‘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.

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

CONTENTS

ABBREVIATIONS

WORD	DESCRIPTION
AI	artificial intelligence
ATI	Algorithmic Transparency Institute
CEDAW COMMITTEE	UN Committee on the Elimination of Discrimination against Women
DSA	EU Digital Services Act
DSC	Digital Services Coordinator
DUE DILIGENCE GUIDANCE	OECD Due Diligence Guidance for Responsible Business Conduct
ECtHR	European Court of Human Rights
EU	European Union
FTC	US Federal Trade Commission
GBV	gender-based violence
GNI	Global Network Initiative
ICCPR	International Covenant on Civil and Political Rights
ISD	Institute for Strategic Dialogue
LLM	Large-Language Model
OECD GUIDELINES	OECD Guidelines for Multinational Enterprises on Responsible Conduct
OHCHR	UN Office of the High Commissioner for Human Rights
SOGIESC	sexual orientation, gender identity and/or expression and sex characteristics
TFGBV	technology-facilitated gender-based violence
UN GUIDING PRINCIPLES	UN Guiding Principles on Business and Human Rights
VLOP	Very Large Online Platform

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.

2. METHODOLOGY

This report is based on research conducted by Amnesty International between March 2024 and April 2025 using a combination of participatory research design approaches, quantitative methods (the research account experiment – see below), and interviews. Amnesty International conducted an analysis of the human rights implications of X's business operations in Poland using publicly available information on X's website and through DSA transparency reports. This analysis was informed by Amnesty International's interviews with subject matter experts, quantitative research of the platform, and desk research. The organization also carried out extensive desk research from open sources, reports from civil society organizations and national and international news media.

This report builds on previous extensive investigations by Amnesty International that found Poland's treatment of the LGBTI community to be discriminatory, and that hostile and stigmatizing rhetoric against LGBTI people, including by high-level officials, have had increasingly harmful consequences that are translating into violence and discrimination on the basis of people's actual or perceived sexual orientation, gender identity and/or expression.[2]

Amnesty International conducted interviews, either in-person or remotely, with 11 affected individuals. Case studies based on these interviews feature in this report and are illustrative and emblematic of the harms to which X has contributed in Poland. Amnesty International conducted five further interviews with subject-matter experts, including digital rights experts and experts on LGBTI rights in Poland. The organization also conducted an in-person participatory workshop with 16 participants from the LGBTI community in Poland.

All interviews were conducted primarily in English, with intermittent Polish translation as required. The information gathered in these interviews was then corroborated with local digital rights and LGBTI organizations, and through Amnesty International's quantitative research on X. All interviewees gave informed consent in advance of being interviewed. Amnesty International did not provide any incentives in exchange for interviews. Due to security risks, some of those interviewed requested anonymity, while others wished to share their identities publicly. For those who chose anonymity, Amnesty International has used pseudonyms and omitted all potentially identifying information from this report.

2.1 QUANTITATIVE EXPERIMENT

Testimonial and participatory approaches are imperative to understanding and foregrounding people's lived experience, as well as serving as a primary source of evidence and a tool for advocacy and awareness-raising. Yet the opaque nature of X's platform mechanics meant that, in order to answer questions on the scale (if any) and prevalence of amplification of certain types of content on the platform, employing quantitative research methods was crucial to this project. Therefore, to complement the participatory and testimonial research, the study also included a quantitative component, conducted in partnership with the ATI.

In an ideal scenario, to explore the way in which X's algorithmic recommender systems serve content to platform users and the prevalence of anti-LGBTI content on the platform, researchers would collect data in a real-world setting, directly from users of the platform. Amnesty International explored this idea, but

[2] Amnesty International, *Targeted by Hate, Forgotten by Law: Lack of a Coherent Response to Hate Crimes in Poland* (Index: EUR 37/2147/2015), 17 September 2015, https://www.amnesty.org/en/documents/eur37/2147/2015/en/; Amnesty International, *"They Treated Us Like Criminals": From Shrinking Space to Harassment of LGBTI Activists* (Index: EUR 37/5882/2022), 20 July 2022, https://www.amnesty.org/en/documents/eur37/5882/2022/en/

concluded that it may not be feasible to collect as much data as necessary within time constraints using this method. The next-best alternative was to create "sock puppets", or simulation research accounts, that mimic users' experience on X.

The quantitative experiment included: (1) a study of the prevalence of anti-LGBTI content on X in Poland, scraping the platform for examples of such content, and (2) a study of the amplification of any harmful content by X's recommender system, using research accounts to simulate users who may be interested in anti-LGBTI content. This was conducted between 1 March and 31 March 2025.

For the experiment, researchers set up 32 research accounts with four different pre-defined "personas" to mimic different users' experiences on X. Personas were set up to follow politicians from specific political parties, with the four groups strategically chosen to represent differing stances on LGBTI issues (alongside other social stances more generally). The 32 research accounts were sorted into one of these four groups, each following 10 politicians from the associated political parties:

- Restrictive on the rights of LGBTI people (PiS, Konfederacja, Kukiz)
- Split on the rights of LGBTI people (PSL). Supports civil partnerships, opposes same-sex marriage (PL2050, KO)
- Supports full rights (Lewica Razem)

In order to limit the extent to which the study promoted potentially harmful content for other users, the research accounts did not like, repost, comment, message, or search for any specific terms.

Each research account was set up to log on to X once per day for a total period of 28 days. Halfway through the study (day 15), each research account was randomly assigned to one of two groups, (1) following the recommended accounts on the "Who to follow" recommendation list, or (2) not following them. The daily workflow for each research account followed this process:

1. Sign into X.
2. Collect all recommended accounts on the "Who to follow" list.
3. Scroll through the first 200 tweets on the reverse-chronological "Following" timeline.
4. Scroll through the first 200 tweets on the algorithmically recommended "For You" feed.
5. Sign out of X.

Finally, the research accounts collected comments and replies from 55 LGBTI-related posts with the highest engagement metrics.

2.2 CONTENT LABELLING

Amnesty International's investigation was concerned with (1) any cumulative effect of X's amplification of anti-LGBTI content, and (2) individual failures of content moderation. Any study of prevalence and amplification of content on platforms presents a challenge by requiring researchers to categorize individual pieces of content.

In collaboration with Amnesty International Poland, researchers developed a framework to categorize individual tweets and comments posted by users. While the framework used in this report reflects feedback from various experts, it can nonetheless only ever present an approximation of the content observed and its potential for harming a person. The categorization scheme that was used sought to provide a transparent description of the dimensions of potentially harmful LGBTI-related content that was present on X.

A total sample of 1,387 individual tweets were manually labelled by researchers according to a framework below:

1. Abusive
2. Homophobic
3. Transphobic
4. Non-consensual intimate image
5. Doxing

6. Harassment
7. Other problem
8. Anti-immigrant
9. Antisemitic
10. Casually racist
11. Anti-LGBTI general
12. Pro-LGBTI general
13. LGBTI other
14. Not relevant

Researchers then deployed a large language model (LLM) to label replies and comments to the 55 individual LGBTI-related posts with the highest engagement metrics. The LLM was trained using examples of the manually labelled content to then classify each reply and comment as to whether it was homophobic/transphobic or not.[3]

In total, across all 32 research accounts over the study period, 163,048 tweets were collected. Of these: 141,563 were not advertisements. As many of the research accounts will have encountered the same content, the number of unique tweets collected is lower, at 33,879, of which 31,951 were not advertisements.

In total, 10,455 non-advertisement unique tweets were collected from the "For You" feed. Content posted from Elon Musk's personal account substantially outweighed content from other accounts which the research accounts did not follow on their respective "For You" feeds. Elon Musk's tweets were seen by the research accounts 1,411 times during the course of the study. The closest other recommended, non-followed account most seen by the research accounts, that of Polish politician Roman Giertych, was seen 494 times.

2.3 METHODOLOGICAL CHALLENGES

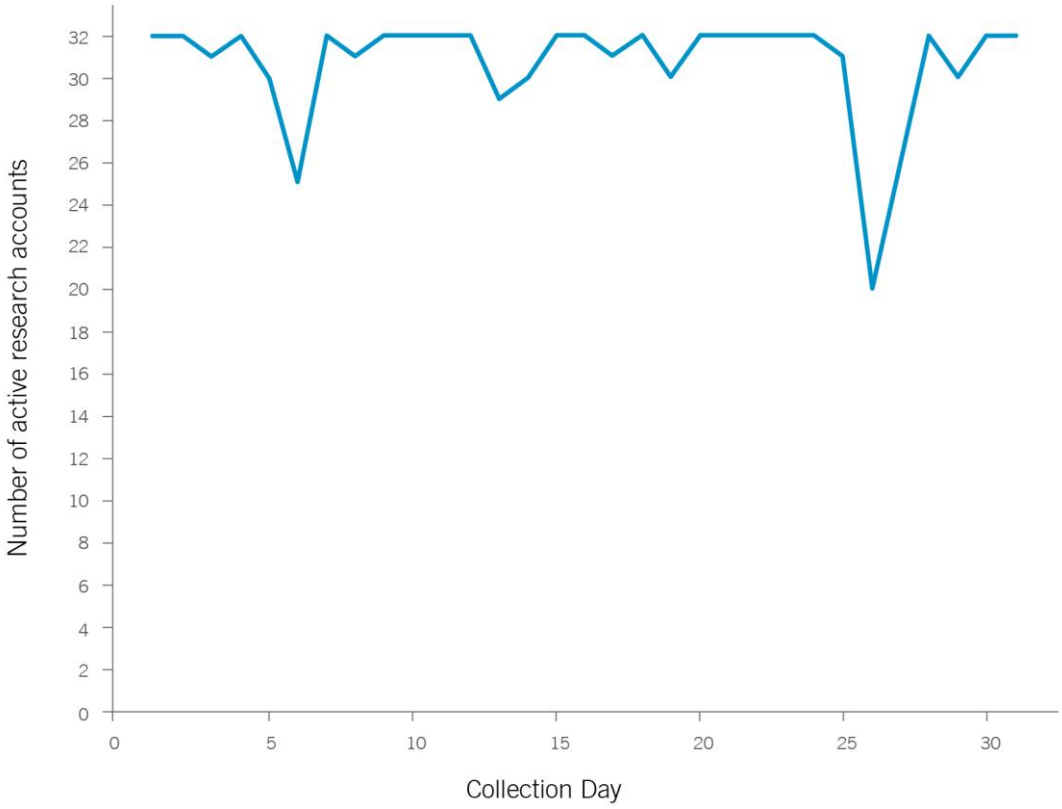
Although the study involved a larger sample compared to previous civil society research in this field, the sample size of 32 research accounts was still not large enough for any statistical hypothesis testing. Therefore, the analysis descriptively compared the volume and frequency of LGBTI-related content, and in particular harmful LGBTI content, seen by each sub-group. The research aimed to build on prior studies by including comparator groups in the form of sub-groups that follow politicians with differing stances on the rights of LGBTI people.

It is highly likely that factors such as the date, time of day and location are used within the recommender algorithm and therefore are influential in determining the content that X serves its users. To ensure the comparisons between sub-groups were as robust as possible, researchers attempted to control these by setting up research accounts to access the platform on specific dates and times from the same location in Poland. Technical issues such as the research account being shut down or malfunctioning presented challenges to this; however, researchers ensured to the greatest extent possible that all data was collected in parallel across each sub-group.

Figure 1 below shows the timeline of how many research accounts were active each day. Note that, on days 6 and 26, a number of technical issues were encountered. To ensure the reliability of the study, the researchers checked and confirmed that the offline research accounts were evenly split between the four sub-groups.

[3] ChatGPT 4.1mini was deployed on 15 April 2025. The LLM was provided the following prompt: *"You are a binary classifier that decides whether particular responses to tweets are homophobic/transphobic, or not. You will be given the original tweet for context and a reply to classify. Only classify the reply, not the tweet. Is the response homophobic/transphobic? Answer only with one number. 1 if the response is homophobic/transphobic and 0 if it is not"*. Each user prompt had the content of the original tweet and the content of the reply. Both were stripped of user handles and links, and the queries did not contain any meta data. Amnesty International notes that LLM categorization is not always stable day to day and therefore the results of this content labelling exercise are specific to the date upon which it was conducted. However, the labels provided by the LLM model are used sparingly throughout the analyses, where the findings do rely upon this, Amnesty International has noted the caveats.

FIGURE 1: NUMBER OF ACTIVE RESEARCH ACCOUNTS PER DAY FOR THE DURATION OF THE DATA COLLECTION PERIOD



Amnesty International wrote to X on 22 August 2024, posing questions regarding the company's actions in relation to its business activities in Poland between 2019 and 2024 but did not receive a response. Amnesty International again wrote to X on 25 June 2025 to inform the company of relevant allegations contained in this report and to give the company an opportunity to respond but did not receive a response.

Throughout this report, "X" is used to refer to the company X.Corp (formerly Twitter), in relation to the period before the company rebranded in April 2023. The term "Twitter" is frequently used interchangeably with "X" by interviewees when talking about both the platform and the company.

3. BACKGROUND

3.1 TARGETED: THE LGBTI COMMUNITY IN POLAND

LGBTI people in Poland have struggled for decades with systemic discrimination by successive Polish governments, which have often implemented restrictive policies under the guise of "traditional values". Between 1985 and 1987, Polish authorities engaged in Operation Hyacinth, a secret mass operation which resulted in the detention of 11,000 people "suspected of or in contact with homosexuality".[4] Compromising details found during this operation were stored by the Secret Police as so-called "Pink Files".[5] The authorities claimed this operation was a preventative measure to counteract sex work and "homosexual criminal gangs".[6] The rights of LGBTI people have remained in the political spotlight since Poland's democratization in the 1990s.[7]

Under the previous PiS government, which held power from 2015 to 2023, Polish authorities took actions which shrank the space for civil society, including undermining the rule of law and attacking the rights of women and LGBTI people, creating an inhospitable environment for LGBTI people and their allies.[8] Hostile and stigmatizing rhetoric against the LGBTI community, including by high-level officials, at times translated into violence and discrimination on the basis of people's actual or perceived sexual orientation, gender identity and/or expression.[9] Politicians from PiS in particular issued numerous anti-LGBTI statements while in government, nurturing an atmosphere that fostered discrimination and afforded social licence for hostility towards LGBTI people in wider society.[10]

An expert in advancing the rights of transgender and non-binary people based in Poland told Amnesty International that the rhetoric against the LGBTI community became noticeably more hostile after PiS ascended to power:

"After PiS started ruling the country, things that were [previously] unacceptable… People started wearing it as a crown. It became great to be hateful."[11]

Many groups promoting anti-rights narratives in particular enjoyed growing impunity when they condoned, advocated for, or used violence or discrimination against LGBTI individuals.[12] Amnesty International has previously found a direct link between the erosion of the rights to freedom of expression, association and peaceful assembly, and the harassment, profiling and targeting of LGBTI activists in Poland.[13]

People in positions of power in the PiS government between 2015 and 2023 and other influential public figures intentionally portrayed LGBTI people as a "threat" to "family values", "the Catholic faith" and "public order".[14] The first high-profile example of a political figure engaging with anti-LGBTI rhetoric in Poland came on 17 April 2018. Speaking as part of a local government campaign, the leader of PiS, Jaroslaw Kaczyński,

[4] Amnesty International, *"They Treated Us Like Criminals"* (previously cited).
[5] *Amnesty International, "They Treated Us Like Criminals"* (previously cited).
[6] Amnesty International, *"They Treated Us Like Criminals"* (previously cited).
[7] The Guardian, "In Poland, the home of 'LGBT-free zones', there is hope at last for the queer community", 1 November 2023, https://www.theguardian.com/commentisfree/2023/nov/01/poland-lgbtq-new-government-law-and-justice-equality
[8] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[9] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[10] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[11] Amnesty International interview with Julia Kata, 29 July 2024.
[12] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[13] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[14] ILGA Europe, "Poland: Anti-LGBTI hate timeline", 2021, https://www.ilga-europe.org/report/poland-anti-lgbti-hate-timeline/; Amnesty International, "*They Treated Us Like Criminals"* (previously cited).

stated that "no homosexual marriages will occur; we will wait peacefully for the European Union countries to sober up".[15]

In a June 2020 speech in the town of Brzeg, the then President, Andrzej Duda, claimed that LGBTI people were an "ideology", stating: "They're trying to tell us that they're people. And it's an ideology. If anyone has any doubts as to whether or not this is an ideology, look back through history and see what it was like to build the LGBT movement around the world... an ideology that is even more destructive to human beings, an ideology that beneath the platitudes of respect and tolerance hides a deep intolerance and elimination, the exclusion of all those who do not want to submit to it."[16]

In the same year, responding to the European Commission's refusal to include self-declared "LGBTI-free zone" Polish towns in the EU's Town Twinning scheme, Poland's Ministry of Justice announced it would financially compensate those towns under its Justice Fund, a fund designed to aid victims of crime.[17] The Ministry reasoned that the towns were "victims" of a lack of EU funding.[18]

In March 2021, Zbigniew Ziobro, serving simultaneously as Minister of Justice and Prosecutor General, proposed a bill to the Sejm (the lower house of Poland's parliament), which it adopted the same year, making adoption by same-sex couples illegal in Poland.[19]

In 2022, Amnesty International found compelling evidence of how attacks on LGBTI people at peaceful gatherings had markedly increased, especially in the wake of the government's hate campaign against LGBTI people, mainly led by politicians from PiS and Konfederacja, which particularly intensified during the 2019 presidential campaign.[20]

As the stigmatization of LGBTI people in Poland deepened, peaceful assemblies such as Equality Marches were repeatedly met with hostility and violence from central and local authorities and law enforcement officials.[21] A prominent example of this is the Bialystok Equality March, which took place on 20 July 2019. Enabled by a lack of police protection for the march, people attending were attacked by a much larger and aggressive crowd of 4,000 counter-demonstrators brandishing bottles, paving stones and firecrackers, and were subjected to homophobic slurs.[22]

A few months later, on 28 September 2019 at the Equality March in Lublin, police deployed water cannons and pepper spray against counter-protesters, making dozens of arrests.[23] It was later revealed that two of the counter-protesters had brought home-made explosives to the march.[24] Amnesty International has also previously documented the use of profiling by police of LGBTI assemblies in Poland to target individuals for detention on the basis of their sexual orientation or gender identity.[25]

The attacks on the rights of LGBTI people in general, and on Equality Marches in particular, were supported by some Polish media outlets, as well as political figures. On 10 October 2019, a documentary aired on public television entitled *LGBT Invasion*, claiming that LGBTI people were paid by foreign NGOs to participate in Equality Marches.[26] The documentary asked questions such as: "Who and for what purpose finances the LGBT marches?", "What methods and sources of funding do they have?" and "How does the LGBT invasion work?"[27]

Julia Kata, a psychologist at the Polish LGBTI civil society organization Fundacja Trans-Fuzja, explained to Amnesty International how the prevalence of anti-LGBTI content in the traditional media provided social licence for open hostility towards LGBTI people:

"If you see these kinds of comments on TV, what's coming to your mind? It's okay to do these kinds of things, it's okay to say these kinds of things. Everybody does that on TV and if it was something wrong and disgusting, they won't show it, right? But they are showing this. So, it's okay."[28]

---

[15] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).
[16] Euronews, "LGBT campaigners denounced President Duda's comments on 'communism'", 15 June 2020, https://www.euronews.com/2020/06/15/polish-president-says-lgbt-ideology-is-worse-than-communism
[17] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).
[18] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).
[19] Equaldex, "LGBT rights in Poland", https://www.equaldex.com/region/poland#adoption (accessed on 2 July 2025).
[20] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[21] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[22] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[23] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).
[24] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).
[25] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[26] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).
[27] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).
[28] Amnesty International interview with Julia Kata, 29 July 2024.

In 2022 the Warsaw District Court ordered the television station to pay a fine and apologize for slandering the LGBTI community in the documentary, affirming such anti-LGBTI programming to be incompatible with media ethics.[29]

As part of his presidential election campaign in June 2020, Andrzej Duda publicly signed the Family Charter which, among other things, committed him to "defend the institution of marriage", prevent the adoption of children by same-sex couples and "protect" children and families from "LGBT ideology", which he described as a foreign ideology "worse than communism" and which he vowed to ban in public institutions.[30] In the same election campaign, Andrzej Duda proposed incorporating a ban on same-sex marriage into Poland's constitution.[31]

Andrzej Duda posited that the LGBTI community was trying to "force" LGBTI rights on Poland.[32] This sentiment was echoed by PiS leader Jaroslaw Kaczyński in an April 2021 interview in which he referred to "LGBT ideology", saying that it "radically limits the freedom of a great number of people who are terrorized to accept this ideology".[33]

In August 2021, a civil bill known as "Stop LGBT" was submitted to the Sejm and subsequently sent for further work after the first reading.[34] The bill proposed imposing a total ban on the "promotion of LGBTI ideology" in public spaces, thus posing a serious threat to the rights to freedom of expression and peaceful assembly, in contravention of international human rights law and standards.[35] The bill was struck down by Poland's ombudsperson in 2024, who said that the attempt to limit the rights of the LGBTI community should be considered unconstitutional.[36]

3.2 "LGBT-FREE ZONES"

In 2019, hostile attitudes towards LGBTI people reached a peak when regions and municipalities joined a government-supported Family Charter, calling for the exclusion of LGBTI people from Polish society.[37] While the local charters were legally unenforceable, they were clear attempts to stigmatize, exclude and discriminate against LGBTI people, sending a clear message that LGBTI individuals were not welcome in those areas. In 2019, around 100 local municipalities in Poland stated that their constituency was "LGBT-free" or banned "LGBT ideology".[38] By early 2020, roughly one-third of the country was covered by "LGBT-free zones".[39]

A 2021 report by the Parliamentary Assembly of the Council of Europe on combating rising hate against LGBTI people in Europe concluded that the zones "deny LGBTI people's right to exist and deprive them of a safe space".[40] By December 2022, more than 90 regional and municipal authorities had declared themselves "LGBT-ideology free" or signed the government-supported Family Charter.[41] The zones were actively supported by members of the Polish government, with activists reporting that the Ministry of Education and Science had sent letters of support to cities or regions adopting resolutions against "LGBT ideology."[42] A weekly newspaper, Gazeta Polska, distributed free stickers bearing the text "LGBT-free zone" to support the adoption of resolutions.[43] LGBT-free zones were also roundly condemned by the EU. In 2021, the European Commission launched infringement proceedings against Poland related to the equality and

[29] ILGA-Europe, "Annual review of the human rights of lesbian, gay, bisexual, trans, and intersex people in Poland covering the period of January to December 2022", 2022, https://www.ilga-europe.org/sites/default/files/2023/poland.pdf
[30] BBC News, "Poland LGBT protests: Three charged with hanging rainbow flags off statues", 5 August 2020, https://www.bbc.co.uk/news/world-europe-53673411; ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).
[31] BBC News, "Poland LGBT protests: Three charged with hanging rainbow flags off statues" (previously cited).
[32] BBC News, "Poland LGBT protests: Three charged with hanging rainbow flags off statues" (previously cited).
[33] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).
[34] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[35] Amnesty International, "*They Treated Us Like Criminals"* (previously cited); Amnesty International, "Pride 'under attack' from new bill in parliament", 28 October 2021, https://www.amnesty.org/en/latest/news/2021/10/poland-pride-under-attack-from-a-new-bill-in-parliament/
[36] Polska Agencja Prasowa, "Ombudsperson strikes down anti-LGBT bill", 6 February 2024, https://www.pap.pl/en/news/ombudsperson-strikes-down-anti-lgbt-bill
[37] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited); Human Rights Watch, "Poland: rule of law erosion harms women, LGBT people", 15 December 2022, https://www.hrw.org/news/2022/12/15/poland-rule-law-erosion-harms-women-lgbt-people
[38] PinkNews, "Poland abolishes last remaining 'LGBT-free' zone in the country", 29 April 2025, https://www.thepinknews.com/2025/04/29/poland-abolishes-lgbt-free-zones/
[39] LGBTQ Nation, "Poland finally repealed the country's last 'LGBT-free zone'", 28 April 2025, https://www.lgbtqnation.com/2025/04/poland-finally-repealed-the-countrys-last-lgbt-free-zone/
[40] Human Rights Watch, "Poland: rule of law erosion harms women, LGBT people" (previously cited).
[41] Human Rights Watch, "Poland: rule of law erosion harms women, LGBT people" (previously cited).
[42] Human Rights Watch, "Poland: rule of law erosion harms women, LGBT people" (previously cited).
[43] ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited).

protection of fundamental rights.[44] The Commission considered that Polish authorities had failed to fully and appropriately respond to its inquiry regarding the nature and impact of the zones. In September 2021 the European Commission put on hold funds to five regions unless they abandoned anti-LGBTI declarations, which resulted in four regions rescinding them.[45]

In 2022, Poland's Supreme Administrative Court deemed the LGBT-free zones unconstitutional and ruled that the effect of the resolutions was a "violation of the dignity, honour, good name and closely related private life of a specific group of residents".[46] As a result of the ruling, most of the local LGBT-free resolutions were repealed.[47]

On 24 April 2025 the last remaining LGBT-free zone was abolished, after officials in Lancut, a town in the south-east of Poland, voted to repeal the regulation.[48]

3.3 LACK OF LEGAL INCLUSION

Polish criminal law specifically provides for the investigation and prosecution of hate crimes motivated by race, ethnicity, nationality, religious affiliation or irreligiousness. However, it does not establish that motivations based on age, disability, gender, gender identity, sexual orientation and social or economic status are also grounds to investigate and prosecute hate crimes.[49] Under international and European human rights law, age, disability, gender, gender identity, sexual orientation and social or economic status are protected characteristics. The absence of hate crime legislation on the basis of sexual orientation and gender identity has meant that homophobic or transphobic motives are rarely considered and these motivations do not play a role in the prosecution of hate crimes in Poland.[50]

This lack of legal inclusion means that there are no institutional mechanisms for dealing with homophobic and transphobic crimes and that, additionally, authorities do not systematically collect official data on homophobic and transphobic crimes, meaning that the extent of hate crimes against the LGBTI community in Poland is unclear.[51] Additionally, LGBTI people who experience violence and other hate crimes are impeded from safely and adequately reporting these acts, further exacerbating the lack of official information on this issue.[52] The combination of these factors means that, despite a documented rise in anti-LGBTI sentiment and international pressure to address this, Poland still does not have a coherent system for reporting and combating hate crimes and hate speech based on sexual orientation and gender identity.[53] As a result, Poland's LGBTI community was ranked the least protected in the EU from 2020 to 2024.[54]

LGBTI people in Poland continue to experience violence and discrimination on the basis of their sexual orientation, gender identity and/or expression, with trans and intersex people often the most harshly affected.[55] This means that many LGBTI people in Poland live in a constant state of vulnerability. In a 2023 survey conducted by FRA, the European Union Agency for Fundamental Rights, 74% of respondents in Poland reported that they often or always avoid holding hands with same-sex partners, and 61% of respondents considered that violence had increased a little or a lot since 2019.[56]

In 2024 the new Polish government took steps to add sexual orientation, gender, age and disability to the categories covered by Poland's hate crime laws.[57] The draft bill was approved by the Council of Ministers and submitted to the Sejm in November 2024, with the first reading of the bill taking place on 19 December

[44] European Commission, "EU founding values: Commission starts legal action against Hungary and Poland for violations of fundamental rights of LGBTIQ people", 15 July 2021, https://ec.europa.eu/commission/presscorner/detail/en/ip_21_3668; ILGA-Europe, "Poland: Anti-LGBTI hate timeline" (previously cited); Human Rights Watch, "Poland: rule of law erosion harms women, LGBT people" (previously cited).
[45] Human Rights Watch, "Poland: rule of law erosion harms women, LGBT people" (previously cited).
[46] TVP World, "Poland's last 'LGBT-free zone' officially abolished", 25 April 2025, https://tvpworld.com/86360798/poland-abolishes-last-lgbt-free-zone-; LGBTQ Nation, "Poland finally repealed the country's last 'LGBT-free zone'" (previously cited).
[47] TVP World, "Poland's last 'LGBT-free zone' officially abolished" (previously cited).
[48] PinkNews, "Poland abolishes last remaining 'LGBT-free' zone in the country" (previously cited).
[49] Amnesty International, *Targeted by Hate, Forgotten by Law* (previously cited).
[50] Amnesty International, *Targeted by Hate, Forgotten by Law* (previously cited).
[51] Amnesty International, *Targeted by Hate, Forgotten by Law* (previously cited).
[52] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[53] Amnesty International, "*They Treated Us Like Criminals"* (previously cited).
[54] The Guardian, "In Poland, the home of 'LGBT-free zones', there is hope at last for the queer community" (previously cited); ILGA-Europe, "Rainbow map 2024", 15 April 2024, https://www.ilga-europe.org/report/rainbow-map-2024/
[55] FRA, "LGBTIQ equality at a crossroads: progress and challenges: EU LGBTIQ survey III", 2024, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2024-lgbtiq-equality_en.pdf
[56] FRA, "LGBTIQ equality at a crossroads: progress and challenges: EU LGBTIQ survey III" (previously cited).
[57] Notes from Poland, "Polish government approves criminalisation of anti-LGBTI hate speech", 28 November 2024, https://notesfrompoland.com/2024/11/28/polish-government-approves-criminalisation-of-anti-lgbt-hate-speech/; Brussels Signal, "Polish government approves criminalisation of anti-LGBT hate speech", 3 December 2024, https://brusselssignal.eu/2024/12/polish-government-approves-criminalisation-of-anti-lgbt-hate-speech/

2024.[58] The Ministry of Justice said that the new regulations "aim to more fully implement the constitutional prohibition of discrimination and to meet international recommendations on standards of protection against hate speech and hate crimes."[59]

However, even this positive development has contentious elements. The initial version of the proposed legislation included "gender identity" as a newly protected category.[60] However, the Ministry of Justice eventually decided that the term "sex/gender" was "sufficient to ensure an appropriate level of protection."[61] This decision was criticized by LGBTI organizations in Poland, with Lambda branding it "disturbing" and Fundacja Trans-Fuzja warning that the change could result in "one of the most excluded and vulnerable groups remaining unprotected" – referring to transgender individuals.[62] There are also concerns from civil society that the absence of a definition of a hate crime is a significant flaw in the draft.[63]

If passed by parliament, the President can sign the bill into law, veto it or pass it to the constitutional court for assessment.[64]

3.4 A CHANGING POLITICAL LANDSCAPE

In October 2023 a new coalition government was elected in Poland comprising the Civic Coalition, the Third Way and the Left, and led by the new Prime Minister, Donald Tusk. Some members of the LGBTI community expressed a sense of hope following the change in administration.[65] The new cabinet included Poland's first ever Minister for Equality.[66] Additionally, polls have consistently shown that the Polish public's attitude to LGBTI rights is becoming increasingly progressive. In 2022, Ipsos polls showed that two-thirds of the population supported marriage equality or civil partnerships and 60% believed that LGBT-free zones should be abolished in order to meet requirements to receive EU funding.[67]

The new government has represented a turning point of sorts for the LGBTI community in Poland, promising to revise some of the policies affecting the community and take steps to address significant protection gaps and bring domestic laws and policies more in line with international norms and standards.[68] In an important gesture, in December 2023 the Ministry of Justice issued a public apology to LGBTI people for the way in which they had been previously maligned by state actors and public media.[69]

However, concerns remain about the length of time that significant legislative changes will take to come into effect.[70] Additionally, the most progressive changes may be frustrated by the fact that there is a lack of unity regarding the rights of the LGBTI community within the ruling coalition.[71]

Nevertheless, the government has taken steps to fulfil some of its promises. In October 2024, it moved a step closer to legalizing civil partnerships with the publication of a draft law.[72] Under the bill, couples in a civil partnership would gain rights to inheritance and medical information about their partners – but not the

[58] KPH, "Amendment to the criminal code on hate crimes and hate speech is about ensuring everyone's safety – NGOs claim", 19 December 2024, https://kph.org.pl/en/amendment-to-the-criminal-code-on-hate-crimes-and-hate-speech-is-about-ensuring-everyones-safety-ngos-claim/
[59] Notes from Poland, "Polish government approves criminalisation of anti-LGBT hate speech" (previously cited).
[60] Notes from Poland, "Polish government approves criminalisation of anti-LGBT hate speech" (previously cited).
[61] Notes from Poland, "Polish government approves criminalisation of anti-LGBT hate speech" (previously cited).
[62] Notes from Poland, "Polish government approves criminalisation of anti-LGBT hate speech" (previously cited).
[63] KPH, "Amendment to the criminal code on hate crimes and hate speech is about ensuring everyone's safety – NGOs claim" (previously cited).
[64] Notes from Poland, "Polish government approves criminalisation of anti-LGBT hate speech" (previously cited).
[65] The Guardian, "In Poland, the home of 'LGBT-free zones', there is hope at last for the queer community" (previously cited).
[66] Reuters, "Rights court rules Poland should recognise same-sex partnerships", 12 December 2023, https://www.reuters.com/world/europe/rights-court-rules-poland-should-recognise-same-sex-partnerships-2023-12-12/
[67] ILGA-Europe, "Annual review of the human rights of lesbian, gay, bisexual, trans and intersex people in Poland covering the period of January to December 2022" (previously cited); The Guardian, "In Poland, the home of 'LGBT-free zones', there is hope at last for the queer community" (previously cited); United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, "Country visit to Poland (18-29 November 2024): End-of-mission statement", 29 November 2024, https://www.ohchr.org/sites/default/files/documents/issues/sexualorientation/statements/2024-11-29-preliminary-observations-ie-sogi-visit-poland.pdf
[68] Euronews, "Queer in Poland: when can the LGBTQ+ community expect equal rights?", 5 April 2024, https://www.euronews.com/2024/04/05/queer-in-poland-when-can-the-lgbtq-community-expect-equal-rights; United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, "Country visit to Poland (18-29 November 2024): End of mission statement" (previously cited).
[69] United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, "Country visit to Poland (18-29 November 2024): End of mission statement" (previously cited).
[70] Euronews, "Queer in Poland: when can the LGBTQ+ community expect equal rights?" (previously cited); Reuters, "Poland publishes civil partnership bill in boost for LGBT couples", 18 October 2024, https://www.reuters.com/world/europe/poland-publishes-civil-partnership-bill-boost-lgbt-couples-2024-10-18/
[71] Euronews, "Queer in Poland: when can the LGBTQ+ community expect equal rights?" (previously cited).
[72] Reuters, "Poland publishes civil partnership bill in boost for LGBT couples" (previously cited).

right to adopt children, a concession thought to be designed to secure the support of the conservative Polish People' Party (Polskie Stronnictwo Ludowe, PSL), which is part of Prime Minister Donald Tusk's ruling coalition.[73] The government came under renewed pressure to advance the legislation in April 2025, following a ruling by the European Court of Human Rights (ECtHR) that Poland must provide legal recognition and protection for same-sex unions to meet the country's obligation to ensure equal rights for all citizens.[74] While issuing the verdict, the ECtHR noted that the absence of legal recognition of same-sex couples who married abroad was a denial of individuals' rights and added that Poland was obliged to legislate for those rights to be respected.[75]

Significant change has already occurred in Poland. In March 2025 the country's Supreme Court issued a landmark ruling on trans rights.[76] Prior to the ruling, transgender individuals – both children and adults – seeking to change their gender on official documents were required to sue their parents, due to a convoluted legal claim that there must be two opposing parties in any civil action, which applies in the case of legal gender recognition in Poland.[77] This added unnecessary distress and legal complexities.[78] The ruling eliminates the requirement for trans people to involve their parents in legal gender recognition proceedings.[79]

Despite this progress, LGBTI rights more broadly, and trans rights in particular, remain a contentious issue in Poland.[80] Although PiS is no longer the ruling party, harmful rhetoric continues to negatively affect legal gender-identity recognition and broader LGBTI rights, in a continuation of what activists and experts have described as "top-down polarization", which has seen political figures reinforce biased views and hostility against the LGBTI community, increasing polarisation on the issue.[81]

As the UN Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity concluded after a 2024 country visit to Poland:

"years of hostile rhetoric and discriminatory practice have left their mark".[82]

3.5 X'S STATUS IN POLAND'S POLITICAL AND INFORMATION LANDSCAPE

At the start of 2025, Poland was home to 34.5 million internet users – almost 90% of the population, and had 29 million social media users, equating to 75.6% of the total population.[83]

Numbers published in X's advertising resource indicate that X had 5.33 million users in Poland in early 2025, equivalent to 13.9% of the total population.[84] Data published by X's own advertising planning tools show that X's potential advertising reach in Poland decreased by 468,000 users (-8.4%) between October 2024 and January 2025.[85] This is line with a global decrease in X's potential audience reach.[86] However, it should be noted that advertising figures are not the same as monthly active user figures, which are not publicly available, and there may be a meaningful difference between the size of X's ad audiences and its total active user database.[87]

[73] Reuters, "Poland publishes civil partnership bill in boost for LGBT couples" (previously cited).
[74] Brussels Signal, "ECHR orders Poland to recognise same-sex partnerships", 28 April 2025, https://brusselssignal.eu/2025/04/echr-orders-poland-to-recognise-same-sex-partnerships/
[75] Brussels Signal, "ECHR orders Poland to recognise same-sex partnerships" (previously cited).
[76] Human Rights Watch, "Landmark Ruling on Trans Rights in Poland", 13 March 2025, https://www.hrw.org/news/2025/03/13/landmark-ruling-trans-rights-poland
[77] Human Rights Watch, "Landmark Ruling on Trans Rights in Poland" (previously cited).
[78] Human Rights Watch, "Landmark Ruling on Trans Rights in Poland" (previously cited).
[79] Human Rights Watch, "Landmark Ruling on Trans Rights in Poland" (previously cited).
[80] Human Rights Watch, "Landmark Ruling on Trans Rights in Poland" (previously cited).
[81] Euronews, "Queer in Poland: when can the LGBTQ+ community expect equal rights?" (previously cited); Human Rights Watch, "Landmark Ruling on Trans Rights in Poland" (previously cited).
[82] United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, "Country visit to Poland (18-29 November 2024): End-of-mission statement" (previously cited), para. 12.
[83] DataReportal, "Digital 2025: Poland", 3 March 2025, https://datareportal.com/reports/digital-2025-poland
[84] DataReportal, "Digital 2025: Poland" (previously cited).
[85] DataReportal, "Digital 2025: Poland" (previously cited).
[86] DataReportal, "X Users, Stats, Data & Trends for 2025", 12 March 2025, https://datareportal.com/essential-x-stats
[87] DataReportal, "Digital 2025: Poland" (previously cited).

X is considered an important place to get news and discuss political issues in Poland. For example, Mateusz Kaczmarek, a board member at the LGBTI organization Grupa Stonewall, told Amnesty International:

"It's the fastest way to find out about new information. When something happens, it could be easier to find out about it on Twitter than on Google or the whole internet."[88]

[88] Amnesty International interview with Matuesz Kaczmarek, 28 July 2024.

4. LEGAL FRAMEWORK

4.1 BUSINESS AND HUMAN RIGHTS STANDARDS

Under international law, states have an obligation to respect, protect and fulfil human rights. States also have the obligation to protect against human rights abuses by private actors, including through the regulation of companies and other economic actors, and to provide effective remedy when corporate actors within their territory or jurisdiction cause or contribute to human rights abuses. This requires taking appropriate steps to prevent, investigate, punish and redress such abuse through effective policies, legislation, regulations and adjudication. This duty is based on human rights treaties that the state has ratified and other international standards. States may breach their international law obligations where such abuse can be attributed to the state, or where they fail to take appropriate steps to prevent, investigate, punish and redress abuse by private actors.

Additionally, the UN Guiding Principles explicitly outline that states have an obligation to protect people from human rights harms linked to corporate activities.[89]

According to the same framework, all companies have a responsibility to respect human rights, regardless of their size, sector or where they operate.[90] This responsibility is independent of a state's own human rights obligations and exists over and above compliance with national laws and regulations protecting human rights.[91]

As an international instrument that is binding on signatory governments, the OECD Guidelines for Multinational Enterprises on Responsible Conduct (OECD Guidelines) reflect the expectation from governments to businesses on how to act responsibly, and OECD member countries such as Poland are required to ensure the OECD Guidelines are implemented and observed.[92] The OECD Due Diligence Guidance for Responsible Business Conduct (Due Diligence Guidance) was created to provide practical support to businesses in implementing the OECD Guidelines. According to the Due Diligence Guidance, an enterprise "contributes" to an adverse impact if its activities, in combination with the activities of other entities, cause, facilitate or incentivize another entity to cause an adverse impact.[93] For a determination of "contribution" to be made, the enterprise's contribution must be substantial, meaning it does not include minor or trivial contributions.[94]

Additionally, the UN Guiding Principles stipulate that, to meet the corporate responsibility to respect human rights, companies should have in place ongoing and proactive human rights due diligence processes to identify, prevent, mitigate and account for how they address their human rights impacts. When conducting this due diligence, a business enterprise might identify that it may contribute – or is already contributing to –

[89] UN Guiding Principles, Principle 1.
[90] This responsibility was expressly recognized by the UN Human Rights Council on 16 June 2011 when it endorsed the UN Guiding Principles, and on 25 May 2011 when the 42 governments that had then adhered to the OECD Declaration on International Investment and Multinational Enterprises unanimously endorsed a revised version of the OECD Guidelines for Multinational Enterprises. See, Human Rights and Transnational Corporations and other Business Enterprises, Human Rights Council, Resolution 17/4, 6 July 2011, UN Doc. A/HRC/RES/17/4; OECD, OECD Guidelines for Multinational Enterprises, 2011, https://www.oecd.org/en/topics/responsible-business-conduct.html
[91] UN Guiding Principles, Principle 11 including Commentary.
[92] OECD, "Guidelines for Multinational Enterprises on Responsible Business Conduct", 2023 edition, https://www.oecd.org/publications/oecd-guidelines-for-multinational-enterprises-on-responsible-business-conduct-81f92357-en.htm p. 18, para. 17.
[93] OECD,"Due Diligence Guidance for Responsible Business Conduct", 2018, https://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-Business-Conduct.pdf, p. 70.
[94] OECD, "Due Diligence Guidance for Responsible Business Conduct" (previously cited), p. 70.

human rights abuses. If such a finding occurs, the business enterprise must prevent or cease the negative human rights impacts.[95]

To verify whether adverse human rights impacts are being addressed, business enterprises should track the effectiveness of their responses.[96] Tracking should be "based on appropriate qualitative and quantitative indicators" and "draw on feedback from both internal and external sources, including affected stakeholders".[97]

4.2 THE RIGHTS OF LGBTI PEOPLE IN BUSINESS CONTEXTS

In 2017 the UN Office of the High Commissioner for Human Rights (OHCHR) outlined standards of conduct for businesses to tackle discrimination against LGBTI people.[98] The standards provide guidance to companies on the responsibility to respect human rights specifically as regards the rights of LGBTI people.[99] The standards also state that "where higher levels of human rights violations against LGBTI people have been documented, including in countries with discriminatory laws and practices, companies will need to undertake more extensive due diligence to ensure they respect the rights of LGBTI people."[100]

In a 2024 report, the UN Working Group on Business and Human Rights reiterated that states and businesses must increase their efforts to address the disproportionate adverse human rights impacts that LGBTI people face in the context of business activities, noting that: "The risks faced by individuals in the LGBTI+ community are diverse, and the structural and intersectional discrimination they endure is often misunderstood or inadequately addressed by both States and businesses".[101]

The Working Group highlighted that LGBTI people suffer discrimination and attacks in various forms, including stigmatization on social media.[102] The report specifies that social media platforms are too often failing in their responsibility to respect the human rights of the LGBTI community, noting that:

"In the technology sector, digital platforms have served as hubs for abusive content. Worse, these platforms rely heavily on automated content moderation systems that overlook both human rights considerations and linguistic diversity. The algorithms used by social media platforms fail to identify hate speech terms in different dialects, rendering the company-based grievance mechanisms ineffective in addressing these issues. To ensure that grievance mechanisms effectively address the barriers to remedy that users routinely face, consulting LGBTI+ persons is crucial."[103]

The Working Group's report provides guidance for businesses to ensure they are meeting their responsibility to respect the human rights of LGBTI people, including through a gender-responsive human rights due diligence process with meaningful engagement from the LGBTI community.[104]

[95] UN Guiding Principles, Commentary to Principle 19.
[96] UN Guiding Principles, Principle 20.
[97] UN Guiding Principles, Commentary to Principle 21.
[98] OHCHR, "Tackling Discrimination against Lesbian, Gay, Bi, Trans & Intersex People: Standards of Conduct for Business", 2017, https://www.unfe.org/sites/default/files/documents/UN-Standards-of-Conduct_0.pdf
[99] OHCHR, "Tackling Discrimination against Lesbian, Gay, Bi, Trans & Intersex People: Standards of Conduct for Business" (previously cited).
[100] OHCHR, "Tackling Discrimination against Lesbian, Gay, Bi, Trans & Intersex People: Standards of Conduct for Business" (previously cited).
[101] OHCHR, "UN Working Group calls for urgent efforts to address violations and abuses of the rights of LGBTI+ persons in business contexts", 1 November 2024, https://www.ohchr.org/en/press-releases/2024/11/un-working-group-calls-urgent-efforts-address-violations-and-abuses-rights
[102] OHCHR, "UN Working Group calls for urgent efforts to address violations and abuses of the rights of LGBTI+ persons in business contexts" (previously cited).
[103] Working Group on the issue of human rights and transnational business enterprises, "Protecting and respecting the rights of lesbian, gay, bisexual, transgender and intersex persons in the context of business activities: fulfilling obligations and responsibilities under the UN Guiding Principles on Business and Human Rights", 18 July 2024, UN Doc. A/79/178, p. 22.
[104] OHCHR, "UN Working Group calls for urgent efforts to address violations and abuses of the rights of LGBTI+ persons in business contexts" (previously cited).

4.3 HUMAN RIGHTS DUE DILIGENCE AND TECH COMPANIES

The UN Guiding Principles provide an important and relevant standard which tech companies should follow, including the need to conduct due diligence on their algorithmic technologies, such as automated content moderation. In 2020, the OHCHR outlined the relevance of the UN Guiding Principles for technology companies, stating that:

"The [UN Guiding Principles] set out a principled approach for all companies – regardless of industry sector, size, structure or operating context – to identify risks to people and to take action to prevent or mitigate them. This includes the expectation that technology companies make efforts to anticipate and mitigate harms that might occur related to the use of their products and services."[105]

The OHCHR specifies that tech companies' due diligence processes must also include addressing situations in which "business model-driven practices and design decisions create and exacerbate human rights risks", and an analysis that looks at the unique human rights risks posed by different products and services, end users and contexts of use.[106] Additionally, "substantive standards for artificial intelligence systems" set out by the Special Rapporteur on the right to freedom of expression specify that: "Companies should orient their standards, rules and system design around universal human rights principles".[107]

Furthermore, in 2021, the OHCHR set out recommendations to companies for assessing the risks related to artificial intelligence (AI), which included:[108]

- Systematically conduct human rights due diligence through the life cycle of the AI systems they design, develop, deploy, sell, obtain or operate. A key element of their human rights due diligence should be regular, comprehensive human rights impact assessments.
- Dramatically increase the transparency of their use of AI, including by adequately informing the public and affected individuals and enabling independent and external auditing of the automated systems. The more likely and serious the potential or actual human rights impacts linked to the use of the AI are, the more transparency is needed.
- Ensure participation of all relevant stakeholders on the development, deployment and use of AI, in particular affected individuals and groups.
- Advance the explainability of AI-based decisions, including by funding and conducting research towards that goal.

International human rights law can also provide important guidance with regard to approaches to content moderation and, if implemented transparently and consistently with meaningful user and civil society input, it can provide a framework for holding both states and companies accountable to users across national borders.[109] Human rights principles also enable companies to create an inclusive environment that accommodates the varied needs and interests of their users while establishing predictable and consistent baseline standards of behaviour.[110]

In 2018, the then UN Special Rapporteur on freedom of expression outlined the steps that companies can take to embed a human-rights-by-default approach to content moderation.[111] Measures suggested in the report included: moving their terms of service away from a discretionary approach rooted in generic and self-serving "community" needs and adopting high-level policy commitments to maintain platforms for users to develop opinions, express themselves freely and access information of all kinds.[112] These commitments should govern their approach to content moderation, ensuring that content-related actions will be guided by

[105] OHCHR, The UN Guiding Principles in the Age of Technology: A B-Tech Foundational Paper", September 2020, https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/introduction-ungp-age-technology.pdf
[106] OHCHR, "Addressing Business Model Related Human Rights Risks: A B-Tech Foundational Paper", July 2020, https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/B_Tech_Foundational_Paper.pdf
[107] Special Rapporteur on the promotion and protection of the right to freedom of expression, Report: *Artificial Intelligence Technologies and Implications for Freedom of Expression and the Information Environment*, 29 August 2018, UN Doc. A/73/348, para. 12.
[108] OHCHR, "The right to privacy in the digital age", 15 September 2021, UN Doc. A/HRC/48/31.
[109] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report: *A Human Rights Approach to Platform Content Regulation*, 6 April 2018, UN Doc. A/HRC/38/25.
[110] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report: *A Human Rights Approach to Platform Content Regulation* (previously cited).
[111] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report: *A Human Rights Approach to Platform Content Regulation* (previously cited).
[112] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report: *A Human Rights Approach to Platform Content Regulation* (previously cited).

the same standards of legality, necessity and legitimacy that bind state regulation of expression.[113] To further align content moderation practices with human rights law, companies should take steps to clarify and specify their content rules so that users can predict with reasonable certainty what content would place them in violation of the platform rules.[114] This could also entail companies disclosing data and examples that provide insight into the factors they assess in determining a violation. In the context of hate speech, explaining how specific cases are resolved may help users to better understand how companies approach difficult distinctions between inoffensive content and incitement to hatred, or how considerations such as the intent of the speaker or the likelihood of violence are assessed in online contexts.[115]

The Global Network Initiative (GNI) – which brings together key stakeholders from academia, civil society, companies and investors – has also developed guidance for content moderation based on human rights principles, with a particular focus on the need to address legitimate public policy concerns around harmful conduct and content online while respecting human rights.[116] GNI's guidance asserts that processes for legislative deliberation on this issue should therefore be open and non-adversarial, drawing on broad expertise to ensure that the results are well-thought out and evidence based.[117] It is particularly important that states take time to understand and consider actions that are consistent with international human rights obligations and appropriate and proportionate to their jurisdiction.[118] Additionally, though companies have responsibilities and an important role to play in addressing online harm, lawmakers should refrain from shifting all the legal liability from those generating illegal content onto intermediaries.[119]

The GNI guidance further recommends that strong transparency, remedy and accountability measures are included in any legislation that addresses content moderation practices and be narrowly tailored to address the services that pose the greatest risk of harm, with relevant exceptions and appropriate safeguards.[120] The importance of human review of content flagged by automated tools is also highlighted in the guidance, as well as provisions ensuring the right to an effective remedy in response to content restrictions.[121] Content moderation decisions must carefully balance the right to freedom of expression with other rights such as the right to non-discrimination and, to this end, the guidance suggests that carving out or providing affirmative defences for particularly vulnerable groups may help ensure that laws are narrowly tailored to meet their objectives.[122]

4.4 THE CORPORATE RESPONSIBILITY TO PROVIDE REMEDY

Access to remedy is a key pillar of the business and human rights framework. The UN Guiding Principles stipulate that, where "business enterprises identify that they have caused or contributed to adverse impacts, they should provide for or cooperate in their remediation through legitimate processes".[123] Potential impacts – or human rights risks – require action to prevent harm or mitigate the risks as far as possible. It is therefore impossible for any business enterprise to meet the responsibility to respect human rights if it contributes to human rights abuses and fails to meaningfully remedy the adverse impact.[124]

---

[113] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report: *A Human Rights Approach to Platform Content Regulation* (previously cited), para. 44; GNI, *Content Regulation and Human Rights*, October 2020, https://globalnetworkinitiative.org/wp-content/uploads/GNI-Content-Regulation-HR-Policy-Brief.pdf
[114] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report: *A Human Rights Approach to Platform Content Regulation* (previously cited).
[115] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report: *A Human Rights Approach to Platform Content Regulation* (previously cited).
[116] GNI, *Content Regulation and Human Rights* (previously cited).
[117] GNI, *Content Regulation and Human Rights* (previously cited).
[118] GNI, *Content Regulation and Human Rights* (previously cited).
[119] GNI, *Content Regulation and Human Rights* (previously cited).
[120] GNI, *Content Regulation and Human Rights* (previously cited).
[121] GNI, *Content Regulation and Human Rights* (previously cited).
[122] GNI, *Content Regulation and Human Rights* (previously cited).
[123] UN Guiding Principles, Principle 22.
[124] OHCHR, "Frequently asked questions about the UN Guiding Principles on Business and Human Rights", 2014, https://www.ohchr.org/en/publications/special-issue-publications/frequently-asked-questions-about-guiding-principles Question 35, p. 36.

4.5 OBLIGATIONS UNDER THE DIGITAL SERVICES ACT (DSA)

The EU has led attempts to regulate social media companies and algorithmic technologies, passing the DSA, a legally binding regulatory framework, in 2022.[125] The DSA is enforced by national authorities, and it is the responsibility of EU member states to designate the authority or authorities in charge of enforcement at the national level.[126] As a legally binding framework, the DSA provides penalties for non-compliance, such as fines amounting to up to 6% of a company's global annual turnover and inspections at their premises, which includes the right to ask the company to give access and explanations in relation to its algorithms, data-handling and business practices.[127]

The DSA mandates social media platforms to be transparent in their content moderation, and social media companies are obliged to disclose their moderation policies and how they are implemented.[128] The legislation also includes provisions around content moderation, including obligations for social media platforms to include effective safeguards for users, such as the possibility to challenge platforms' content moderation decisions based on obligatory information that platforms must provide to users when their content is removed or restricted.[129]

Article 34(1) of the DSA introduces obligations on VLOPs to assess and mitigate systemic risks that arise from the "design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services".[130]

Article 34(1)(b) stipulates that the risk assessment should be specific to the services provided by VLOPs and that they should take into account the severity and probability of risks, including:

"any actual or foreseeable negative effects for the exercise of fundamental rights, in particular the fundamental rights to human dignity enshrined in Article 1 of the Charter, to respect for private and family life enshrined in Article 7 of the Charter, to the protection of personal data enshrined in Article 8 of the Charter, to freedom of expression and information, including the freedom and pluralism of the media, enshrined in Article 11 of the Charter, to non-discrimination enshrined in Article 21 of the Charter, to respect for the rights of the child enshrined in Article 24 of the Charter and to a high level of consumer protection enshrined in Article 38 of the Charter."

Similarly to the international standards outline above, the DSA requires that relevant stakeholders are included in social media companies' due diligence processes around identifying and mitigating systemic risks and that they test their assumptions with groups most affected by the risks.[131]

The DSA requires VLOPs to put in place reasonable, proportionate and effective mitigation measures, tailored to identified systemic risks and with particular consideration given to the impacts of such measures on fundamental rights.[132] Such measures may include adapting the design, features or functioning of their services, including their online interfaces, adapting content moderation processes, and testing and adapting their algorithmic recommender systems.[133]

Providers of VLOPs are subject to yearly independent audits to assess compliance with their due diligence operations.[134] The obligations contained in the DSA have applied to designated VLOPs (including X) since August 2023, and to all digital platforms since February 2024.[135]

---

[125] Amnesty International, *What the EU's Digital Services Act Means for Human Rights* (Index: POL 30/5830/2022), 7 July 2022, *https://www.amnesty.org/en/documents/pol30/5830/2022/en/*
[126] Amnesty International, *What the EU's Digital Services Act Means for Human Rights* (previously cited).
[127] Amnesty International, *What the EU's Digital Services Act Means for Human Rights* (previously cited).
[128] European Commission, "Questions and answers on the Digital Services Act", 23 February 2024, https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348
[129] European Commission, "Questions and answers on the Digital Services Act" (previously cited).
[130] Digital Services Act, Article 34 (1)
[131] Digital Services Act, Recital 90.
[132] Digital Services Act, Article 35.
[133] Digital Services Act, Article 35.
[134] Amnesty International, *What the EU's Digital Services Act Means for Human Rights* (previously cited).
[135] European Commission, "The Digital Services Act", https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en (accessed on 2 July 2025).

4.6 THE RIGHT TO LIVE FREE FROM GENDER-BASED VIOLENCE

International human rights law obliges states to uphold the right to live free from GBV. GBV encompasses a wide range of violence, including physical, sexual and psychological violence, threats, abuse and coercions that are rooted in and produce gender inequality, power asymmetry and harmful gender stereotypes and social norms. GBV has a disproportionate effect on women and girls but also affects other people owing to their real and/or perceived gender, sexual orientation, gender identity and/or expression. GBV is a form of discrimination and may in some cases amount to torture or other ill-treatment. The definition of GBV also covers violence "occurring online and in other digital environments".[136]

UN human rights mechanisms and bodies have increasingly recognized that discrimination based on sexual orientation, gender identity and/or expression and sex characteristics (SOGIESC) plays a crucial role in shaping and exacerbating GBV, including technology-facilitated gender-based violence (TfGBV). Amnesty International understands TfGBV to be any act of GBV, or threat thereof, perpetrated by one or more individuals that is committed, assisted, aggravated and/or amplified in part or fully using information and communication technologies or digital media. While TfGBV disproportionately affects women and girls, it can also affect other people based on their real and/or perceived gender or SOGIESC, causing physical, psychological, economic, social and sexual harm.

In its General Recommendation 35 on GBV against women, the UN Committee on the Elimination of Discrimination against Women (CEDAW Committee) reaffirmed its stance that forms of discrimination against women are intersectional, being "inextricably linked to other factors" which include being a lesbian, bisexual or transgender women or an intersex person.[137] In a report about SOGIESC-based discrimination to the UN Human Rights Council in November 2011, the then High Commissioner for Human Rights further acknowledged that homophobic and transphobic attacks constitute GBV.[138] Furthermore, the Yogyakarta Principles relating to the application of international law to SOGIESC protect the human rights of LGBTI people in relation to information and communication technology. Principle 36 of the Yogyakarta Principles states that LGBTI people are entitled to the same level of protection online and offline and that LGBTI people have the right to use and access communication technology without violence and discrimination based on SOGIESC.[139]

International human rights law requires states to ensure that both state and non-state actors respect LGBTI people's right to live free from GBV, including TfGBV.[140] States must also take all necessary steps to protect those subjected to GBV, including TfGBV,[141] investigate these offences, bring perpetrators to justice, and provide survivors with access to justice and timely and appropriate reparation.[142] Additionally, states must take measures to prevent TfGBV, by raising awareness about this issue and establishing support services for all people who have experienced GBV.[143] In doing so, it is fundamental to take into account, with an intersectional approach, the ways in which race, ethnic background and socio-economic status can shape experiences of TfGBV in varying contexts.[144]

The right to live free from GBV is indivisible from and interdependent on other human rights, including but not limited to the rights to privacy, freedom of expression, freedom of peaceful assembly and freedom of association.[145]

When people experience TfGBV, it becomes more dangerous for them to engage and participate in online conversations and to benefit from digital technologies, such as social media platforms.[146] It can also lead to severe psychological harms that affect their mental health, including experiencing depression, anxiety and thoughts of self-harm. TfGBV can force targets to withdraw and can limit their ability to use the internet, with

[136] CEDAW Committee, General Recommendation 35: Gender-Based Violence Against Women, Updating General Recommendation 19 (1992), 26 July 2017, UN Doc. CEDAW/C/GC/35, para. 20.
[137] CEDAW Committee, General Recommendation 35 (previously cited), para.12.
[138] OHCHR, Report: *Discriminatory Laws and Practices and Acts of Violence Against Individuals Based on Their Sexual Orientation and Gender Identity*, 17 November 2011, UN Doc. A/HRC/19/41, para. 20.
[139] The Yogyakarta Principles +10, Principle 36.
[140] UN Special Rapporteur on violence against women, Report: *Online violence against women and girls from a human rights perspective*, 18 June 2018, UN Doc. A/HRC/38/47, para. 22.
[141] UN Special Rapporteur on violence against women, Report: *Online violence against women and girls from a human rights perspective* (previously cited), para. 67.
[142] CEDAW Committee, General Recommendation 35 (previously cited), para. 29.
[143] CEDAW Committee, General Recommendation 35 (previously cited), para. 31(iii).
[144] CEDAW Committee, General Recommendation 35 (previously cited), para. 12.
[145] CEDAW Committee, General Recommendation 35 (previously cited), para.15.
[146] Amnesty International, *Human Rights Implications of Technology-Facilitated Gender-Based Violence: Submission to the Human Rights Council Advisory Committee* (Index: IOR 40/9284/2025), 24 April 2025, https://www.amnesty.org/en/documents/ior40/9284/2025/en/

implications for a broad range of human rights, including the realization of the rights to education, to freedom of association and assembly, to participate in social, cultural and political life, to health, to an adequate standard of living, to work and to social and economic development.[147]

Amnesty International uses the term TfGBV to refer to violence against LGBTI people because these forms of violence are 'gender-based', where gender is understood to be a set of socially constructed social norms, roles and behaviours associated with a person's assigned sex at birth, which serves to uphold cis-heteropatriarchy[148]. Subsequently, while GBV does disproportionately affect women and girls, it also affects others when the root cause of the violence is to preserve, uphold and maintain gendered roles, norms, social systems and power structures, even as its manifestation may vary across different groups.

4.7 THE PROHIBITION OF ADVOCACY OF HATRED UNDER INTERNATIONAL HUMAN RIGHTS LAW

Under international human rights law and standards, advocacy of hatred must be prohibited, although hateful expressions ought to be considered in light of both the right to freedom of expression and the rights to equality and non-discrimination. The right to freedom of expression protects many forms of speech, even speech which may be considered deeply offensive, shocking or disturbing.[149] However, the right to freedom of expression is not absolute and it can be restricted under certain circumstances, including when it is necessary and proportionate to protect the rights of others.

The right to equality and non-discrimination, a critical component of international human rights law, constitutes a "basic and general principle relating to the protection of human rights".[150] Individuals whose right to non-discrimination is violated must have access to effective remedy. This is affirmed by the Toronto Declaration – a civil society-led statement based on international human rights law outlining principles of this fundamental right in the use of machine learning and AI.[151] The declaration states:

"Companies and private sector actors designing and implementing machine learning systems should take action to ensure individuals and groups have access to meaningful, effective remedy and redress. This may include, for example, creating clear, independent, visible processes for redress following adverse individual or societal effects, and designating roles in the entity responsible for the timely remedy of such issues subject to accessible and effective appeal and judicial review."[152]

As made clear by Article 20 of the International Covenant on Civil and Political Rights (ICCPR), advocacy of hatred is more than just the expression of ideas or opinions that are hateful towards members of a particular group. It requires a clear showing of intent to incite others to discriminate, be hostile (experience intense or irrational emotions of opprobrium, enmity and detestation) toward, or commit violence against, the group in question. Laws prohibiting advocacy of hatred must also comply with the ICCPR's provisions on freedom of expression, and must meet the requirements of legality, legitimate aim, necessity and proportionality.

The Rabat Plan of Action on the prohibition of national, racial and religious hatred constituting incitement to discrimination, hostility or violence suggests a six-part threshold test to guide states' implementation of the prohibition of advocacy of hatred. The six factors that should be considered when determining if an expression constitutes advocacy of hatred are: context; the speaker's position or status; intent; content and form; the extent of the speech act; and the likelihood – including imminence – of harm.[153]

The Rabat Plan of Action also distinguishes between forms of expression that advocate hatred that constitute incitement to violence, hostility or discrimination that must be prohibited; and forms of expression that are not criminal but still raise concerns in terms of tolerance, civility and respect for the convictions of others.[154]

---

[147] Amnesty International, *Human rights implications of technology-facilitated gender-based violence: Submission to the Human Rights Council Advisory Committee* (previously cited).
[148] Cis-heteropatriarchy refers to a social system where cisgender, heterosexual males hold a dominant position of power and privilege, influencing society structures and norms to their benefit, often at the expense of women, LGBTI individuals, and other marginalized groups.
[149] UN Human Rights Committee (HRC), General Comment 34: Article 19: Freedoms of Opinion and Expression, 12 September 2011, UN Doc. CCPR/C/GC/34, para. 11.
[150] HRC, General Comment 18: Non-Discrimination, 1989, UN Doc. RI/GEN/1/Rev.9 Vol I, para. 1.
[151] Amnesty International and Access Now, *The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems* (Index: POL 30/8447/2018), 17 May 2018, https://www.amnesty.org/es/documents/pol30/8447/2018/en/
[152] Amnesty International and Access Now, *The Toronto Declaration* (previously cited), para. 53.
[153] UN Human Rights Council, Rabat Plan of Action, 11 January 2013, UN Doc. A/HRC/22/17/Ad.4, para. 29.
[154] Rabat Plan of Action (previously cited), para. 20.

For the purposes of this report, the use of the term "advocacy of hatred" refers to expression that constitutes incitement to discrimination, hostility or violence that must be prohibited in law in accordance with Article 20 of the ICCPR. In addition, the report also addresses the spread of expression that may not reach the threshold of "advocacy of hatred" but still raises concerns in terms of tolerance, civility and respect for others, affecting the right to non-discrimination and equality.

Amnesty International has not sought to make determinations about whether specific pieces of content on X should be considered "advocacy of hatred". Rather, this report is intended to provide an analysis of X's overall contribution to human rights abuses against LGBTI people in Poland, due to its failure to adequately mitigate the risks of the platform.

5. THE ROLE OF X IN SPREADING TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE IN POLAND

> **"The hate resonates there."**
>
> Aleksandra Herzyk, activist.

This section outlines the role that X has played in the spread of anti-LGBTI content, including hateful content, between 2019 and 2025.

5.1 TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE ON X

Attacks against LGBTI users on English-speaking X have increased substantially since Elon Musk bought the company in October 2022.[155] A week after Elon Musk's takeover, anti-rights figures appeared to begin testing X's boundaries for anti-LGBTI speech.[156] 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

[155] Amnesty International USA, "Hateful and abusive speech towards LGBTQ+ community surging on Twitter under Elon Musk", 9 February 2023, https://www.amnestyusa.org/press-releases/hateful-and-abusive-speech-towards-lgbtq-community-surging-on-twitter-under-elon-musk/; NBC News, "Twitter is the 'most dangerous platform for LGBTQ people', GLAAD says" (previously cited).
[156] NBC News, "A timeline of Elon Musk's takeover of Twitter", 17 November 2022, https://www.nbcnews.com/business/business-news/twitter-elon-musk-timeline-what-happened-so-far-rcna57532

went on to say, "I was suspended for this exact tweet a month ago so we will see if Twitter is now free."[157] The conservative podcaster Matt Walsh tweeted: "We have made huge strides against the trans agenda. In just a year we've recovered many years' worth of ground conservatives had previously surrendered. The liberation of Twitter couldn't have come at a more opportune time. Now we can ramp up our efforts even more."[158]

In a 2022 survey of LGBTI activists and organizations, conducted by Amnesty International USA and the LGBTI rights organizations GLAAD and Human Rights Campaign, 60% of respondents reported they had experienced an increase in abusive and hateful speech on X since October 2022.[159] The remaining 40% reported that they experienced the same level of abusive and hateful speech as before.[160] None of the respondents reported a decrease in abusive and hateful speech.[161] Moreover, 60% of all respondents said that hateful and abusive speech had affected how they used the platform, including posting to X less frequently, sharing less information regarding their work, and limiting with whom they interact on the platform.[162] Additionally, 65% of the respondents said they believed there was more hateful and abusive speech on X compared to other platforms they use.[163]

Many of these sentiments were echoed by the LGBTI community members interviewed by Amnesty International in Poland in July and August 2024.

Jakub Szymik, a gay man based in the capital, Warsaw, told Amnesty International that he limits what he posts on X to avoid being targeted with hate:

"I definitely try to keep a professional profile, so I focus on things that are not controversial, or I don't share opinions online under my name. I have anonymous accounts where I am freer, but I won't do it in my own name... I really feel that checking news online or opinions online will make me more vigilant... and it's hard to relax."[164]

Misza, a non-binary person living in Poznan, also explained that anonymity was key to enabling them to continue using X:

"On Twitter, I'm not [using] my real name, I'm [using] my nickname. I don't take photos of my face or my family's face because I am afraid."[165]

Maja Heban, a trans woman and activist based in Warsaw, told Amnesty International she believes that there is more hate on X than on other social media platforms:

"I get an insane amount of hate on social media and Twitter. This is just way beyond anything else, I sometimes get negative comments on Facebook or on Instagram, but it's nothing compared to Twitter. And Twitter is basically a never-ending stream of deadnaming, misgendering, insults, death wishes."[166]

Interviewees gave Amnesty International examples of the anti-LGBTI content they have seen circulating on the platform:

"There are lots of comments about this community being sick, destroying national values... use of emojis that are connected to vomit, to shit, that sort of thing. There is lots of content that is public threats, like saying 'you're next'".[167]

"[They say] LGBTQ people will be in gas chambers, or they talk like we are trash, and they think that we have to be cleansed."[168]

---

[157] NBC News, "Far-right figures appear to be testing Twitter's boundaries for anti-LGBTQ speech", 2 November 2022, https://www.nbcnews.com/feature/nbc-out/open-season-lgbtq-people-far-right-celebrates-liberation-twitter-rcna54542
[158] NBC News, "Far-right figures appear to be testing Twitter's boundaries for anti-LGBTQ speech" (previously cited).
[159] Amnesty International USA, "Hateful and abusive speech towards LGBTQ+ community surging on Twitter under Elon Musk" (previously cited).
[160] Amnesty International USA, "Hateful and abusive speech towards LGBTQ+ community surging on Twitter under Elon Musk" (previously cited).
[161] Amnesty International USA, "Hateful and abusive speech towards LGBTQ+ community surging on Twitter under Elon Musk" (previously cited).
[162] Amnesty International USA, "Hateful and abusive speech towards LGBTQ+ community surging on Twitter under Elon Musk" (previously cited).
[163] Amnesty International USA, "Hateful and abusive speech towards LGBTQ+ community surging on Twitter under Elon Musk" (previously cited).
[164] Amnesty International video call with Jakub Szymik, 5 August 2024.
[165] Amnesty International video call with Misza (pseudonym), 24 July 2024.
[166] Amnesty International interview with Maja Heban, 30 July 2024.
[167] Amnesty International video call with Jakub Szymik, 5 August 2024.
[168] Amnesty International video call with Misza, 24 July 2024.

"Twisted people, that we are broken, that we are sick. I think sick is the most used word."[169]

"[They say] these people are not normal, they are against Polish families, they are destroying Polish families, they are not people, they are [an] ideology."[170]

Aleksy told Amnesty International that posts on X by the LGBTI rights organization for which he works were often subject to homophobic and transphobic comments:

"There are private users who are publishing homophobic things, or some comments... I have an impression that under almost any tweet about Pride marches, there is a lot of disgusting comments... We usually try to make our social media a safe space so we try to hide these more triggering comments but on Twitter, when it's happening so fast, it's very difficult."[171]

5.2 THE ROLE OF POLAND'S PUBLIC AND POLITICAL FIGURES IN SPREADING ANTI-LGBTI CONTENT

Globally, X is viewed as a platform on which to engage with and comment on the news cycle.[172] Aleksy, who works at a Polish LGBTI rights organization, explained the role of X in Poland's news and information landscape to Amnesty International, saying:

"Twitter is used mostly by politicians and journalists in Poland... it's one of the social media [platforms] where if you talk to a politician, he or she is more likely to answer you than on others... and it's fast. So, a lot of decisions are communicated first on Twitter. It is a good tool to know what is happening and to make pressure [on politicians]."[173]

Aleksy added that, during PiS's time in government, the LGBTI community was often targeted:

"We faced a lot of homophobic hate speech by some politicians, for example, the LGBTI community was [described] not as people but as an ideology."[174]

Jakub Szymik explained that X was a prominent platform for politicians to foment anti-LGBTI sentiment:

"I think Twitter is a tool, like in a very broad spectrum it is one cog in the system. Over the last eight years in Poland the majority of this hate speech in the public space was fuelled by political actors and obviously they used Twitter because it allows for the fast spread of information."[175]

This fast spread of information on X, powered by engagement-centric algorithms, has long been a feature of the platform's business model as the company sought to create the sense that it was a real-time news feed.[176] In 2009, during the earliest days of Twitter, co-founder Biz Stone wrote that the platform would become a "new kind of information network."[177]

Andrzej Duda, the former President of Poland from 2015 to 2025, is an active X user, and at the time of writing in 2025, had 1.9 million followers on the platform.[178] His active use of X is perhaps indicative of the importance of X in Poland's political sphere, and of an understanding among political actors that social media platforms, and in particular X, are key battlegrounds in shaping the narrative on political and social issues.

Content featuring Andrzej Duda making anti-LGBTI statements has been posted on X. For example, on 13 June 2020, a clip of Andrzej Duda claiming that LGBTI people are not human began to circulate on the platform.[179] The clip is indicative of the political rhetoric utilized by PiS ahead of the June 2020 presidential

[169] Amnesty International video call with Nathan Bryza, 9 August 2024.
[170] Amnesty International video call with Jolanta Prochowicz, 29 July 2024.
[171] Amnesty International interview with Aleksy (pseudonym), 26 July 2024.
[172] X, "How many people come to Twitter for news? As it turns out, a LOT", 12 September 2022, https://blog.x.com/en_us/topics/insights/2022/how-many-people-come-twitter-for-news
[173] Amnesty International interview with Aleksy (pseudonym), 26 July 2024.
[174] Amnesty International interview with Aleksy (pseudonym), 26 July 2024.
[175] Amnesty International video call with Jakub Szymik, 5 August 2024.
[176] The Verge, "How Twitter broke news", 12 December 2023, https://www.theverge.com/c/features/23993135/twitter-breaking-news-history
[177] X, "What's Happening?", 19 November 2009, https://blog.x.com/en_us/a/2009/whats-happening
[178] Andrzej Duda's X profile, https://x.com/AndrzejDuda
[179] Bartosz T. Wielinski, X post, 13 June 2020, https://x.com/Bart_Wielinski/status/1271773086833094656

election, in which the rights of LGBTI people became one of the main political issues.[180] As well as election-related clips circulating online, anti-LGBTI content was also posted directly to the platform by politicians. For example, Joachim Brudziński, a Member of the European Parliament for PiS and head of Andrzej Duda's re-election campaign, tweeted in June 2020 that "Poland is the most beautiful without LGBT".[181] At the time of writing, the post remains available on X.

After the PiS victory in the 2020 presidential election, which returned Andrzej Duda to power, prominent figures in the Polish anti-rights populist movement continued to post anti-LGBTI content on X, possibly emboldened by the success of the election rhetoric.[182] For example, two weeks after the presidential election, right-wing figure Rafał A. Ziemkiewicz (who had 334,200 followers on X as of May 2025) tweeted about an action where LGBTI activists draped a rainbow flag over a statue of Jesus: "I encourage all those who, whether out of meanness or naivety,say that the sodomite banner imposed on Jesus 'does not offend' to throw pork into the mosque and then go to their Muslim brothers and tell them that pigs are cool and cuddly, and healthy meat cannot offend anyone".[183] At the time of writing, the post remains available on X.

5.3 ANALYSIS: HOW TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE CONTRIBUTES TO OFFLINE HARM IN POLAND

Anti-LGBTI content – like other content targeting marginalized groups – does not exist in a vacuum or an online-only space, separated from the offline world. Jolanta Prochowicz, a lesbian woman based in Lublin, explained this connection to 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."[184]

In 2024, the LGBTI advocacy organization GLAAD noted: "[A]s we have seen over and over again – there is a direct line from dangerous online rhetoric and targeting to violent offline behaviour against the LGBTI community."[185] This sentiment is echoed in a 2024 Online Extremism report from the US Government Accountability Office: "Research suggests the occurrence of hate crimes is associated with hate speech on the internet [and] suggests individuals radicalized on the internet can perpetuate violence".[186] Amnesty International has also previously found that the harms of TfGBV specifically are not confined to digital spaces, and that the online-offline continuum of GBV means that threats online can often have offline consequences, including physical violence.[187]

Academic scholars have argued that hate speech also has grave implications on a societal level as it can "poison societies by threatening individual rights, human dignity and equality, reinforcing tensions between social groups, disturbing public peace and public order and jeopardizing peaceful co-existence".[188]

TfGBV should be understood to be a continuum of GBV that exists between the online and offline and, as such, some forms of TfGBV can also result in the physical violation of rights offline.[189] Recent research by the Centre for Hate Studies at the University of Leicester, UK, highlighted that, while much of the anti-LGBTI

[180] Politico, "Poland's LGBTQ community in the political crosshairs", 19 June 2020, https://www.politico.eu/article/poland-lgbtq-community-in-the-political-crosshairs-elections-duda/
[181] Joachim Brudzinski, X post, 11 June 2020, https://x.com/jbrudzinski/status/1271167309269413892
[182] Global Project Against Hate and Extremism (GPAHE), "Twitter, YouTube allowing anti-LGBT hate speech, antisemitism and racism to thrive in Poland", 18 August 2020, https://globalextremism.org/post/twitter-youtube-allowing-antisemitism-racism-and-anti-lgbt-hate-speech-to-thrive-in-poland/
[183] Rafal A. Ziemkiewicz, X post, 12 August 2020, https://x.com/R_A_Ziemkiewicz/status/1293565844128178177
[184] Amnesty International video call with Jolanta Prochowiz, 29 July 2024.
[185] GLAAD, *Social Media Safety Index*, 2024, https://glaad.org/smsi/social-media-safety-index-2024/ p. 2.
[186] United States Government Accountability Office, "Online extremism: more complete information needed about hate crimes that occur on the internet", January 2024, https://www.gao.gov/assets/d24105553.pdf p. 43.
[187] Amnesty International, *"Being Ourselves is Too Dangerous": Digital Violence and the Silencing of Women and LGBTI Activists in Thailand* (Index: ASA 39/7955/2024), 16 May 2024, https://www.amnesty.org/en/documents/asa39/7955/2024/en/; Amnesty International, *"Everybody Here is Having Two Lives or Phones": The Devastating Impact of Criminalization on Digital Spaces for LGBTQ People in Uganda* (Index: AFR 59/8571/2024), 23 October 2024, https://www.amnesty.org/en/documents/afr59/8571/2024/en/
[188] Judit Bayer and others, "Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States", 18 November 2021, European Parliament, LIBE Committee, Policy Department for Citizens' Rights and Constitutional Affairs, http://dx.doi.org/10.2139/ssrn.3409279
[189] European Parliamentary Research Service, "Combating gender-based violence: cyber violence", March 2021, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662621/EPRS_STU(2021)662621_EN.pdf

sentiment in Poland circulates online, its effects also manifest offline in public spaces. For example, in several cities, homophobic stickers were found in public areas, some of which included hidden razor blades designed to injure those trying to remove them.[190]

Additionally, many types of TfGBV such as online stalking and harassment have offline equivalents, and online violence – especially when normalized by political figures – can manifest as offline violence.[191]

Jakub Szymik shared that he felt he had to continue to engage with the hateful content on X, despite the effect on his well-being:

"I really feel that checking the news online or opinions online will make me more vigilant in what I can expect and it's harder to relax. I feel like I need to be connected all the time because it might make me prepared for some weird, undisclosed things that might happen in real life."[192]

Aleksy, who works for an LGBTI organization in Poland, expressed concerns about the effect that TfGBV on X has on younger people's right to freedom of expression:

"There are people who are underage, for example, or who are not sure or are afraid to come out and I think if they see all that hate speech, it is very difficult. It might be very difficult for them."[193]

Jakub Szymik, an LGBTI rights activist, told Amnesty International that the hate he sees on X has caused him to use the platform less frequently:

"Now I feel like it is a source of opinions and polarized opinions and this is something that discourages me from using the platform."[194]

TfGBV in Poland has also contributed to the significant mental distress of some members of the LGBTI community. Psychologist Julia Kata told Amnesty International that people who have been turning to the LGBTI rights organization Fundacja Trans-Fuzja for psychological support will often cite online hate as one of the issues they are struggling with:

"Especially for those who, the only sense of community, support and acceptance they have is online – if they're exposed to this kind of content, it just takes [away] all the feeling of security from them. They feel even more alienated. They are alienated in real life, and then in the place that they saw as their safe space."[195]

The mental health struggles of the Polish LGBTI community have also been highlighted by the UN Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity. After his 2024 country visit to Poland, the Independent Expert noted that these challenges are related to:

"sustained discrimination, instances of violence, or the threat of violence, as well as social ostracism and stigma".[196]

Julia Kata mentioned how this pervasive sense of vulnerability due to discrimination and violence is exacerbated by exposure to TfGBV:

"People don't want to have interactions with other people. Like in the real world. Because they are scared they will hear something harmful, hateful, or someone will be just not nice. It's just alienating on so many levels."[197]

The spread of anti-LGBTI content has been central to normalizing an ideology that dehumanizes LGBTI people in Poland, and at times inciting violence and discrimination against the community. The mass dissemination of these messages on an important social media platform like X has played a key role in stigmatizing Poland's LGBTI community – a stigmatization which continues to impact on the rights of LGBTI people in the country today.

---

[190] University of Leicester, "Centre for Hate Studies submits report on Anti-LGBTI hate crime in Poland for United Nations Review", 29 October 2024, https://le.ac.uk/news/2024/october/lgbt
[191] European Parliamentary Research Service, "Combating gender-based violence: cyber violence" (previously cited).
[192] Amnesty International video call with Jakub Szymik, 5 August 2024.
[193] Amnesty International interview with Aleksy (pseudonym), 26 July 2024.
[194] Amnesty International video call with Jakub Szymik, 5 August 2024.
[195] Amnesty International interview with Julia Kata, 29 July 2024.
[196] United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, "Country visit to Poland (18-29 November 2024): End-of-mission statement" (previously cited), para. 23.
[197] Amnesty International interview with Julia Kata, 29 July 2024.

5.4 CUTS TO CONTENT MODERATION

The spiralling levels of hateful content on X may in part be due to the drastic staff cuts made by the X owner and then CEO Elon Musk, which has resulted in a lack of content moderators for the platform.[198] In the early days of Elon Musk's takeover, as top executives were fired and half of the company's staff were laid off, a content moderation council was formed and tasked with reviewing account reinstatements.[199]

Shortly after buying the company, Elon Musk disbanded the Trust and Safety Council, an advisory group comprising almost 100 civil society, human rights and other organizations that sought to address child exploitation, suicide, self-harm and hate speech on the platform.[200] It is estimated that he also laid off 80% of the engineers dedicated to trust and safety.[201] Despite these cuts, the company's current policy on "Defending and respecting the rights of people using our service" positions the same much-depleted Trust and Safety team as a key partner for stakeholders across the company, working "in tandem with product, engineering, user services, sales, public policy, and legal to help X keep the people and organizations using our service front and center when we make decisions".[202]

The company previously also had a dedicated team working to combat coordinated disinformation campaigns, but experts and former staff claim that most of these specialists either resigned or were laid off after October 2022.[203]

Digital rights experts have raised concerns that the layoffs could compromise the platform's capacity to police harmful content – including forms of harassment – and advertisers said they would pull back due to "uncertainty" about Elon Musk's strategy.[204]

In late 2022, it was reported that Elon Musk planned to lean heavily on automation to moderate content, doing away with certain manual reviews and favouring restrictions on algorithmic distribution rather than removing certain speech completely.[205] This approach, known as "visibility filtering", involves leaving certain tweets visible that violate the company's policies but barring them from appearing in places like the home timeline and search.[206]

In early 2023, X reportedly began planning a new system to keep the most "undesirable", or harmful, content off the platform.[207] This new plan involved building a smaller, in-house team of content moderators, based in a new Trust and Safety Center of Excellence in Austin, Texas, and envisioned as a specialized safety net to prevent the most egregious content from slipping through without compromising too much on X's prioritization of free speech.[208]

It was reported that the Center would house 100 content moderators, significantly smaller than the 500-person team that was originally envisioned (or the approximately 1,500 content moderators X had in 2020).[209] However, by the time the Center opened, it was reported to be unclear whether X had managed to hire more than a dozen people.[210]

---

[198] NBC News, "Twitter is the 'most dangerous platform for LGBTQ people', GLAAD says" (previously cited).
[199] ABC News, "A timeline of Elon Musk's tumultuous Twitter acquisition", 11 November 2022, https://abcnews.go.com/Business/timeline-elon-musks-tumultuous-twitter-acquisition-attempt/story?id=86611191
[200] Amnesty International, "Twitter's decision to disband safety council threatens wellbeing of users", 13 December 2022, https://www.amnesty.org/en/latest/news/2022/12/global-twitters-decision-to-disband-safety-council-threatens-wellbeing-of-users/
[201] Fortune, "Inside the shifting plan at Elon Musk's X to build a new team and police a platform 'so toxic it's unrecognisable'", 6 February 2024, https://fortune.com/2024/02/06/inside-elon-musk-x-twitter-austin-content-moderation/
[202] X, "Defending and respecting the rights of people using our service", https://help.x.com/en/rules-and-policies/defending-and-respecting-our-users-voice (accessed on 3 July 2025).
[203] BBC News, "Twitter pulls out of voluntary EU disinformation code", 27 May 2023, https://www.bbc.co.uk/news/world-europe-65733969
[204] ABC News, "Potential mass layoffs at Twitter could cripple content moderation, some experts say", 23 October 2022, https://abcnews.go.com/Business/potential-mass-layoffs-twitter-cripple-content-moderation-experts/story?id=91856973; NBC News, "A timeline of Elon Musk's takeover of Twitter" (previously cited).
[205] Reuters, "Exclusive: Twitter leans on automation to moderate content as harmful speech surges", 5 December 2022, https://www.reuters.com/technology/twitter-exec-says-moving-fast-moderation-harmful-content-surges-2022-12-03/
[206] Reuters, "Exclusive: Twitter leans on automation to moderate content as harmful speech surges" (previously cited).
[207] Fortune, "Inside the shifting plan at Elon Musk's X to build a new team and police a platform 'so toxic it's unrecognisable'" (previously cited).
[208] Fortune, "Inside the shifting plan at Elon Musk's X to build a new team and police a platform 'so toxic it's unrecognisable'" (previously cited).
[209] Paul M. Barrett, "Who moderates the social media giants? A call to end outsourcing", June 2020, https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf
[210] Fortune, "Inside the shifting plan at Elon Musk's X to build a new team and police a platform 'so toxic it's unrecognisable'" (previously cited).

X's policies on harmful content, including content which may constitute TfGBV, have also shifted since Elon Musk became the owner. For example, in April 2023, X quietly removed its policy against the "targeted misgendering and deadnaming of transgender individuals".[211] This policy was reinstated in 2024.[212]

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 speech that stops short of violating the domestic law of countries in which it operates.[213]

Speaking to Fortune magazine, a source familiar with trust and safety at X highlighted the effect that unclear or rapidly changing policies have on the platform's ability to adequately tackle harmful content: "[T]he number of humans at computers matters less, in some ways, than having clear policies rooted in harm reduction strategies, and the tools and systems necessary to implement these policies at scale".[214]

Critics have also argued that X's misinformation policies and detection technologies have allowed inflammatory posts to be boosted on the platform, in part because so-called "Twitter Blue" accounts belonging to users who have paid for verification now have their posts and accounts amplified by X's algorithm even if these accounts are used to spread false or harmful content.[215]

An independent audit of X's DSA-mandated risk assessment found inconsistencies in how content moderation rules were applied.[216] Accounts with large followings or "verified" status appeared to be treated differently from regular users.[217] The audit recommended that X implement standardized enforcement procedures to ensure fairness and transparency for all users.[218]

5.4.1 COMMUNITY NOTES

Since 2021, X has become increasingly reliant on the platform's Community Notes function, which allows a decentralized network of approved users to add notes with additional context to posts, as a form of content moderation for its millions of active users.[219] The Community Notes function has more than 100,000 contributors across the EU and was made available in Poland in 2023.[220]

Community Notes contributors begin their contributions by rating others' notes and later earn the ability to write notes themselves.[221] Anyone who has an X account in "good standing" with no violations of X's rules and a verified phone number can sign up to be a contributor. Contributors are chosen by X every week in a randomized process.[222]

Contributors rate the helpfulness of notes based on a series of heuristics such as whether it is written in neutral language, contains a high-quality citation or directly refers to the post's claim.[223]

---

[211] NBC News, "Twitter is the 'most dangerous platform for LGBTQ people', GLAAD says" (previously cited).
[212] GLAAD, *Social Media Safety Index* (previously cited).
[213] ABC News, "Potential mass layoffs at Twitter could cripple content moderation, some experts say" (previously cited); NBC News, "Elon Musk, new owner of Twitter, tweets unfounded anti-LGBTQ conspiracy theory about Paul Pelosi attack", 30 October 2022, https://www.nbcnews.com/news/us-news/elon-musk-new-owner-twitter-tweets-unfounded-conspiracy-theory-paul-pe-rcna54717
[214] Fortune, "Inside the shifting plan at Elon Musk's X to build a new team and police a platform 'so toxic it's unrecognisable'" (previously cited).
[215] Foreign Policy, "Elon Musk's Twitter is becoming a sewer of disinformation", 15 July 2023, https://foreignpolicy.com/2023/07/15/elon-musk-twitter-blue-checks-verification-disinformation-propaganda-russia-china-trust-safety/; The Independent, "How Twitter became a far-right misinformation vehicle for the riots – and how it could be saved", 6 August 2024, https://www.independent.co.uk/tech/twitter-x-riots-elon-musk-b2592074.html
[216] FTI Consulting, "X Independent Audit": Article 37 of the Regulation (EU) 2022/2065 (Digital Services Act)", 2024, https://transparency.x.com/content/dam/transparency-twitter/dsa/dsa-audit/TIUC-DSA-Audit-Report-2024-08-27.pdf
[217] FTI Consulting, "X Independent Audit" (previously cited).
[218] FTI Consulting, "X Independent Audit" (previously cited).
[219] Tech Crunch, "Elon Musk takes Twitter out of the EU's Disinformation Code of Practice", 27 May 2023, https://techcrunch.com/2023/05/27/elon-musk-twitter-eu-disinformation-code/; Fortune, "Inside the shifting plan at Elon Musk's X to build a new team and police a platform 'so toxic it's unrecognisable'" (previously cited); House of Commons Science Innovation and Technology Committee, "Oral evidence: social media, misinformation and harmful algorithms", HC 441, 25 February 2025, https://committees.parliament.uk/oralevidence/15413/pdf/
[220] Community Notes, X post, 27 July 2023, https://x.com/communitynotes/status/1684667112869052416; X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, August 2024, https://transparency.x.com/content/dam/transparency-twitter/dsa/dsa-sra/dsa-sra-2024/TIUC-DSA-SRA-Report-2024.pdf
[221] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441(previously cited).
[222] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441(previously cited).
[223] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441(previously cited).

The decision of whether a note is placed on a post is entirely in the hands of the approved users. X employees do not place or remove notes.[224] As of July 2024, users can request a Community Note on a post that they believe would benefit from additional context.[225] Every post and user is subject to and eligible for Community Notes, including advertisers, and content that receives a Community Note is demonetized, meaning that the author will not be able to earn revenue from the post.[226]

Post authors can request a review of a Community Note attached to their content, which is undertaken by the Community Note contributors.[227] If the additional review changes the rating of the Community Note to no longer be helpful, it will be taken down.[228] X does not have oversight of the review process.[229] If a note remains up for two weeks, it is locked onto the post and cannot be removed.[230]

The efficacy of these approaches to countering harmful content on the platform, has been criticized by digital rights groups. There have also been concerns over the quality of some of the contributions, and whether this approach can match the pace at which content is created on X.[231] X claims that independent academic research has shown that accounts or posts that receive Community Notes are 80% more likely to be deleted, and 50% to 61% less likely to be shared.[232] While X has claimed that this helps reduce the possibility that such posts become 'viral', the company has also said that Community Notes do not affect reach,[233] meaning that a post with a Community Note only receives less shares due to user behaviour, rather than algorithmic downranking.

Additionally, the Community Notes feature is slow. A note is placed on a post only when two contributors who have previously disagreed agree that a note would be helpful. X has acknowledged that this presents issues of speed and scale.[234] While X reports that the speed of Community Notes is improving, the median time for a note to appear after a post is created in a crisis situation – such as the 7 October Hamas attacks on Israel – is around five hours; sufficient time for a post to be seen by potentially hundreds of thousands of people before a note is added.[235] A research study conducted in 2021 using data from Birdwatch, X's predecessor to Community Notes, found evidence that users are more likely to write negative evaluations of tweets from people with whom they have political differences, and are more likely to rate evaluations of tweets from people with different political perspectives as unhelpful – suggesting that users preferentially challenge content from those with whom they disagree politically.[236] While not necessarily indicating that this method of fact-checking is ineffective for identifying misleading content, this does suggest that partisanship can play a significant role in content evaluation.[237] The same 2021 research study also found that users who followed accounts with a similar political leaning to a post's author – and who therefore may be more likely to come across the post organically in their news feed – were more likely to be critical (that is, rate as unhelpful) notes that marked the post as misleading.[238] Therefore, people who were more likely to view the Community Note on a certain tweet were the least likely to find it helpful .[239]

Paired with the drastic reduction in resources for content moderation, Community Notes appears to be an attempt by X to shirk its human rights responsibilities, effectively outsourcing the moderation of the platform to its users. This is patently inadequate under human rights standards, as the lack of clear rules and policies around what content is subject to a Community Note means that users cannot adequately predict whether their content will be flagged. Additionally, the Community Notes function does not appear to adequately

[224] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441(previously cited).
[225] 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).
[226] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441(previously cited).
[227] X, Community Notes, https://communitynotes.x.com/guide/en/contributing/additional-review
[228] X, Community Notes (previously cited)
[229] X, Community Notes (previously cited)
[230] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441(previously cited).
[231] Tech Crunch, "Elon Musk takes Twitter out of the EU's Disinformation Code of Practice" (previously cited).
[232] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441(previously cited).
[233] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441(previously cited).
[234] House of Commons Science Innovation and Technology Committee, "Oral evidence: Social media, misinformation and harmful algorithms", HC 441(previously cited).
[235] 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).
[236] Jennifer Allan and others, "Birds of a feather don't fact-check each other: partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program", CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, Article No. 245, https://dl.acm.org/doi/pdf/10.1145/3491102.3502040
[237] Jennifer Allan and others, "Birds of a feather don't fact-check each other" (previously cited).
[238] Jennifer Allan and others, "Birds of a feather don't fact-check each other" (previously cited).
[239] Jennifer Allan and others, "Birds of a feather don't fact-check each other" (previously cited).

consider the issues of legality, necessity and legitimacy which, under human rights standards such as Article 19 of the ICCPR and the GNI guidance, must be taken into account when limiting freedom of expression, since all that is necessary for a post to receive a note is for two contributors who have previously disagreed to agree. Furthermore, the Community Notes function does not provide any particular protection for the rights of marginalized groups.

Community Notes effectively allows content on the platform to be moderated based on user opinions, rather than robust, clear and carefully considered policies developed in alignment with human rights standards. Furthermore, Community Notes contributors do not receive any training in fact-checking, and the Notes are not quality controlled in any way by X. It is therefore of significant concern that the Community Notes feature is presented as a key risk mitigation measure in X's most recent DSA risk assessment.[240]

5.5 X'S CHANGING APPROACH TO HATEFUL CONTENT

At the beginning of Elon Musk's tenure at X, accounts which were previously banned for breaking the platform's policies were reinstated, such as those of US Congress Representative Marjorie Taylor Green (banned for violating Covid-19 misinformation policies), satirical news site The Babylon Bee (banned for posting a transphobic story violating the hateful conduct policy) and social media personality Andrew Tate (banned for saying that women should "bear some responsibility" for being sexually assaulted).[241]

Civil society organizations focusing on LGBTI rights, such as US-based GLAAD, have also noted a change in the way X responds to posts that violate the platform rules.[242] Previously, offending posts were removed; however, at the time of writing, posts violating platform rules are sometimes only 'restricted' from amplification, rather than being removed from the platform altogether.[243]

In October 2023, X made several changes to its Community Guidelines policy, which included a significant softening of its Violent Content policy.[244] Violent content is defined as content containing "violent speech" or "violent media".[245] Violent speech is defined in the policy as content that threatens, incites, glorifies or expresses desire for violence or harm, whereas violent media is visual material depicting graphic, violent or excessively gory content, including sexual violence.[246]

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".[247] 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".[248] Additionally, the Violent Content policy does not contain any mention of the risks that such content poses to marginalized communities, including risks associated with TfGBV.[249] However, X's Hateful Conduct policy outlines that it is prohibited to attack people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability or serious disease.[250]

In 2023, researchers at the Center for Countering Digital Hate reported 300 posts to X for hate speech. One week later, X continued to host 86% of these posts on the platform.[251] Researchers received notifications from X that three of the accounts which had posted hateful content were "locked", stating that "they can't post, repost or Like content, and we'll ask them to remove the reported content if they want to regain full access to their account". Nevertheless, the reported posts remained visible on the platform.[252]

---

[240] 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).
[241] Fortune, "Inside the shifting plan at Elon Musk's X to build a new team and police a platform 'so toxic it's unrecognisable'" (previously cited).
[242] NBC News, "Twitter is the 'most dangerous platform for LGBTQ people', GLAAD says" (previously cited).
[243] NBC News, "Twitter is the 'most dangerous platform for LGBTQ people', GLAAD says" (previously cited).
[244] Lab Platform Governance, Media and Technology (PGMT), "X (formerly Twitter), softens its violent speech policy", 2 November 2023, https://platform-governance.org/2023/x-formerly-twitter-softens-its-violent-speech-policy/
[245] X, "Violent Content policy", February 2025, https://help.x.com/en/rules-and-policies/violent-content
[246] X, "Violent Content policy" (previously cited).
[247] PGMT, "X (formerly Twitter) softens its violent speech policy" (previously cited). (Emphasis added).
[248] X, "Violent Content policy" (previously cited). (Emphasis added).
[249] X, "Violent Content policy" (previously cited); The Verge, "Twitter rewrites its rules on violent content under Elon Musk", 28 February 2023, https://www.theverge.com/2023/2/28/23619262/twitter-violent-speech-policy-zero-tolerance
[250] X, "Hateful Conduct policy", April 2023, https://help.x.com/en/rules-and-policies/hateful-conduct-policy
[251] Center for Countering Digital Hate, "X content moderation failure: how Twitter/X continues to host posts reported for extreme hate speech", September 2023, https://counterhate.com/wp-content/uploads/2023/09/230907-X-Content-Moderation-Report_final_CCDH.pdf
[252] Center for Countering Digital Hate, "X content moderation failure: how Twitter/X continues to host posts reported for extreme hate speech" (previously cited), p. 6.

Under the DSA, VLOPs such as X must submit transparency reports describing their content moderation activities. At the time of writing, X was the most inconsistent platform in its reporting, based on an academic research audit of the DSA Transparency Database, which covered all records submitted by the eight largest social media platforms in the EU during the first 100 days of the database's operation.[253]

In X's most recent DSA transparency report, the company states that its content moderation policies "are designed and tailored to mitigate systematic risks without unnecessarily restricting the use of our service and fundamental rights, especially freedom of expression. Content moderation activities are implemented and anchored on principled policies and leverage a diverse set of interventions to ensure that our actions are reasonable, proportionate and effective. Our content moderation systems blend automated and human review paired with a robust appeals system that enables our users to quickly raise potential moderation anomalies or mistakes."[254]

The platform also acknowledges that "[v]iolence, harassment, and other similar types of behaviour discourage people from expressing themselves, and ultimately diminish the value of global public conversation."[255]

X also outlined the factors that influence enforcement decisions on the platform, including: if the behaviour is directed at an individual, group or protected category of people, whether the report has been filed by the target of the abuse or a bystander, if the reported user has a history of posting violating content, the severity of the violation, and if the subject of the reported post may be a topic of legitimate public interest.[256]

The DSA requires that the transparency reports of all large social media platforms include data on the human resources dedicated to content moderation, broken down into each of the EU's official languages. The first tranche of reports, released in November 2023, revealed that X had just one Polish-speaking content moderator, while only 8% of the platform's content moderators were proficient in an official EU language other than English.[257] In the most recent DSA transparency report at the time of writing (from April 2025), X had only two Polish-language content moderators, one of whom spoke Polish as a second language.[258]

X disclosed in its transparency report that, in situations where additional language support is needed, the company uses machine translation tools.[259]

Jakub Szymik, an LGBTI activist, told Amnesty International that he believes the lack of Polish-speaking content moderators has led to a sluggish response to harmful content reported on the platform:

"I did it [reported content] a few times, but I did it when I saw some content about someone I knew personally, and nothing was taken down. I received notices that the content will stay there. I don't think there's any point wasting time on this. There is one Polish-speaking content moderator for 8 million users in Poland… I have no trust in this platform… I don't report anymore."[260]

Similarly, Mateusz Kaczmarek of Grupa Stonewall didn't notice any changes as a result of his reporting:

"Sometimes when I am using Twitter, I report some tweets. And I think most of my reports don't change anything so the tweets were left there, and nothing happened."[261]

Maja Heban, a trans woman and activist, also shared with Amnesty International that she no longer reports content to X due to the poor response from the platform:

"I stopped reporting a year ago. I tried at the beginning, when Musk took over. But quickly everybody realized that it's just not going to work anymore. However, even before him, it was still unhelpful… there were times when I was reporting really viral stuff and it was just found to be OK, according to Twitter… The most vile stuff you can come up with and it was ok. So very quickly it became apparent that maybe if someone posts CSAM [child sexual abuse material], maybe this will be taken down, but apart from that you can say anything, and it will stay on."[262]

---

[253] Amaury Trujillo and others, "The DSA Transparency Database: auditing self-reported moderation actions by social media", 1 August 2024, Proceedings of the ACM on Human-Computer Interaction, Volume 9, Issue 2, https://dl.acm.org/doi/10.1145/3711085
[254] X, *DSA Transparency Report – October 2024*, https://transparency.x.com/dsa-transparency-report.html
[255] X, *DSA Transparency Report – October 2024* (previously cited).
[256] X, *DSA* Transparency Report – October 2024 (previously cited).
[257] Global Witness, "How big tech platforms are neglecting their non-English language users", 30 November 2023, https://www.globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/
[258] X, *DSA Transparency Report – April 2025*, https://transparency.x.com/dsa-transparency-report-2025-april.html
[259] X, *DSA Transparency Report – October 2024* (previously cited).
[260] Amnesty International video call with Jakub Szymik, 5 August 2024.
[261] Amnesty International interview with Maetusz Kaczmarek, 28 July 2024.
[262] Amnesty International interview with Maja Heban, 20 July 2024.

Psychologist Julia Kata explained to Amnesty International that the comments on posts were also difficult to report:

"You will receive a lot of comments. It's sometimes easier to just take down the content, instead of [accepting] the comments. It's very difficult to report that [the comments]. They don't care. They don't."[263]

Piotr Pjotrowicz, a gay man living in the city of Sosnowiec who uses X for activism, told Amnesty International that he believes X's poor moderation is indicative of a broader pattern of technology platforms neglecting the needs of non-English speaking users:

"I think it's a global issue. I don't think it's a Polish issue. I think they don't care in multiple countries, and they don't have enough people to deal with what is happening."[264]

While X claims that each content moderator goes through extensive training and refresher courses, it is unclear whether this includes specific training on the rights of LGBTI people. GLAAD noted that, of all the companies it evaluated in its Social Media Safety Index, X was the only platform that did not disclose any information on whether it has training in place to educate content moderators on the needs of LGBTI users.[265] Amnesty International wrote to X on 22 August 2024 requesting information on what type and level of LGBTI rights training is available to the platform's content moderators, but did not receive a reply.[266]

5.6 RELUCTANCE TO COMPLY WITH EU RULES AND STANDARDS

In 2018, previous management at X (then known as Twitter) signed the platform up to the voluntary EU Code on Disinformation.[267] The Code committed X to taking steps to combat the spread of false information on its service by targeting associated advertising revenue, tackling bots and fake accounts, providing consumers with the tools to report disinformation and empowering researchers to study the platform.[268]

In June 2022 the European Commission unveiled an improved version of the EU Code on Disinformation and announced the establishment of a transparency centre to monitor adherence to it.[269] The European Commission also announced that adhering to the Code would be one aspect of a company's DSA compliance.[270] In May 2023, X pulled out of the voluntary Code, sparking a row with the European Commission, which took the view that X had chosen "confrontation" with the move.[271] At the time, Elon Musk maintained that there was "less misinformation rather than more" since he acquired the platform.[272]

X has already been subject to the first-ever investigation under the DSA.[273] In December 2023 the European Commission opened infringement proceedings against X after it was subject to repeated claims that it was not doing enough to curb the spread of disinformation and hate speech online.[274] Four investigations were launched focusing on X's failure to comply with EU rules to counter illegal content and disinformation as well as rules on transparency around advertising and data access for researchers.[275] In July 2024 the European Commission informed X of its preliminary view that the platform is in breach of the DSA in areas linked to "dark patterns" (deceptive techniques used by online platforms to manipulate users behaviour, often

[263] Amnesty International interview with Julia Kata, 29 July 2024.
[264] Amnesty International video call with Piotr Pjotrowicz, 7 August 2024.
[265] GLAAD, *Social Media Safety Index* (previously cited); X, *DSA Transparency Report – October 2024* (previously cited).
[266] Amnesty International letter to X, 22 August 2024.
[267] Tech Crunch, "Elon Musk takes Twitter out of the EU's Disinformation Code of Practice" (previously cited).
[268] Tech Crunch, "Elon Musk takes Twitter out of the EU's Disinformation Code of Practice" (previously cited).
[269] Tech Crunch, "Elon Musk takes Twitter out of the EU's Disinformation Code of Practice" (previously cited).
[270] Tech Crunch, "Elon Musk takes Twitter out of the EU's Disinformation Code of Practice" (previously cited).
[271] BBC News, "Twitter pulls out of voluntary EU disinformation code" (previously cited); Tech Crunch, "Elon Musk takes Twitter out of the EU's Disinformation Code of Practice" (previously cited); Euronews, "Twitter has chosen 'confrontation' with Brussels over disinformation code of conduct", 5 June 2023, https://www.euronews.com/my-europe/2023/06/05/twitter-has-chosen-confrontation-with-brussels-over-disinformation-code-of-conduct
[272] BBC News, "Twitter pulls out of voluntary EU disinformation code" (previously cited).
[273] Politico, "X vs. EU: Elon Musk hit with probe over spread of toxic content", 18 December 2023, https://www.politico.eu/article/elon-musks-x-probed-by-eu-for-failing-to-curb-toxic-content/
[274] Politico, "X vs. EU: Elon Musk hit with probe over spread of toxic content" (previously cited).
[275] Politico, "X vs. EU: Elon Musk hit with probe over spread of toxic content" (previously cited).

without their knowledge or consent), advertising transparency, and data access for researchers.[276] Elon Musk has reportedly frequently referred to the DSA as a "censorship tool".[277]

In April 2025 it was reported that the European Commission was considering issuing a large fine against X for being in violation of the DSA, as well as requesting changes to the way X functions.[278] It is still possible that the EU and X could reach a settlement if the company agrees to changes that satisfy the regulator's concerns.[279]

X responded to media reports of the impending fine by posting that enforcement actions against it would be "an unprecedented act of political censorship and an attack on free speech".[280]

Dorota Głowacka, an advocacy and litigation expert at Panoptykon Foundation, a Polish digital rights NGO, explained to Amnesty International the importance of robust DSA enforcement:

"We feel there is no other effective way, like we tried other measures before – self regulation, codes of conduct, all sorts of measures, you know. And it just didn't seem to work or bring any meaningful changes."[281]

5.7 SYSTEMIC ISSUES

While there has been a large amount of scrutiny of X since Elon Musk's 2022 takeover, it is important to note that, even before the company changed hands, X (then known as Twitter) was often criticized for its failure to properly police potentially harmful content, including content which may constitute TfGBV.[282]

In 2018, Amnesty International found X to be failing to respect women's rights online by not properly mitigating online abuse. Women of colour, women from ethnic or religious minorities, lesbian, bisexual and transgender women, non-binary individuals and women with disabilities were found to be exposed to the most abuse on the platform.[283] The research demonstrated that, even in 2018, the volume of harmful content on X was perceived to be higher than on other platforms. Jessical Valenti, a US journalist and writer interviewed for the research, told Amnesty International: "The content feels pretty similar across the platforms but the sheer volume of it on Twitter is what's different."[284]

Also in 2018, Amnesty International conducted research on X to understand the extent of abusive content on the platform targeted at women politicians and journalists.[285] At the time the research was conducted, one in 10 tweets mentioning Black women politicians and journalists based in the UK and USA in a sample analysed by Amnesty International was found to be abusive or problematic.[286] These findings led Amnesty International to describe X as "a place where racism, misogyny and homophobia are allowed to flourish basically unchecked".[287]

A lack of transparency was also an issue in X's operations long before Elon Musk's takeover. In 2018, Amnesty International repeatedly asked X to publish data regarding the scale and nature of abuse on the platform, but at the time the company failed to do so.[288] Amnesty International subsequently recommended that "Twitter must start being transparent about how exactly they are using machine learning to detect abuse and publish technical information about the algorithms they rely on".[289]

---

[276] European Commission, "Commission send preliminary findings to X for breach of Digital Services Act", 12 July 2024, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761
[277] Tech Policy Press, "Understanding the EU's Digital Services Act Enforcement Against X", 5 April 2025, https://www.techpolicy.press/understanding-the-eus-digital-services-act-enforcement-against-x/
[278] New York Times, "E.U. prepares major penalties against Elon Musk's X", 3 April 2025, https://www.nytimes.com/2025/04/03/technology/eu-penalties-x-elon-musk.html
[279] New York Times, "E.U. prepares major penalties against Elon Musk's X" (previously cited).
[280] Global Government Affairs, post on X, 4 April 2025, https://x.com/GlobalAffairs/status/1907963263419297893
[281] Amnesty International interview with Dorota Głowacka, 29 July 2024.
[282] The Independent, "How Twitter became a far-right misinformation for the riots – and how it could be saved" (previously cited).
[283] Amnesty International, "Toxic Twitter – a toxic place for women", 21 March 2018, https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1-1/
[284] Amnesty International, "Toxic Twitter – a toxic place for women" (previously cited).
[285] Amnesty International, "Crowdsourced Twitter study reveals shocking scale of online abuse against women", 18 December 2018, https://www.amnesty.org/en/latest/press-release/2018/12/crowdsourced-twitter-study-reveals-shocking-scale-of-online-abuse-against-women/
[286] Amnesty International, "Crowdsourced Twitter study reveals shocking scale of online abuse against women" (previously cited).
[287] Amnesty International, "Crowdsourced Twitter study reveals shocking scale of online abuse against women" (previously cited).
[288] Amnesty International, "Crowdsourced Twitter study reveals shocking scale of online abuse against women" (previously cited).
[289] Amnesty International, "Crowdsourced Twitter study reveals shocking scale of online abuse against women" (previously cited).

In follow-up research published in 2020, Amnesty International found that, while X had made some progress on addressing the issue of online violence and abuse against women since 2018, the company continued to fall short of its human rights responsibilities and needed to do more to protect women's rights online.[290]

5.7.1 PREVALENCE OF TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE ON X IN 2025

TfGBV on X is a global problem. A 2023 academic study of homophobia and transphobia on the platform found that there was widespread use of anti-LGBTI language in all seven languages examined.[291] The study concluded that, despite the use of automated hate speech detection systems on X, there remained a significant amount of anti-LGBTI content on the platform.[292]

In Amnesty International's quantitative study conducted on X in Poland in 2025, it became apparent that abusive content, including content constituting TfGBV, remains a significant issue on the platform. When analysing the sample of 1,387 tweets, the findings suggest that homophobic and transphobic content is highly prevalent on X, particularly for accounts that follow politicians that do not support LGBTI rights. Amnesty International found that almost 4% of tweets categorized using our methodology from such accounts (which followed anti-LGBTI politicians)were homophobic or transphobic and, that more than 25% of all the LGBTI related tweets these accounts see were homophobic or transphobic (See Figure 2 below). By comparison, only 1.2% of tweets which were categorized from research accounts in the 'Full Rights' group (which followed politicians supportive of the rights of LGBTI people) were considered homophobic or transphobic, constituting 11% of all LGBTI-related content they saw

FIGURE 2

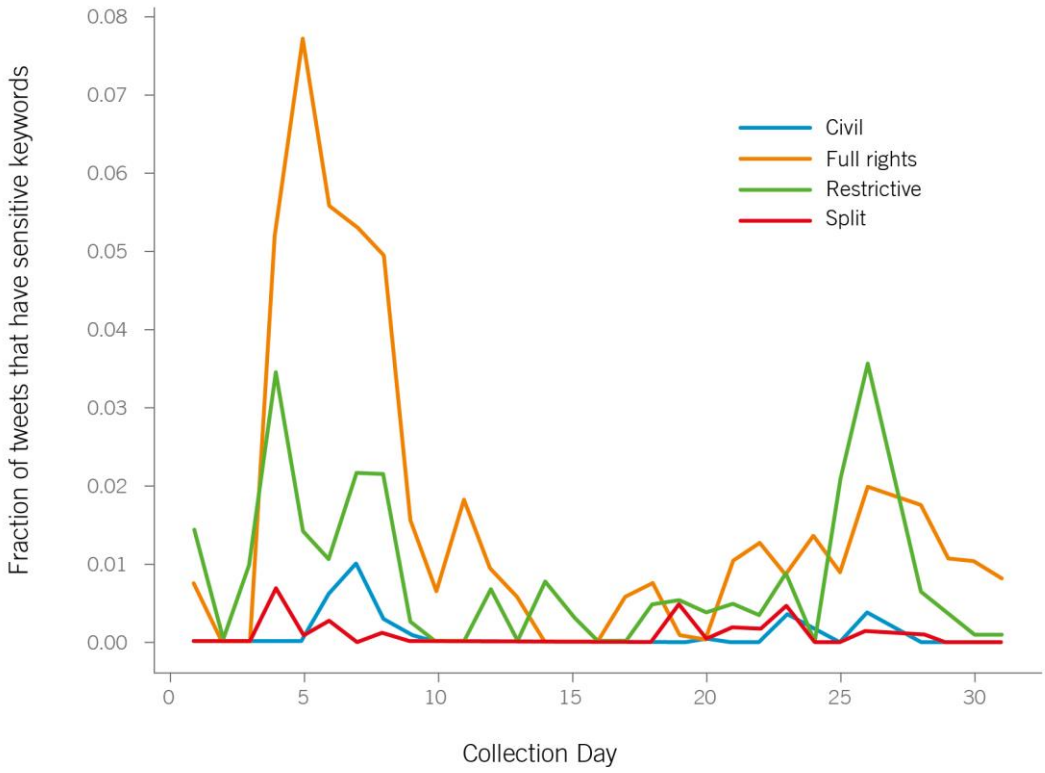
GROUP	% of all content categorized as homophobic or transphobic	% of LGBTI-related content categorized as homophobic or transphobic
CIVIL	0	0
FULL RIGHTS	1.2	11
RESTRICTIVE	3.7	28
SPLIT	0.5	23

Amnesty International first analysed the percentage of all collected tweets for their relevance to LGBTI issues. Using a list of keywords,[293] Amnesty International found around 0.6% of all tweets were relevant to LGBTI issues across all 32 research accounts. While only constituting one in every 1700 pieces of content, it is important to note that this figure is an average across all content collected from both the "For You" algorithmically determined timeline, and the reverse chronological "Following" timeline. Since the research accounts only followed politicians, and did not engage with any content on the platform, this suggests LGBTI issues remain a substantive topic of discussion on X in Poland, with the average user who engages with politicians regularly encountering LGBTI-related content each time they log on to X.

When examining by sub-group, research accounts in the 'Full Rights' group (those following politicians that support equal rights for LGBTI people) were on average shown double the volume of LGBTI-related content (around 1.6%) compared to the three other sub-groups (around 0.8%). This means they were most exposed to LGBTI-related content, as shown in Figure 3 below. This was most notable between days five and 10 of the study, which aligned with the Polish government passing a new regulation on hate speech.

[290] Amnesty International, *Twitter's Scorecard: Tracking Twitter's Progress in Addressing Violence and Abuse Against Women Online,* (Index: AMR 51/2993/2020), 22 September 2020, https://www.amnesty.org/en/documents/amr51/2993/2020/en/
[291] Davide Locatelli and others, "A cross-lingual study of homotransphobia on Twitter", 2023, Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), ACL Anthology, https://aclanthology.org/2023.c3nlp-1.3/
[292] Davide Locatelli and others, "A cross-lingual study of homotransphobia on Twitter" (previously cited).
[293] The terms on the list were: LGBT, Queer, gender, homo, gay, lesbian, same-sex, lesbij, equality parade, trans, lesb, lezb, plciow, plec, plci, hetero, cispl, pedal, ciot, niebnarn, interseks, fembo.

FIGURE 3: FRACTION OF TWEETS THAT HAVE SENSITIVE KEYWORDS



Analysis of the comments collected by Amnesty International's research accounts under posts related to LGBTI issues found that 26.8% of the responses were homophobic or transphobic. Notably, Amnesty International's analysis showed that this figure was particularly pronounced in comments and replies on tweets presented to the 'Full Rights' sub-group. This suggests that pro-LGBTI rights-related content received substantially more homophobic and transphobic replies than other tweets in the sample.

5.7.2 ALGORITHMIC AMPLIFICATION AND THE CHALLENGES OF MEASUREMENT

As outlined in the Methodology section, Amnesty International's quantitative research attempted to study the extent to which, if at all, X was amplifying anti-LGBTI content. This could be done because X's interface provides a clear comparison between the "For You" and "Following" feeds.

As detailed in the Methodology section above, the quantitative research was conducted simply by creating research accounts which followed the X accounts of Polish politicians. The research accounts did not interact with any of the content to avoid potentially amplifying harmful content. Therefore, the study was unable to measure the effect that engagement with content would have on algorithmic curation or amplification.[294] Within the constraints of this quantitative research, we found no evidence of algorithmic amplification of either LGBTI-related content or anti-LGBTI content, based simply on following accounts.

The algorithmic "For You" timeline presents fewer LGBTI-related tweets than the "Following" timeline. This suggests the politicians followed by research accounts, regardless of their political partisanship or stance on

[294] Algorithmic curation refers to the automated process of selecting and organizing content online using algorithms. Algorithmic amplification is the way in which algorithms tend to amplify certain types of content and suppress others, often based on the inferred interests of platform users.

social issues, posted more about LGBTI issues than was algorithmically recommended to the research accounts, and therefore not implying any noticeable amplification of the content. Additionally, during the study period, Poland's Supreme Court issued a landmark ruling eliminating the requirement for trans people to involve their parents in gender recognition proceedings, which may have increased the number of posts by political figures on the rights of LGBTI people.

Figure 4 below details the results of the experiment looking at the "Following" and "For You" timelines. It does not present the results by sub-group but the findings are consistent across all four groups.

FIGURE 4: PERCENTAGE OF LGBTI-RELATED CONTENT ON THE "FOLLOWING" VERSUS "FOR YOU" TIMELINES

TIMELINE	% of anti-LGBTI content	% of content categorized as homophobic or transphobic	% of pro-LGBTI content
FOLLOWING	3.7	2.3	2.5
FOR YOU	0.6	0.5	1.8

While the exact weightings and inner workings of X's recommender system are opaque, the weightings disclosed in X's publicly available recommender code (discussed in more detail in Chapter 7) states that the platform heavily prioritizes content in users' "For You" feed which the recommender system predicts they will engage with, particularly in terms of replies. Given that the research accounts did not interact or engage with any posts, Amnesty International's working hypothesis is that simply following accounts with particular political stances or social attitudes is not sufficient to lead users into rabbit holes of content, nor to measure amplification.

It is likely that the recommender system would require accounts to interact, share, like or comment on specific pieces of content for them to bear significant weighting in what it algorithmically recommends to accounts. Given the ethical risks of amplifying harmful content, Amnesty International researchers chose not to do this within this study. However, this does present a challenge for future research into algorithmic amplification on X.

Even within the limitations of the quantitative experiment, Amnesty International researchers were able to find evidence that content related to LGBTI issues remains prevalent on X, a high amount of the content related to LGBTI issues contains homophobic and transphobic content (whether in posts or in replies to posts) and that users who are following politicians who support the rights of LGBTI people are most exposed to these replies. The prevalence of the content has significant adverse impacts on the human rights of LGBTI people, who fall in this category, which will be explored in further detail in the following chapter.

6. ON ALERT: THE EFFECTS OF TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE ON LGBTI INDIVIDUALS

> **"It's hard for me to talk about this because my whole life revolves around what is happening, so I always have to be on alert."[295]**

As outlined above, X has played a key role in normalizing dehumanizing rhetoric against the LGBTI community in Poland and the resultant "top-down polarization" which lingers in Polish society despite the change in government. This chapter outlines five cases of LGBTI individuals who experienced TfGBV – including threats of violence, online harassment, doxing (the sharing of private or identifying information about a particular individual on the internet, with malicious intent) and targeted online hate on X, and the effect this has had on their ability to freely express themselves, to live free from discrimination and to feel safe in Polish society. The cases are not exhaustive of the issues faced by LGBTI people in relation to TfGBV in Poland, but provide illustrative examples of the nature of X's role in normalizing anti-LGBTI sentiment in Poland and its various impacts.

[295] Amnesty International video call with Ali (pseudonym), 26 July 2024.

6.1 ALEKSANDRA HERZYK'S STORY

Aleksandra Herzyk is an asexual woman living in Krakow, the second largest city in Poland. She is a comic book artist and author, who uses social media to showcase her work and to raise awareness of human rights and social justice issues. After experiencing TfGBV on X, she no longer uses the platform, logging out permanently in early 2024.

Aleksandra spoke about her asexuality on X, hoping to provide guidance to younger people who may be struggling with their asexuality. She told Amnesty International that the responses she received to her posts about asexuality on X were "very, very aggressive".[296] However, the post which triggered an 18-month-long campaign of targeted hate directed at Aleksandra between 2023 and 2024 related to her experience of having breast reduction surgery. Aleksandra shared with Amnesty International screenshots with examples of this content.

Writing about her surgery led some people on X to perceive Aleksandra as a trans woman. As a result, she was targeted with transphobic hate. Aleksandra told Amnesty International that, while she initially found the comments humorous, it soon became clear that the hate was a serious issue:

"On the one hand it was funny, but I knew that I was taking the hate that was [meant] for trans people. And sometimes people wished very bad things for me. There were people saying that if they saw me in the gym or something, then they will break my bones, that they wished someone would kill me."[297]

Aleksandra reported some extreme examples of hate she received on X – such as comments suggesting that trans people should kill themselves – with inconsistent results. Some posts were taken down, while others with similar harmful calls for violence remained on X.

Aleksandra told Amnesty International that she only realized the effect that experiencing TfGBV had had on her well-being after she permanently left X:

"I was constantly prepared to be witty, sarcastic, to respond in a way that was very unpleasant. I thought I was having fun and in some sense I was. But when I logged out of Twitter, I thought that it wasn't worth it. It wasn't worth the time. 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."[298]

6.2 ALI'S STORY

Ali is a 24-year-old non-binary person living in Warsaw, who uses he/him pronouns. He is an abortion activist and has also used social media to post about his life as a non-binary person in Poland. He also buys chest-binding bandages, known as 'binders', for transgender youth and works with a collective which provides the emergency contraceptive pill. Although Ali did not use X for his activism, most of the harmful content targeting him came from the platform.

Ali described how experiencing TfGBV affected his sense of safety:

"If there is more hate speech online, then people are more open to it. And they usually have more courage to apply it."[299]

Ali noted that the online targeting of the LGBTI community in general became more intense in 2019:

"I think a lot of homophobic acts started in 2019 because of the presidential campaign. A lot of politicians were saying publicly or online that we are not humans, and we are some sort of ideology. Not really humans... We were dehumanized. But also, if it comes from the authorities, it creates an atmosphere that [people think] we can do this. This is allowed. This is actually sometimes even appreciated."[300]

Generally, Ali found that he was targeted after speaking out publicly on LGBTI issues.[301] He told Amnesty International how, in 2020, he became a target of TfGBV when he spoke out during a protest at the

[296] Amnesty International interview with Aleksandra Herzyk, 27 July 2024.
[297] Amnesty International interview with Aleksandra Herzyk, 27 July 2024.
[298] Amnesty International interview with Aleksandra Herzyk, 27 July 2024.
[299] Amnesty International interview with Ali (pseudonym), 26 July 2024.
[300] Amnesty International interview with Ali (pseudonym), 26 July 2024.
[301] Amnesty International interview with Ali (pseudonym), 26 July 2024.

university at which he was studying. The protest was sparked by the university making it more difficult for students to use their chosen names and pronouns.

"I received a lot of hate. They [the online abusers] also found me on Facebook, on Instagram, and they were writing to me. I saw on Twitter they were talking to each other in tweets, saying 'okay, this is her [sic] profile. Okay we should write something'. And they were writing that when I get to the university and they see me, they are going to rape me, they are going to beat me, they are going to kill me, that I should resign from my studies because nobody wants this freak here and it was all in my private messages."[302]

At one point, Ali was doxed, including by having his place of work posted on X:

"Some people came to my work, shouted at me and were being really aggressive."[303]

During the doxing, information about his family was also shared on X:

"They found my mother's information and her Facebook account. I wasn't out [to my family] and I was a student, I was financially dependent on them. These people wrote to my parents and said, 'do you know what your daughter [sic] is doing online?' And sent them screenshots of information I was posting about myself online."[304]

Ali explained to Amnesty International that because he had left X at the time of the attacks, he was reliant on friends to let him know about threats being made against him on the platform:

"They were just sending me stuff and saying 'hey Ali, I don't know if you want to see this but there's this post on Twitter and they are talking about how they want to harm you, so I think that you should know'... I had a Twitter account for like two months but the amount of hate there is just so overwhelming."[305]

Ali felt he had no choice but to engage with the hateful content:

"I felt like – if I don't read what they are saying, I won't be prepared for their next move. So I felt like I had to expose myself to that [content]."[306]

Ali noted that his friends reported the content which contained his personal information to X, but the posts were not taken down.

He described the lingering sense of unease that the experience has left him with:

"I have to always be on alert."[307]

6.3 MAGDA DROPEK'S STORY

Magda Dropek is a 42-year-old LGBTI activist who recently moved to Warsaw from Krakow. She has participated in LGBTI activist spaces since 2011. In 2020, Magda also began to engage more in the abortion rights movement in Poland. She co-organized strikes in Krakow and began to do more activism around human rights, taking an intersectional approach. In February 2024, Magda took up a position in the Department for Equal Treatment in the Prime Minister's office.

Magda explained to Amnesty International that she uses X primarily because it is an important tool for her activism:

"I'm doing queer activism and political activism on Twitter mostly. I know how important a tool it is when it comes to communication and activism... and also to speak about what you are doing and why and engage people, because community building and community organizing was always important to me when it came to what I was doing locally."[308]

At the same time, Magda told Amnesty International that X was a place where the LGBTI community faced a lot of hate:

---

[302] Amnesty International interview with Ali (pseudonym), 26 July 2024.
[303] Amnesty International interview with Ali (pseudonym), 26 July 2024.
[304] Amnesty International interview with Ali (pseudonym), 26 July 2024.
[305] Amnesty International interview with Ali (pseudonym), 26 July 2024.
[306] Amnesty International interview with Ali (pseudonym), 26 July 2024.
[307] Amnesty International interview with Ali (pseudonym), 26 July 2024.
[308] Amnesty International interview with Magda Dropek, 24 July 2024.

"Earlier it was more like public, pure hate when it comes against queer people. You'd be reading all the worst things about yourself because you are a queer person or an LGBTI activist. But in the last two years, it's getting more personal. So, it's getting more about, how do you look, how do you present yourself? How do you speak? So, it's more about trying to humiliate you."[309]

Magda also described to Amnesty International that the problem on X was not only the anti-LGBTI content, but the targeting of LGBTI activists speaking out on the platform:

"In my experience, if I [or others] write about something queer... very often, what I see is there is a tweet that has 30, 40 likes or five comments, and then after a day or a few hours, there's like 500 [anti-LGBTI] comments."[310]

Magda told Amnesty International about her experience of posting a tweet supporting the Black Lives Matter and Women's Strike movements, which resulted in thousands of hateful comments and even threatening phone calls:

"I felt the threat in the real world. Especially because of the calls. But I think it made me more aware of what I'm saying, writing and why. So it happened that I caught myself thinking 'maybe I shouldn't write it or maybe I shouldn't say it', because it is something that can be used against me and become another flow of hate."[311]

Magda explained that she feels that the persistence of TfGBV on X prevents people in the LGBTI community fully enjoying their right to freedom of expression:

"It's mostly about silencing and to show that this is not your place, your place is to be silent, not to be visible. Not to be active, engaged, or doing things when it comes to activism especially."[312]

6.4 MAJA HEBAN'S STORY

Maja Heban is a 34-year-old trans woman living in Warsaw. She has been living openly as a trans woman for the past 15 years. She started engaging in LGBTI activism in 2019, when she took part in a social media protest against homophobia during the presidential election, in which people used the hashtag #IMLGBT to come out in large numbers. Maja publishes commentary on social media and transgender rights. X is one of the main platforms she uses for her work.

Maja noted that the level of hate on X was extremely high for some years:

"I think even before Elon Musk took over, I guess moderation and reporting barely worked... I remember that many times I would make a fuss about reporting something really, really, extremely negative, extremely hateful. And not getting proper action on this, so not being able to take down the comments. Not even [taking down] the users, but specific tweets about how I should kill myself."[313]

Maja explained to Amnesty International that X is an important political platform in Poland and as a result many activists including herself continue to use it, despite experiencing TfGBV.

Maja believes that X has created an environment in which anti-LGBTI sentiment is increasingly permitted. She described some of the content that targets the LGBTI community:

"Whatever you can say, it's okay – comparing LGBT people to animals, to rapists, to paedophiles. Anything goes."[314]

Maja reflected that, to some extent, the volume of hate she has received has desensitized her:

"Because I am so out, I basically treat hate, getting hate speech, being misgendered, being deadnamed as just part of life. And from time to time I stop and think, well, maybe some people don't live like this, you know? Like maybe they are open about being trans, but at the same time they don't expect to hear that [hate] every day. They don't really expect strangers to tell them to kill themselves every day. Maybe I shouldn't be used to this."[315]

---

[309] Amnesty International interview with Magda Dropek, 24 July 2024.
[310] Amnesty International interview with Magda Dropek, 24 July 2024.
[311] Amnesty International interview with Magda Dropek, 24 July 2024.
[312] Amnesty International interview with Magda Dropek, 24 July 2024.
[313] Amnesty International interview with Maja Heban, 30 July 2024.
[314] Amnesty International interview with Maja Heban, 30 July 2024.
[315] Amnesty International interview with Maja Heban, 30 July 2024.

Maja described a particularly egregious example of transphobia on X to Amnesty International, when in 2022, photos from a trans support group on Facebook were posted to X by an anti-trans activist:

"He infiltrated and downloaded pictures of mastectomy results from a young trans man who was posing with his scar tissue from the mastectomy. He anonymized the photos, but he basically posted them on to Twitter. He said that 'this is a 14-year-old girl who had her breasts hacked off and was mutilated'... It had hundreds of thousands of views."[316]

6.5 NATHAN BRYZA'S STORY

Nathan Bryza is a 21-year-old nonbinary, trans person who uses male pronouns living in the city of Wrocław. He works for a transport company.

Nathan described the way that TfGBV on X had affected his sense of safety:

"Twitter is a specific place that is just a small circle of hell. Lately as the Olympics started, it's even worse because there is a lot of hate towards transgender people even though a non-trans woman is competing... There's rising panic that makes me scared people will see I'm trans and attack me on the street."[317]

Nathan described the type of anti-LGBTI hate he regularly sees on X:

"[They say we are] twisted people, that we are broken, that we are sick. I think that sick is the most used word. We are sinners. That we are lustful. That we are trying to make ourselves special."[318]

TfGBV causes Nathan to feel unsafe in the offline world:

"I've changed my workplace recently and every day when I come to work, I think 'when will be the day people will know I'm queer?' When they look at me, I think they suspect but I don't speak openly about it and I'm scared that someday I will come to work, and I will hear the things I see online about myself and the people I know. And it's scary because it makes me feel really unsafe."[319]

Nathan's experience of feeling unable to fully express his identity at work is commonplace among the LGBTI community in Poland. According to one academic research study, around 35% of LGBTI employees felt it was necessary to hide their identity at work, for fear of discrimination.[320]

Nathan explained the emotional toll this has taken on him:

"I feel really sad about it. I feel anger that, no matter how much we speak for ourselves, no one believes us. It upsets me."[321]

Nathan also spoke about the lack of effective content moderation in the Polish language, and his view that this is exacerbating the problem of online hate:

"As far as I know there is only one moderator on Polish Twitter... I don't see anything taken down. I don't know how much gets reported. Everything stays no matter how hateful it is."[322]

The experiences of Poland's LGBTI community highlighted in this chapter demonstrate the harms to which X has contributed through its failure to adequately address TfGBV on its platform. The rampant TfGBV on X has contributed to members of the LGBTI community in Poland feeling unsafe both online and offline, negatively affecting their mental health and their ability to freely express themselves. Exposure to a near-constant stream of harmful content on X has undermined their rights to freedom of expression, non-discrimination and to live free from GBV. These case studies clearly show that X is failing in its responsibilities to respect the human rights of LGBTI people using its platform in Poland, and that X has also failed to adequately mitigate risks related to GBV on its platform, a requirement of the DSA.

The following chapter explores how X's business model risks further facilitating the spread of anti-LGBTI content.

---

[316] Amnesty International interview with Maja Heban, 30 July 2024.
[317] Amnesty International video call with Nathan Bryza, 9 August 2024.
[318] Amnesty International video call with Nathan Bryza, 9 August 2024.
[319] Amnesty International video call with Nathan Bryza, 9 August 2024.
[320] United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, "Country visit to Poland (18-29 November 2024): End-of-mission statement" (previously cited), para. 38.
[321] Amnesty International video call with Nathan Bryza, 9 August 2024.
[322] Amnesty International video call with Nathan Bryza, 9 August 2024.

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]

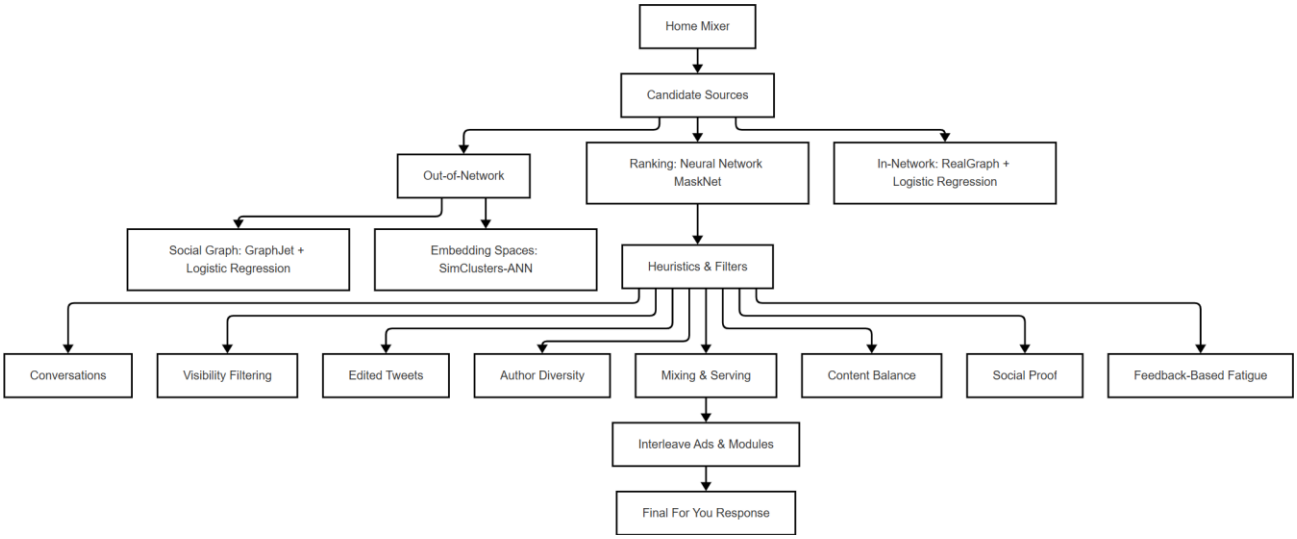
The recommendation pipeline has three main features:[366]

- 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.

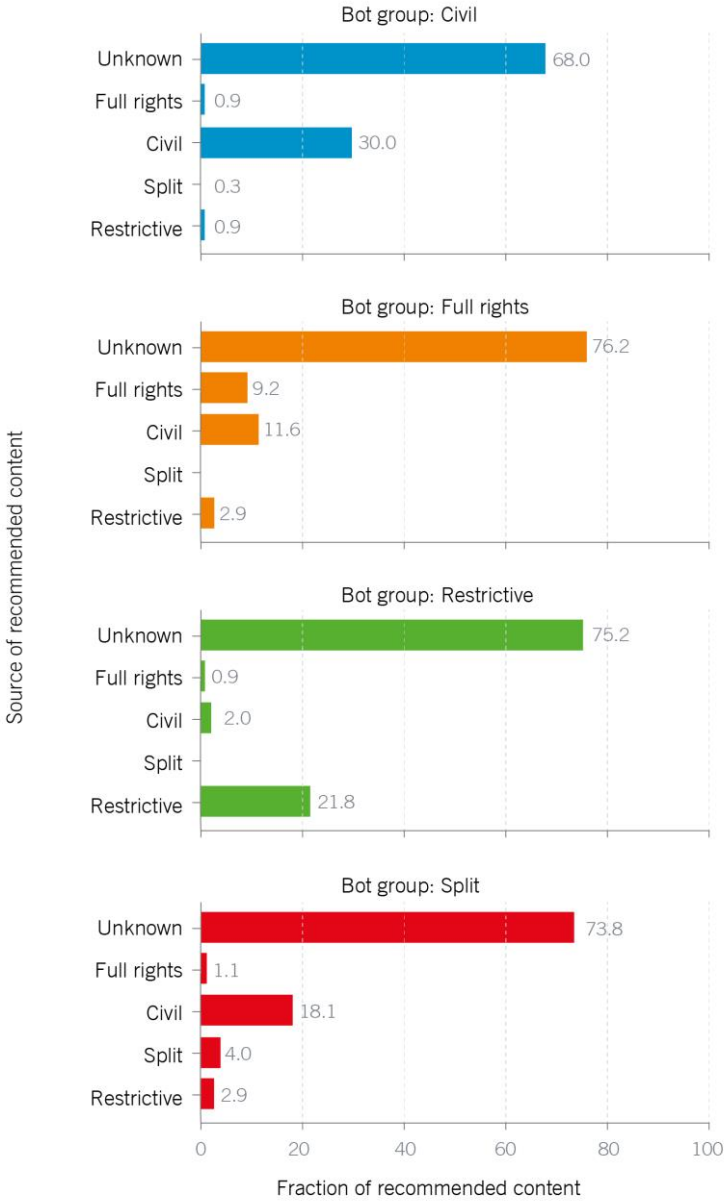
[399] Amnesty International interview with Aleksy (pseudonym), 28 July 2024.
[400] Manuel Pratelli and others, "Entropy-based detection of Twitter echo chambers" (previously cited).
[401] Manuel Pratelli and others, "Entropy-based detection of Twitter echo chambers" (previously cited).
[402] Manuel Pratelli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

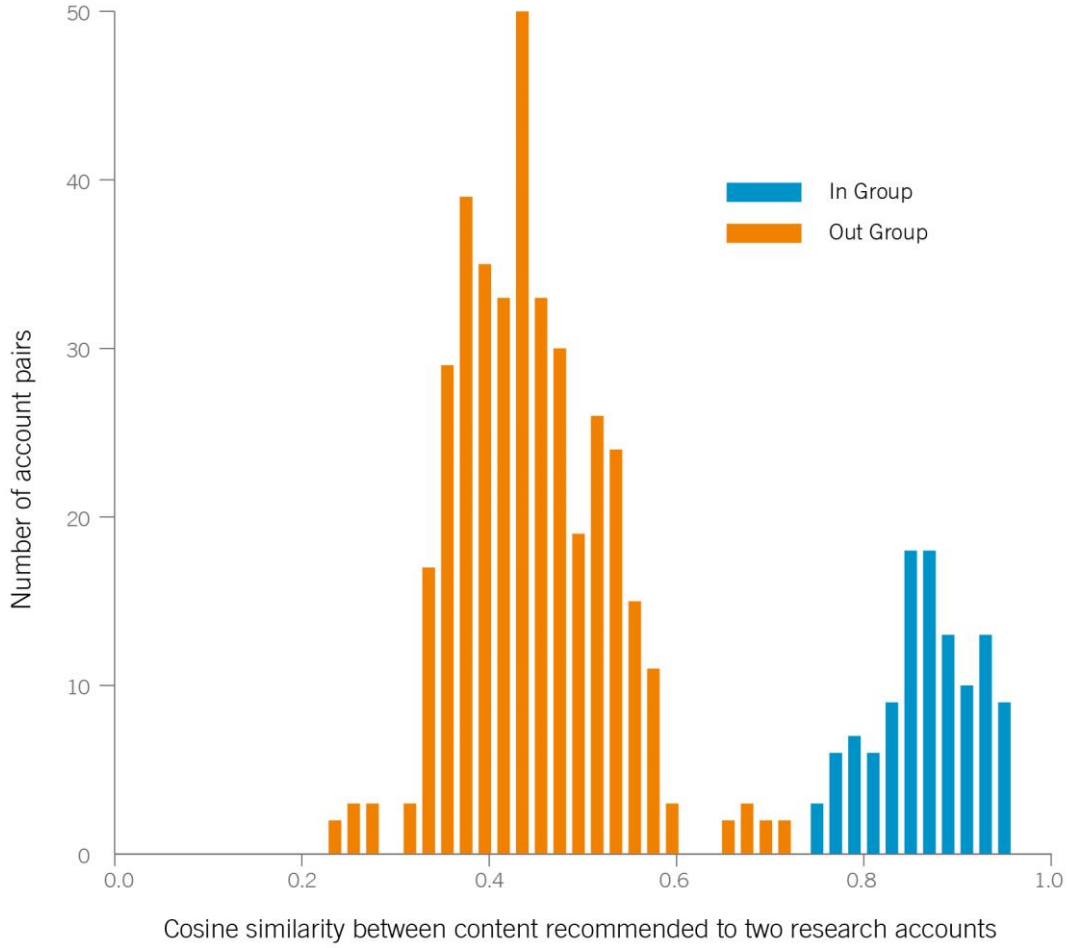
FIGURE 8



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).

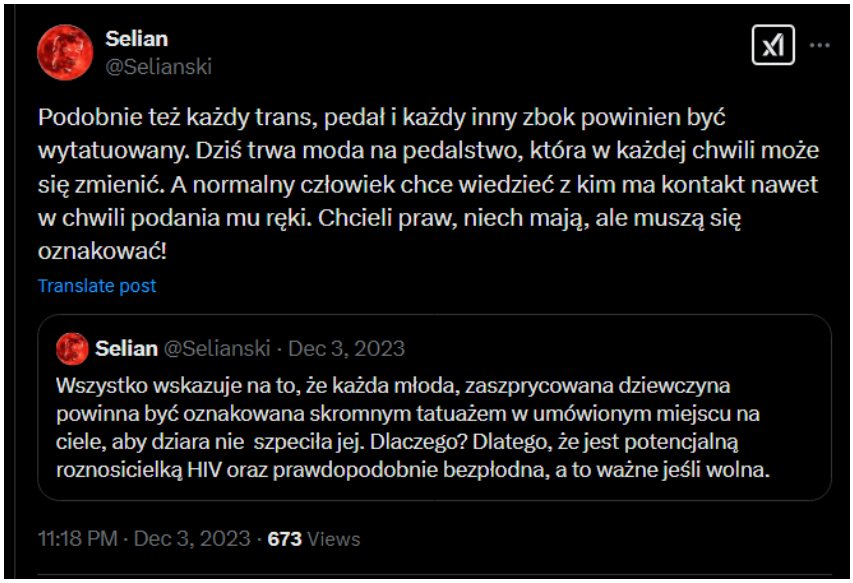
Several posts reported by Never Again Association during the year to August 2024 remained visible on the platform as of May 2025. These tweets are documented below and include posts that portray the LGBTI community as deviants, use slurs and call for discrimination against the LGBTI community.

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 from an X user, arcziwyspa, which features a photo of an LGBTI Polish politician. The text reads: "Get out of Gdansk you whore, such shame is brought to your father by you, you faggot."[419]



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!"[420]

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

[419] arcziwyspa, X post, 17 October 2023, https://x.com/ArcziWyspa/status/1714198139186385141
[420] Selian, X post, 3 December 2023, https://x.com/Selianski/status/1731452869792924087

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.[421] 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.[422] 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.[423]

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:

"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."[424]

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.[425] 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.[426]

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.[427] 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,[428] which do not adequately address the risks inherent in X's business model, including a focus on algorithmically optimizing for

[421] UN Guiding Principles, Principle 19 including Commentary.
[422] Amnesty International interview with Aleksy (pseudonym), 26 July 2024.
[423] Amnesty International interview with Mateusz Kaczmareck, 28 July 2024.
[424] Amnesty International interview with Julia Kata, 29 July 2024.
[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*, August 2024 (previously cited).

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.[429] Additionally, X has a company policy, introduced in March 2023, to remove violent hate speech from the platform.[430] 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 [431].

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",[432] 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.[433] Furthermore, according to X's latest risk assessment, it views hate speech as "illegal content" under the DSA framework.[434] However, as Poland does not specifically prohibit or criminalize hate speech targeting LGBTI people,[435] 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.[436] For slurs and tropes in particular, the company uses glossaries specific to EU languages.[437] This is, however, far from constituting adequate resourcing for content moderation, particularly as X has just two Polish-speaking content moderators.[438]

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."[439] 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."[440]

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

---

[429] 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).
[430] X, "Violent Content policy" (previously cited).

[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), p. 36.
[433] 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).
[434] 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).
[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, *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).
[438] X, *DSA Transparency Report – April 2025* (previously cited).
[439] 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.
[440] Amnesty International video call with Jakub Szymik, 5 August 2024.

"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'."[441]

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.[442]

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."[443]

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,[444] 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 moderation to X users.[445] 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.[446] 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".[447]

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".[448] 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."[449] 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.

---

[441] Amnesty International video call with Jakub Szymik, 5 August 2024.
[442] 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).
[443] Amnesty International interview with Magda Dropek, 24 July 2024.
[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).
[447] 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.
[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).
[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. 55.

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".[450] 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".[451]

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.[452] 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".[453]

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.

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.[454] 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:

---

[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. 22.
[451] 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.
[452] 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/
[453] UN Guiding Principles, Principle 13 including Commentary.
[454] Amnesty International letter to X, 22 August 2024.

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.

8. REMEDY AND AVOIDANCE OF FUTURE HARM

As outlined in Chapter 7, X has contributed to, and continues to contribute to, TfGBV-related human rights harms suffered by the LGBTI community in Poland. As a result, in accordance with the UN Guiding Principles, the company has a responsibility to provide effective remedy to those who have been adversely affected by its operations in Poland.

Additionally, under the DSA, X has an obligation to identify and mitigate systemic risks to fundamental rights which its operations may present.

This chapter explores how X can meet its responsibilities under international business and human rights standards and European law.

8.1 X'S RESPONSIBILITY TO PROVIDE REMEDY

8.1.1 INTERNATIONAL HUMAN RIGHTS STANDARDS

Companies that have contributed to adverse human rights impacts have a responsibility to adequately remediate those affected.[455] The appropriate type of remediation depends on the nature of the harm and may take a range of forms, including apologies, restitution, rehabilitation, financial or non-financial compensation, and justice (through criminal or administrative mechanisms), as well as guarantees of non-repetition for the prevention of future harm.[456]

A public apology is an important form of remediation, which acknowledges the facts and acceptance of responsibility, and could be accompanied by verification of the facts and full and public disclosure of the truth.[457]

An equally important form of remediation is a guarantee of non-repetition, which is intended to prevent abuses from occurring in the future. In this context, the prevention of further abuses can be achieved through several measures including regulatory and accountability measures taken by states, and actions taken by the companies themselves – any and all of which could contribute to guaranteeing non-repetition.[458] Among other things, efforts to guarantee non-recurrence could include committing adequate

[455] OECD Due Diligence Guidelines; UN Guiding Principles, Principle 22.
[456] UN Guiding Principles, Interpretive Guide, p.7.
[457] UN Basic Principles and Guidelines on the Right to Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law, 21 March 2006, UN Doc. A/RES/60/147, Principle 22.
[458] Amnesty International, *Injustice Incorporated: Corporate Abuses and the Human Right to Remedy* (Index: POL 30/001/2014), 7 March 2014, https://www.amnesty.org/en/documents/pol30/001/2014/en/, p. 18.

resourcing to X's content moderation operations in Poland and meaningfully engaging with civil society groups to better understand and address the issue of TfGBV.

8.1.2 PROVISIONS UNDER THE DSA

Under the UN Guiding Principles, states are required to take "appropriate steps to prevent, investigate, punish and redress" business-related human rights abuses within their territory or jurisdiction.[459] To this end, the DSA provides some recourse to remedy for platform users. In the case of infringements of the DSA, users and any organizations mandated to exercise their rights on their behalf have the right to make a complaint against the company to the Digital Services Coordinator (DSC) in their country of residence.[460] Platform users also have the right to receive compensation from companies against damage or loss stemming from infringements of the DSA.[461]

8.2 X'S COMPLIANCE WITH RESPONSIBILITIES UNDER THE DSA

Under Article 34 of the DSA, VLOPs such as X are required to undertake systemic risk assessments of their services in the EU due to their size and the potential impact they can have on society.[462] The Article 34 obligation requires X to "diligently identify, analyse and assess any systemic risks in the Union stemming from the design and functioning of their service and its related systems, including algorithmic systems, and from the use made of their services".[463] This should include any "actual and foreseeable negative effects for the exercise of fundamental rights such as freedom of expression, media freedom and pluralism, discrimination, consumer protection and children's rights".[464] When conducting its latest risk assessment, X should have taken into account, in particular, whether and how factors such as the design of its recommender systems, any other relevant algorithmic systems and its content moderation systems (among other things) influence any systemic risk to fundamental rights, as well as risks to GBV and physical and mental well-being.

However, X's 2024 risk assessment did not adequately consider these risks in general. It also did not consider risks to the rights of the LGBTI community stemming from its recommender systems, nor indeed from its lack of content moderation resources. Under Article 35 of the DSA, VLOPs should put in place reasonable, proportionate and effective mitigation measures tailored to the specific systemic risks identified pursuant to Article 34. These measures could include: adapting the design, features or functioning of their services, including their online interfaces; adapting content moderation processes; and testing and adapting their algorithmic systems, including their recommender systems.[465] Although X has not adequately identified the systemic risks it presents to the LGBTI community, these measures could be considered as a suite of options to ensure non-repetition of harms as part of an effective remedy for the community.

A crucial aspect of conducting human rights due diligence exercises such as risk assessments is stakeholder engagement. Recital 90 of the DSA emphasizes that stakeholder engagement with civil society and with specifically affected groups is an essential part of ensuring risk assessments and mitigations are based on the "best available information".[466] In particular, any assumptions about systemic risk must be tested with groups most affected by those risks.[467] However, none of the civil society organizations interviewed by Amnesty International for this research had been consulted by X as part of its risk assessment, and the risk assessment reveals very limited engagement with civil society organizations in general.

---

[459] Amnesty International, *What the EU's Digital Services Act Means for Human Rights* (previously cited).
[460] Amnesty International, *What the EU's Digital Services Act Means for Human Rights* (previously cited).
[461] Amnesty International, *What the EU's Digital Services Act Means for Human Rights* (previously cited).
[462] Digital Services Act, Article 34; European Commission, "DSA: Very large online platforms and search engines", 12 February 2025, https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops#:~:text=Those%20designated%20as%20VLOPs%20or,respond%20to%20the%20auditor%27s%20recommendations
[463] Digital Services Act, Article 34.
[464] Digital Services Act, Article 34; European Commission, "DSA: Very large online platforms and search engines" (previously cited).
[465] Digital Services Act, Article 35.
[466] Digital Services Act, Recital 90.
[467] Digital Services Act, Recital 90.

8.3 PENALTIES UNDER THE DSA

Under the DSA, the European Commission has direct supervision and enforcement powers and can, in the most serious cases, impose fines of up to 6% of the global turnover of a service provider.[468] However, the Commission's enforcement mechanism is not limited to fines: the DSC and the Commission have the power to require immediate actions where necessary to address very serious harms, and the platforms may offer commitments on how they will remedy them.[469]

If the Commission suspects that a VLOP has infringed any of the DSA's provisions, it can adopt a decision to open a formal proceeding.[470] Indeed, the Commission has already opened an investigation into X for possible infringement of the DSA in areas linked to risk management, content moderation, advertising transparency and data access for researchers.[471] At the time of writing, this investigation is ongoing.

Should the Commission conclude during the proceeding that there is an infringement of the DSA, it can take further enforcement steps which may include:[472]

- *Interim measures*: where there is urgency due to the risk of serious damage for users, the Commission can require immediate actions to address such harm. Any measure taken should be proportionate and temporary to mitigate such a risk. Examples of interim measures can be changes to recommender systems, increased monitoring of specific keywords or hashtags, or orders to terminate or remedy alleged infringements.
- *Binding commitments*: platforms involved in enforcement proceedings can make commitments to the Commission to ensure compliance with the DSA. Should the Commission consider them effective, it can accept these commitments by adopting a decision.
- *Non-compliance decision*: if the Commission finds that the DSA, the ordered interim measures, or the commitments made have been breached, it can adopt a non-compliance decision, after which the Commission can impose fines of up to 1% of the company's annual turnover.

[468] European Commission, "Questions and answers on the Digital Services Act" (previously cited).
[469] European Commission, "Questions and answers on the Digital Services Act" (previously cited).
[470] European Commission, "Questions and answers on the Digital Services Act" (previously cited).
[471] European Commission, "Commission opens formal proceedings against X under the Digital Services Act", 18 December 2023, https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709
[472] European Commission, "Questions and answers on the Digital Services Act" (previously cited).

9. CONCLUSION AND RECOMMENDATIONS

9.1 CONCLUSION

This report, based on a thorough investigation of X's role in facilitating TfGBV experienced by the LGBTI community in Poland, has firmly established that the company contributed to human rights harms and, therefore, has a corresponding responsibility to provide remedy to the community and to take additional mitigation measures to prevent the recurrence of harm in the future.

As a global company operating in numerous countries inside and outside of the EU, there is a significant risk that X's operations could fuel TfGBV in other contexts, particularly in non-English speaking countries. The risk is particularly acute due to the combination of the platform's overly permissive approach to content moderation, and its use of engagement-centric recommender systems.

In a context where anti-LGBTI sentiment had been present for several years, and with the community in the crosshairs of the former PiS government, X's lack of adequate content moderation resources and poor human rights due diligence helped to normalize hate, violence and discrimination against the LGBTI community. Despite claiming to be fostering a "town square" for free expression, X has allowed Poland's LGBTI community to be targeted with TfGBV on the platform, with very little recourse to remedy. At the same time, the company has consistently failed to engage directly with the LGBTI community, civil society organizations working on LGBTI issues in Poland, and even its own Trusted Flagger based in Poland.

These failures fomented anti-LGBTI sentiments and allowed for the targeting of the community to become increasingly normalized, resulting in X contributing to the TfGBV and human rights harm suffered by the community.

In 2018, Amnesty International similarly found that X (then Twitter) was failing to meet its human rights responsibilities regarding violence and abuse targeting women, including LGBTI women, on the platform.[473] In 2020, Amnesty International found that X (then Twitter) had failed to sufficiently address the prevalence of TfGBV on its platform.[474] The pervasive nature of this issue suggests a systemic problem which the company must urgently resolve.

The fact that X has continued to contribute to TfGBV in the five years since Amnesty International's last investigation – and that X has, in fact, changed its content policies to adopt an even more permissive approach to harmful content – strongly suggests that the company is neglecting its human rights responsibilities, and raises serious questions about whether any meaningful human rights due diligence is being conducted ahead of sweeping policy changes being announced. It also raises serious concerns around the company's willingness to take appropriate and effective mitigation measures, including adequately resourcing content moderation in non-English languages. The company's apparent failure to consider the systemic risks its operations present to the LGBTI community in its latest DSA risk assessment – and indeed its failure to seriously consider risks to rights other than the right to freedom of expression – is especially concerning and raises concerns about X's willingness or ability to engage with accountability measures.

[473] Amnesty International, "Toxic Twitter – a toxic place for women" (previously cited).
[474] Amnesty International, *Twitter's Scorecard* (previously cited).

Although X has taken several mitigation measures, such as allowing users to create block lists and to limit who can reply to their posts, these reforms are significantly below the level required to adequately mitigate the negative human rights impacts of its operations in Poland. Moreover, these measures are too limited in scope and therefore insufficient to provide a guarantee of non-repetition, particularly when moderation of Polish-language content remains extremely under-resourced.

The risk that X could contribute to further harm targeting the LGBTI community is heightened by the platform's engagement-centric business model. X's own blogs and policies show that engagement is a pillar of the platform's operations. This, combined with the existing prevalence of anti-LGBTI content, means that the possibility of algorithmically amplifying instances of TfGBV on the platform is a risk which must be considered in future mitigation measures.

This failure to properly assess and mitigate the risk to the LGBTI community in Poland suggests that X is not fulfilling its obligations under the DSA, in particular Articles 34 and 35. This analysis is supported by the independent audit of X's risk assessment, which found that the company's risk assessment process was not rigorous enough and that, while X demonstrated an awareness of risk-management activities, there was insufficient evidence of involvement in decision making in this area by the management body.[475] It is vital that X's next systemic risk assessment includes a thorough analysis and mitigation strategy for risks to human rights beyond freedom of expression, and that it gives appropriate consideration to the risks its operations present to marginalized communities, particularly members of Poland's LGBTI community, who have already experienced the harms of TfGBV on the platform.

The deterioration in the platform's ability to stem the prevalence of TfGBV on X since 2020 suggests an unwillingness or inability to make the necessary improvements to sufficiently and efficiently mitigate these harms. X should urgently change course and seek to improve its operations to uphold its responsibility to respect human rights, including through adequate resourcing of content moderation and building more meaningful relationships with civil society organizations in Poland and in all the countries in which it operates.

The unregulated development of the Big Tech sector has resulted in grave human rights consequences for marginalized communities around the world. However, the EU has taken a step to rebalance the scales through its adoption of the DSA. It is more crucial than ever that the European Commission, European Parliament and member states honour their obligation to protect human rights – including the right to live free from GBV – and robustly enforce the legislation. X's combination of weak content-moderation practices and surveillance-based business model is a ticking time bomb. Strong enforcement of the DSA can ensure that it does not explode.

9.2 RECOMMENDATIONS

9.2.1 RECOMMENDATIONS TO X

REMEDY AND PREVENTION OF FUTURE HARM IN POLAND

- Provide remedy to the LGBTI community in Poland for the TfGBV they have faced on X through a public apology, as well as by changing content moderation practices and the surveillance-based business model that will guarantee non-repetition in the future.
- Reform the "Trusted Flagger" programme in Poland, giving civil society organizations and human rights defenders more opportunity for involvement in content-related decision making.
- Significantly expand X's capacity to moderate Polish language content, including by hiring more Polish-speaking content moderators and ensuring that working conditions adhere to human rights standards.

HUMAN RIGHTS DUE DILIGENCE

- Undertake a comprehensive review and overhaul of human rights due diligence at X, including by mainstreaming human rights considerations throughout all of X's operations.

[475] FTI Consulting, "X Independent Audit" (previously cited).

- Ensure that human rights due diligence policies and processes address the systemic and widespread human rights impacts of X's business model as a whole, and be transparent about how risks and impacts are identified and addressed.
- Undertake a human rights impact assessment of X's operations in Poland and make the findings of any such assessment public in full.
- Ensure that human rights impact assessments are conducted in relation to the design and deployment of algorithmic systems in non-English speaking markets, to include meaningful public consultation and engagement prior to the finalization and deployment of a product or a service, with civil society, human rights defenders and representatives of marginalized or under-represented communities.

SURVEILLANCE-BASED BUSINESS MODEL

- Cease the collection of invasive personal data which undermines the right to privacy and threatens a range of other human rights.
- End the practice of using targeted advertising and embrace less harmful alternative business models, such as contextual advertising.
- Improve transparency in relation to the use of content shaping and content moderation algorithms, ensuring that their mechanics are publicly available in clearly understandable terms and are regularly updated as changes are made to algorithmic systems.
- Enable independent researchers to access and review data that is in the public interest, including data pertaining to content moderation and algorithmic systems.

CONTENT MODERATION

- Ensure appropriate investment in local-language resourcing throughout the world, with a particular emphasis on resolving existing inequalities which disproportionately affect non-English speaking countries.
- Ensure content moderators are given training which specifically addresses the needs of LGBTI platform users, and the risks associated with TfGBV.
- Ensure that content moderation guidelines are based on, and consistent with, international human rights law and standards, including on GBV.
- Ensure that content constituting TfGBV is restricted in line with international human rights law and standards, which allows for restrictions of freedom of expression to protect the rights of others, provided that these restrictions are necessary and proportionate to that aim.
- Ensure that reporting mechanisms are adequate, accessible to all users, including in widely spoken languages other than English, sufficiently clear, responsive and timely.
- Ensure that fact-checking is adequately resourced, that content moderation is not wholly outsourced to the Community Notes function, and that this function is supplementary to content moderation which is led by policies in line with international human rights law and standards, rather than acting as a replacement.
- Immediately review content policy changes post-2022 and amend policies to ensure that they are in line with international human rights law and standards, with consideration given to how content policies may affect marginalized or vulnerable communities.

9.2.2 RECOMMENDATIONS TO THE EUROPEAN COMMISSION

- Ensure that the Digital Services Act as a whole is robustly enforced, including by ensuring that countries without Digital Services Coordinators appoint them at the earliest opportunity.
- Introduce guidelines for Very Large Online Platforms conducting systemic risk assessments to ensure that the information contained in the assessments is appropriate and can be used by civil society to scrutinize the platforms.
- Expand the European Commission's current investigations into X to include an investigation into the company's ability to efficiently address the risk of TfGBV on the platform.

- Ensure that penalties issued as a result of non-compliance decisions are sufficient to encourage the non-repetition of harm.

9.2.3 RECOMMENDATIONS TO THE POLISH GOVERNMENT

- Fully implement the Digital Services Act by appointing a Digital Services Coordinator as soon as possible.
- Ensure that, once a Digital Services Coordinator is appointed, the position is effectively resourced in terms of expertise, capacity and funding.

AMNESTY INTERNATIONAL IS A GLOBAL MOVEMENT FOR HUMAN RIGHTS. WHEN INJUSTICE HAPPENS TO ONE PERSON, IT MATTERS TO US ALL.

CONTACT US

contactus@amnesty.org

+44 (0)20 7413 5500

JOIN THE CONVERSATION

www.facebook.com/amnesty

@Amnesty

‘A THOUSAND CUTS’

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

In 2019, the LGBTI community became a key target for Polish politicians and anti-rights activists during that year’s election campaign. Many politicians and activists used X to spread anti-LGBTI rhetoric and incite violence and discrimination against the community.

This report is an investigation into X’s contribution to harms suffered by the LGBTI community in Poland between 2019 and 2025. It reveals the toll that X’s engagement-based business model has taken on LGBTI individuals, and particularly those targeted with hate on the platform. It shows that, despite the company’s proclaimed desire to protect freedom of expression, online abuse on X has contributed to the LGBTI community in Poland living in fear of being their true selves – both online and offline.

Despite an obligation to mitigate systemic risks under the EU’s Digital Services Act, X has failed to engage meaningfully in human rights due diligence or mitigation measures in Poland. However, Amnesty International’s analysis makes clear that its business model is a threat to marginalized community globally, and underlines the necessity of robust enforcement of the DSA and tech regulation globally, to ensure that marginalized communities are safeguarded from the adverse impacts of the business model – whoever they are, wherever they are.

INDEX: EUR 37/0098/2025
JULY 2025
LANGUAGE: ENGLISH

amnesty.org

AMNESTY INTERNATIONAL