounts' For You feeds. Amnesty International was not easily able to determine political partisanship and level of support for the rights of LGBTI people for all accounts shown on the research accounts 'For You' feed. However for accounts where this categorization was possible there is a clear alignment between accounts recommended to the research accounts via the "For You" feed, and the partisanship of the politicians those same research accounts follow.

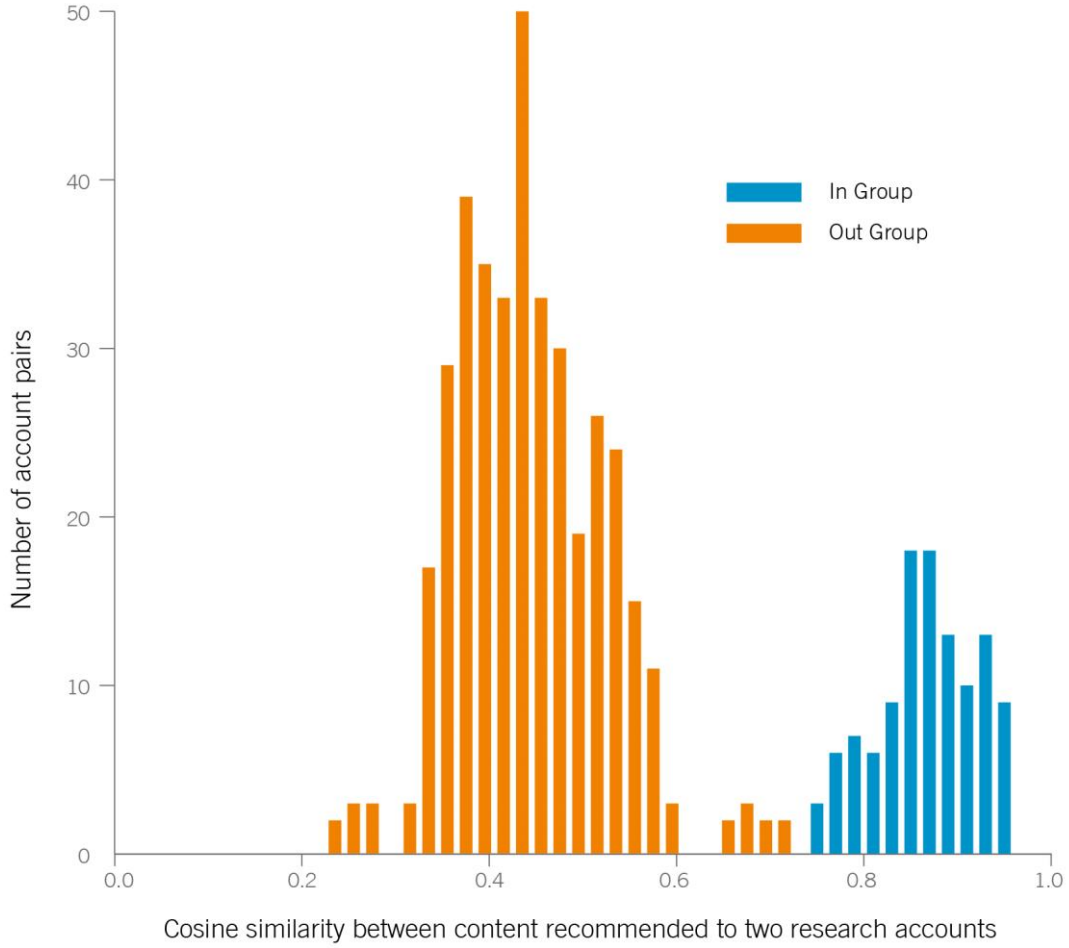
FIGURE 8



Further, Amnesty International also compared the similarity of the content presented to research accounts within the same sub-group, compared to those in different sub-groups. Amnesty International researchers compared the recommended content for pairs of research accounts (eg 'Civil' research account 1 versus 'Full Rights' research account 1).

Figure 9 below shows that research accounts who follow the same set of accounts (blue bars) are recommended more similar content than if they do not (orange bars). While not explicitly commenting on the nature of the content, this finding confirms that the recommended content is indeed personalized based on which accounts the research accounts follow and is not random.

FIGURE 9



7.5 PRIORITIZING THE 'TOWN SQUARE' OVER MITIGATION MEASURES

CONTENT WARNING

This section contains examples of content which include graphic calls for violence and discrimination, which may be distressing for some readers.

X's mitigation strategies appear to be based on preserving the platform's position as a digital 'town square' through allowing unfettered freedom of expression in a manner that is patently inconsistent with international human rights law and standards. In the first risk assessment produced under the DSA, the platform made clear that this remains a key priority in its decision-making processes, reporting that "X strives to be the town square of the internet by promoting and protecting freedom of expression. We have always understood that to reach this goal we must give everyone the power to create and share ideas and information instantly,

without barriers."[403] The second risk assessment produced under the DSA describes how X gives "special consideration" to the effect on freedom of expression when choosing mitigation measures.[404]

The absolutist approach to freedom of expression taken by X is at odds with international human rights law and standards. While the right to freedom of expression must be protected, it is not an absolute right and must be balanced with other rights such as the right to non-discrimination and the right to live free from GBV. The decision by X to allow freedom of expression with very few restrictions presents an unacceptable level of risk to platform users from marginalized communities, including the LGBTI community in Poland.

This inappropriate prioritization of freedom of expression over other rights has led X to approach content moderation outside of what it calls a "binary, absolutist take down/leave up approach", with many of its mitigation strategies for harmful content being focused on limiting the reach of content which violates platform policies.[405] According to X, restricted posts receive 81% less reach or impressions, on average, than an unrestricted post and the platform also seeks to prevent adverts from appearing adjacent to content which has been labelled as harmful.[406]

In its 2024 risk assessment, the platform acknowledged: "There is a risk that exposure of private content could impact an individual's physical safety, emotional wellbeing, psychological health and financial security."[407]

The independent audit of X's risk assessment, which was submitted to the European Commission as part of the company's obligations under the DSA, concluded that X's risk assessment process was not rigorous enough.[408] The audit found that X needs to conduct a full risk assessment for each of its recommender systems to identify systemic risks, define the role and purpose of the recommender systems, establish metrics for effectiveness and continuously monitor the risks posed by these systems.[409] The audit also recommended that X conducts a risk assessment on what it calls its "Freedom of Speech, Not Reach" system.[410] Similarly, the audit found that X's risk mitigation measures are ineffective at reducing systemic risks and found a lack of mitigation measures relating to algorithmic systems, among other things.[411]

X's irresponsible and cavalier approach to harmful content is evidenced in a report published in 2024 by the Polish civil society organization Never Again Association. The organization is registered as a Trusted Flagger by an online monitoring project financially supported by the EU's Citizens, Equality, Rights and Values programme, Between August 2023 and August 2024, Never Again Association reported 343 examples of hateful content to X over a 12 month period.[412] The organization initially reported the posts through X's regular user interface and, if it received no response or the content was not removed, it then reported the cases through the Never Again Association's X account.[413] In most of the cases, X either refused to remove the posts (Never Again Association reported only a 10% removal rate on its reports) or ignored the reports.[414] The posts – which included text, image and videos, could be seen as inciting hatred against minorities, including LGBTI people.[415]

Some of the posts which Never Again Association reported to X but were not removed specifically targeted the LGBTI community and could be considered incitement to violence and advocacy of hatred. For example, one post which was reported but received no action from X read: "Fuck gender. Fuck the perverted whores. Fuck transvestites. Load those whores into the furnace!!"[416]

---

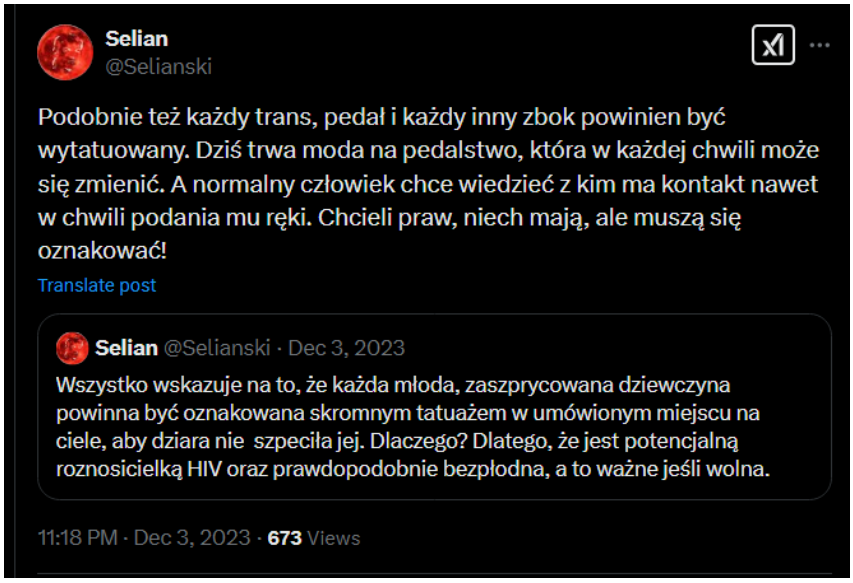
[403] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023, https://transparency.x.com/content/dam/transparency-twitter/dsa/dsa-sra/dsa-sra-2023/TIUC-DSA-SRA-Report-2023.pdf
[404] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).
[405] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited), p. 8.
[406] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited).
[407] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).
[408] FTI Consulting, "X Independent Audit" (previously cited).
[409] FTI Consulting, "X Independent Audit" (previously cited).
[410] FTI Consulting, "X Independent Audit" (previously cited).
[411] FTI Consulting, "X Independent Audit" (previously cited).
[412] Never Again Association, "The Twitter standards of hate (August 2023 – August 2024)", 2 September 2024, https://www.nigdywiecej.org/docstation/com_docstation/172/the_twitter_standards_of_hate.pdf
[413] Never Again Association, "The Twitter standards of hate (August 2023 – August 2024)" (previously cited).
[414] Never Again Association, "The Twitter standards of hate (August 2023 – August 2024)" (previously cited).
[415] Never Again Association, "The Twitter standards of hate (August 2023 – August 2024)" (previously cited).
[416] Never Again Association, "The Twitter standards of hate (August 2023 – August 2024)" (previously cited).

Several posts reported by Never Again Association during the year to August 2024 remained visible on the platform as of May 2025. These tweets are documented below and include posts that portray the LGBTI community as deviants, use slurs and call for discrimination against the LGBTI community.

A post from an X user, Antoni Kocemba, which translates as: "They are just leftist faggots. We will not get far with them."[417]

A post from the Konfederacja party, which translates as: "We don't want deviants, promoters of deviance and ostentatious professional sodomites teaching our children tolerance."[418]

[417] Antoni Kocemba, X post, 31 July 2023, https://x.com/antoni_kocemba/status/1685950832657797120
[418] Konfederacja, X post, 28 July 2023, https://x.com/KONFEDERACJA_/status/1684882568087543808



A post by X user Selian which reads: "Similarly, every trans, faggot and every other person should be tattooed. And a normal person wants to know whom he is in contact with, even when shaking hands. They wanted rights, let them have them, but they have to label themselves!"[419]

These posts, still circulating on the platform as of May 2025, are clear evidence of the harmful content which has become normalized on X due to its unfettered approach to freedom of expression, which X uses to justify a negligent approach to content moderation. Even when receiving reports of content which could be considered incitement to violence and advocacy of hatred towards the LGBTI community, X appears to ignore the prevalence of harmful content on the platform, without considering the risk that this content presents to the rights of marginalized individuals. This includes their own right to freedom of expression, since many of the LGBTI community members interviewed by Amnesty International referred to their self-censorship on X. The lack of serious consideration for rights other than freedom of expression is reflected throughout X's risk assessments for 2023 and 2024, which do not meet an acceptable level of human rights due diligence under international human rights standards.

7.6 LACK OF ENGAGEMENT WITH CIVIL SOCIETY

An important factor in assessing X's responsibility for undermining the human rights of the LGBTI community in Poland is the foreseeability of the company contributing to human rights harms. According to international human rights standards, if a company knows or *should know* that it risks contributing to human rights harms, then it has a responsibility to take the necessary steps to cease or prevent its contribution and use its leverage to mitigate any remaining negative effects to the greatest extent possible.[420] To this end, companies are encouraged to engage with relevant stakeholders to identify and mitigate risks. Stakeholder engagement is also a necessary element of producing risk assessments under the DSA.

However, Amnesty International found that, since at least 2022, X has had very little proactive engagement with Polish civil society organizations working with the LGBTI community to discuss mitigating risks on the platform. For example, an interviewee working at one of the most prominent LGBTI civil society organizations in Poland told Amnesty International that he was unaware of any communication between the organization and X.[421] Similarly, Mateusz Kaczmarek, a board member at Grupa Stonewall, told Amnesty International that X had never reached out to the group to discuss possible risks or risk mitigation measures.[422]

Julia Kata, a psychologist at the LGBTI organization Fundacia Trans-Fuzja, told Amnesty International she was not aware of any consultation between X and LGBTI civil society organizations in Poland:

[419] Selian, X post, 3 December 2023, https://x.com/Selianski/status/1731452869792924087
[420] UN Guiding Principles, Principle 19 including Commentary.
[421] Amnesty International interview with Aleksy (pseudonym), 26 July 2024.
[422] Amnesty International interview with Mateusz Kaczmareck, 28 July 2024.

"We [Polish LGBTI organizations] are in this together so, more or less, we do speak to each other and probably if X approached one, two or three organizations, everybody would know, and they would ask to pass on their contact details because we would love to talk to them."[423]

Limited engagement with civil society is reflected in X's 2024 DSA risk assessment, which notes that the company has had a handful of engagements with civil society organizations, without providing detail on how many engagements were conducted nor on which areas of expertise or particular affected communities were involved in this exercise.[424] Furthermore, X's description of civil society engagements seems to focus on engagements that focus on teaching civil society organizations to better use the platform's reporting tools, rather than X drawing on the organizations' expertise regarding harmful content and marginalized communities.[425]

7.7 FAILURE TO ADEQUATELY MITIGATE SYSTEMIC RISKS

In X's 2024 DSA risk assessment, the company reported that its existing controls reduce the level of risk in most areas identified to a low or medium level.[426] However, the current and planned mitigations outlined in the risk assessment are limited to improvements to policies, content moderation systems (including enforcement and detection) and Community Notes awareness-raising measures,[427] which do not adequately address the risks inherent in X's business model, including a focus on algorithmically optimizing for engagement, or even the risks its current operations present to marginalized communities, for example through its poor content moderation resourcing.

X states that its recommender systems are designed to exclude harmful and "violating" content by integrating with visibility filtering systems and other systems, using content health prediction models to prevent harmful and violating content from ranking higher.[428] Additionally, X has a company policy, introduced in March 2023, to remove violent hate speech from the platform.[429] However, it appears that if recommender systems incorrectly allow harmful content to be algorithmically boosted, there are few robust mitigation measures to minimize harm since, according to its own risk assessment, the platform relies heavily on user controls such as muting notifications or limiting replies to posts [430].

The reliance on improvements to policies – particularly in a context where an increasingly permissive approach to harmful content has led to policies being degraded – has shown to be inadequate in mitigating systemic risks on the platform. For example, despite a policy to remove violent hate speech, most of the LGBTI activists interviewed by Amnesty International reported seeing, or being directly targeted with, such speech on the platform – repeatedly, and over several years.

Additionally, X acknowledges a risk that "personalisation of recommended content could in some circumstances also contribute to information bubbles, limiting users' access to pluralistic sources of information",[431] but does not outline any specific mitigation measures to address this.

X notes that comments, as well as posts, may present a risk to platform users who are purposefully exposed to hateful commentary, as tagging the author of the original post will notify the author.[432] Furthermore, according to X's latest risk assessment, it views hate speech as "illegal content" under the DSA framework.[433] However, as Poland does not specifically prohibit or criminalize hate speech targeting LGBTI

[423] Amnesty International interview with Julia Kata, 29 July 2024.
[424] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).
[425] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).
[426] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).
[427] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).
[428] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited).
[429] X, "Violent Content policy" (previously cited).

[431] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited), p. 36.
[432] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited).
[433] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).

people,[434] it is not clear how the platform would handle hateful content targeting LGBTI individuals if this was not linked to a call for violence.

This is of additional concern because X relies heavily on automated detection of violations of policies.[435] For slurs and tropes in particular, the company uses glossaries specific to EU languages.[436] This is, however, far from constituting adequate resourcing for content moderation, particularly as X has just two Polish-speaking content moderators.[437]

X states that, because of these mitigation strategies, its "data has shown that 99.99% of post impressions are on content that is deemed 'healthy'. Less than 0.01% of post impressions contain hateful language."[438] However, the company does not provide a breakdown of these figures by language or country.

Jakub Szymik told Amnesty International that he had seen how damaging hateful comments on X could be:

"I work with one LGBTQ organization that is led by someone with strong visibility on the platform and they use Twitter to amplify their work. And I see how he is impacted by swarms and masses of anonymous comments but also people using their names and sending things calling for violence or threats on the platform publicly. The mass communication of this and the waves of very violent content impacts him and the organization very much."[439]

He told Amnesty International that he often sees comments targeting LGBTI activists on X:

"Most situations I encounter are focused on a specific person speaking out and they get multiple comments that are very violent in nature and saying, 'someone should shoot you, someone should kill you, you shouldn't be able to speak up'."[440]

The mitigation measure for the risk of harmful content in comments is reply controls, which allow a user to limit who replies to their posts by either only allowing users mentioned in the post to reply or by turning off replies altogether.[441]

However, LGBTI rights activist Magda Dropek told Amnesty International that these tools were pre-emptive in nature and insufficient to adequately address the harm:

"What I have noticed on my social media in the last years – of course, it's very difficult to do something with very hateful messages. In my case for example, if someone is writing to me 'kill yourself', 'no one wants you here', 'you're like garbage for this country', and for example I have hundreds of messages like this and comments like this. For me, I have the tools to cope with it. But what is important for me is that very often the community which is following me will see those messages. This is something [to which] I feel completely vulnerable because especially after Twitter became X, it's like the tools [on the platform] are very difficult now."[442]

X is well aware of the risk of individuals and groups being targeted with hateful content or abuse on the platform. In its 2024 DSA risk assessment, X reports that this could create a sense of fear and intimidation and lead to self-censorship, and notes that the platform may be misused to promote hate or incite hostility, discrimination and violence,[443] as experienced by the LGBTI community members interviewed by Amnesty International.

However, once again, the mitigation measures for these risks are wholly inadequate, being limited to reviews of policies and processes and the Community Notes function, which essentially outsources content

[434] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited).
[435] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited).
[436] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited).
[437] X, *DSA Transparency Report – April 2025* (previously cited).
[438] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited), p. 25.
[439] Amnesty International video call with Jakub Szymik, 5 August 2024.
[440] Amnesty International video call with Jakub Szymik, 5 August 2024.
[441] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited).
[442] Amnesty International interview with Magda Dropek, 24 July 2024.
[443] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).

moderation to X users.[444] As discussed in section 5.4.1 the Community Notes feature is seriously limited and flawed.

7.8 X'S KNOWLEDGE OF SYSTEMIC RISKS

Based on its latest DSA risk assessment, X is clearly aware that its platform represents systemic risks to a range of human rights, including risks at the "societal level" and specifically to marginalized communities.[445] X highlights that its "approach to assessing and mitigating risks associated with harmful content continues to be based on a framework that considers physical, psychological, informational, economic and societal harms, allowing us to analyse the potential real-world harm of content and behaviour that may occur on X".[446]

While algorithmic amplification and recommendation are key parts of X's business model, the platform maintains that its algorithms do not intentionally promote content containing "slurs" and "hateful terms".[447] Nonetheless, the company acknowledges that previous research has shown that "in certain circumstances our recommender systems could lead to accounts from specific ideological leanings to be amplified over others. However, while there was a risk of bias in these systems, the research highlighted that there are no clear, singular factors in this effect and that in different circumstances the same algorithm produced different impacts on political content."[448] This underlines the imperative for X to perform country-specific human rights due diligence on the potential harmful impacts of its recommender systems, if they indeed function differently in different contexts.

X is also aware that some of its design features, such as mentions and quote posts, may be leveraged for harassment, "contributing to a risk to human dignity, non-discrimination, and the respect for private and family life".[449] X further accepts that "the digital gender divide may have also contributed to women and members of the LGBTQ+ community being a target of hate and abuse".[450]

7.9 ASSESSING X'S CONTRIBUTION TO TFGBV AGAINST POLAND'S LGBTI COMMUNITY

According to the UN Guiding Principles, a business enterprise has contributed to an adverse human rights impact when its activities (including omissions) materially increase the risk of the specific impact which occurred – even if the business enterprise's activities would not have been sufficient in and of themselves to result in that impact.[451] To fulfil its responsibility to respect human rights, X must "avoid causing or contributing to adverse human rights impacts through their own activities" and to "seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts".[452]

Between 2019 and 2023, X was used by a range of actors including Polish government officials, regional government officials and anti-LGBTI activists to post content which targeted the LGBTI community. Some of this content incited violence and discrimination. While the political rhetoric around the LGBTI community has improved since the 2023 election, the effect of years of hate lingers on the platform, with LGBTI people continuing to be targeted with TfGBV.

---

[444] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).
[445] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited).
[446] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024 (previously cited), p. 7.
[447] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited).
[448] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited), p. 55.
[449] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited), p. 22.
[450] X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, September 2023 (previously cited), p. 65.
[451] Business & Human Rights Resource Centre, *Practical Definitions of Cause, Contribute, and Directly Linked to Inform Business Respect for Human Rights,* 9 February 2017, https://www.business-humanrights.org/en/latest-news/practical-definitions-of-cause-contribute-and-directly-linked-to-inform-business-respect-for-human-rights/
[452] UN Guiding Principles, Principle 13 including Commentary.

X's contribution to the negative human rights impacts suffered by the LGBTI community stems from the fact that X's mitigation measures – such as content moderation – have not adequately addressed the prevalence of TfGBV including threats of violence, online harassment and doxing on the platform.

The effects of this were made more acute because X is an important platform in Poland, particularly for political discourse, and a source of information for journalists and activists. X can also be considered to have contributed to adverse human rights impacts due to the foreseeability of the risk its operations presented in X. Despite well-documented attacks on the LGBTI community from senior political figures in Poland, X failed to adequately mitigate the human rights risks of its operations in Poland.

There are numerous additional steps that X could have taken to prevent the spread and prevalence of content targeting the LGBTI community on the platform, such as more proactively engaging with content moderation mechanisms. Amnesty International sent a letter to X in August 2024 asking for information on X's staffing and resources for its Poland operations between 2019 and 2024, including the number of country-specific content moderators, their proficiency in Polish, and their physical location, but the company did not provide a response.[453] As detailed in this report, X was not able to adequately moderate content in Poland. Additionally, the platform was slow to respond to feedback from platform users and a civil society organization monitoring hate speech online, reporting content which should be considered TfGBV - and in some cases, failed to respond at all. This resulted in harmful content being allowed to circulate on the platform and some members of the LGBTI community no longer reporting TfGBV due to the lack of a response from X.

Despite its obligation to identify systemic risks under the DSA, there is little evidence that X has made meaningful efforts to adequately identify or mitigate the risks its platform presents to the LGBTI community in Poland.

Amnesty International's analysis of X's role in human rights abuses suffered by the LGBTI community in Poland from 2019 to the present day, based on international human rights standards including the UN Guiding Principles, leads to the following conclusions:

1. As a key platform in Poland for politicians, journalists and activist communities, members of the Polish government, Polish political parties and anti-LGBTI activists have used X to post content targeting the LGBTI community. Some of this content has incited violence and discrimination.
2. X's failures of content moderation in Poland allowed content which incited violence and discrimination against the LGBTI community to remain prevalent on the platform.
3. X knew, or should have known, that it risked contributing to human rights abuses in Poland, particularly as its Polish content moderation efforts are not as well-resourced as those in other European countries.
4. X failed to engage in adequate human rights due diligence, which could or should have identified the risks that its operations presented in Poland. X also failed to enact adequate and appropriate mitigation measures which may have prevented or mitigated the harm in Poland.
5. In the case studies outlined in Chapter 6, X's failures of due diligence regarding the prevalence of content inciting violence, discrimination and hate in Poland and its inadequate content-moderation operations, contributed to violations of a range of human rights, including the right to freedom of expression, the right to equality and non-discrimination, and the right to health.

X contributed to TfGBV suffered by the Polish LGBTI community and therefore has a corresponding responsibility to remediate the harm.

[453] Amnesty International letter to X, 22 August 2024.