REPORT SETTING OUT THE RESULTS OF TWITTER INTERNATIONAL UNLIMITED COMPANY RISK ASSESSMENT PURSUANT TO ARTICLE 34 EU DIGITAL SERVICES ACT

AUGUST 2024

TABLE OF CONTENTS

I. Executive Summary

With over 45M monthly active users in the EU, X was designated as a Very Large Online Platform (VLOP) under the EU Digital Services Act (DSA) on April 25, 2023. In accordance with DSA Article 34, we have conducted a comprehensive assessment that identifies, analyses and assesses any systemic risks to the Union stemming from the design or functioning of our service, its related systems (including algorithmic systems) and from the use made of our services.

In keeping with our legal obligations under EU law, we have taken into consideration the following factors: the dissemination of illegal content through our service; any actual and foreseeable negative effects to the exercise of fundamental rights; any actual or foreseeable negative effects in relation to civic discourse, electoral processes, public security; and any actual or foreseeable negative effects in relation to gender based violence, the protection of public health and minors and serious negative consequences to the physical and mental well-being of individuals. In accordance with Article 34(2), the risk assessment also addresses our recommender systems, content moderation systems, applicable terms and conditions, systems for the selection and presenting of advertisements and any of X's data related practices. This risk assessment covers TIUC's designated service[1] as of June 30, 2024.

In this DSA Risk Assessment summary report, X summarises the outcomes of its second annual systemic risk assessment exercise. As this exercise builds on the first risk assessment, X uses 'Y1' to refer back to the risk assessment exercise and report submitted in 2023, and 'Y2' to refer to the risk assessment conducted in 2024 and the current report. This report summarises X's consideration of new inherent risks since August 2023, new and improving controls in place, the residual risk that remains on the platform, and further routes that X could explore to tackle the residual risk.

Our Y1 methodology aimed to serve as a blueprint for future risk assessments. In Y2, we have enhanced the methodology with further learnings from academia, industry best practices, regulatory guidance, and internal stakeholder feedback. In accordance with DSA Article 34, our risk assessment covers the four systemic risk areas, and provides a granular assessment through 13 individual assessments.

For each identified risk area, we assessed how our platform's design, functioning, use, or potential misuse, could result in inherent risks in Y2; mapped existing and new controls and remediations against these inherent risks; and assessed the residual risk that remains on our platform in Y2. Following our assessment, we found that our controls bring down the level of risk for most areas to a low or medium level. We look to improve our existing controls and explore further measures, to continue to mitigate this residual risk. Our measures are designed to address Article 34 systemic risks and are proportional to X's capacity, while avoiding unnecessary restrictions on service use. Special consideration is given to the impact on freedom of expression. Acknowledging that these systemic risks are continuously evolving and can be

[1] Twitter International Unlimited Company (TIUC) is the service provider of the X VLOP (X) in the EU. Throughout this report, we will use "X" to refer both to the designated VLOP service and its service provider.

impacted by intentional coordinated exploitation, we remain committed to continuing to monitor and mitigate these risk areas.

We have conducted this DSA systemic risk assessment utilising our knowledge, resources, and understanding of DSA requirements. Internal teams across the globe, including X management, the DSA Leadership team, Safety, Product Engineering, Legal, Privacy & Data Protection, Global Government Affairs (GGA), the Independent Compliance Function, the TIUC Board, along with external resources, were relied on in this cross-functional exercise. This second assessment serves as a continuation of our efforts to maintain platform safety in an evolving and iterative process, as envisaged by the DSA.

II. Introduction

X's mission is to promote and protect the public conversation, serving as a trusted digital public town square. With more than 45M monthly active users in the EU, X was designated as a Very Large Online Platform (VLOP) under the EU Digital Services Act (Regulation 2022/2065; the DSA) on 25 April, 2023.

In 2024, we have seen major European elections, including the EU elections and national elections in the Union, alongside emerging public narratives on significant events, such as the Israel/Hamas conflict post-October 7th. As a platform that facilitates public conversation, X has responded to this changing risk environment by addressing the online conversations stemming from these off-platform events in a proportionate manner - balancing freedom of expression while ensuring that our platform and users remain safe. Balancing human rights, including the right to freedom of speech, are the foundation of how we think about and iterate on policy and enforcement. X's approach to policy and enforcement factors in potential impacts on human rights, including negative impacts to physical safety, privacy, and freedom of expression being most significant and ones to prevent and mitigate. We believe it is our responsibility to keep users on our platform safe from content violating our Rules.

Last year, we developed our DSA risk assessment methodology with reference to multiple existing frameworks, including, but not limited to, the UN Guiding Principles on Business and Human Rights and the DTSP Safe Assessments Framework, and adapted them to the unique environment of X.

In consideration of new guidance, we introduced a more robust methodology for our score calculations across the four systemic risks identified by Article 34(1) of the DSA. We further identified subcategories of each risk to facilitate more granular analysis. Additionally, we standardised our evidence base, enabling a more precise scoring system and better comparability across risk areas. Notably, we adjusted our scales to consider vulnerable groups and X users in the EU, providing a more nuanced understanding of how such content manifests on the platform and its reach. These changes are further detailed in VI. Methodology.

Our risk assessment, consistent with last year's approach, involved analysing existing controls to reduce inherent risks and considering additional measures to mitigate systemic risks identified in the assessment. A summary of the results of this exercise can be found in VII. Summary of risk assessments. In identifying further mitigation measures, we considered the residual risks, our economic capacity, and the impact on fundamental rights, particularly freedom of expression. These measures are detailed in VIII. Considerations for further mitigations.

We conducted this risk assessment using our expertise, resources, and understanding of the DSA requirements, while also considering established and emerging cross-industry standards. As the risk assessment and management framework is a continuous exercise, we refer back to our Y1 report and take into consideration the Y1 scores, in order to track the evolution of risks.

III. The DSA & X

With over 111M average monthly users in the EU[2], and 250M daily users globally,[3] X continues to be an indispensable platform for the world.[4] Since August 2023, we have adopted and reinforced a vast number of measures to improve our safety mechanisms and empower users in the EU. In compliance with the DSA, this has included a dedicated illegal content reporting form and appeal form for users in the EU, updated communications and statement of reasons to users following enforcement actions, biannual DSA transparency reports, and increased transparency to users about our ads and recommender systems. We have also onboarded designated trusted flaggers, and collaborated with civil society organisations in preparation for and during the elections that took place in the EU over the past year.

While balancing freedom of expression, our cooperation with law enforcement for information requests, removal orders, and proactive referrals in cases of suspicions of criminal activity is ongoing and we have established dedicated points of contact for both EU authorities and users to contact us with their DSA inquiries. Our Terms of Service and various Help Centre pages have also been updated following the DSA, to clearly reflect summaries of our terms, as well as new information to help our users understand our recommender systems and give them more control over their experience on X.

Our ads transparency center also provides EU users a look into all advertisements and commercial communications present on the platform with instructions on how to get started. We have also opened an application process for qualified researchers to apply for X API access to conduct research related to DSA systemic risks, separate to our subscriptions for general academic research.

Our product development process has been enhanced to consider dark patterns in a broader context, having historically focussed on dark patterns arising in a data protection context. We also conduct assessments of products that may have a critical impact on systemic risks in the EU, both at a pre-deployment stage and throughout the product's lifetime. This is also core to our risk assessment and risk management process, which we see as a continuous effort over time to mitigate potential risks on X.

Although many of these risks may be manifestations on the platform of existing offline issues, we recognise the role that online platforms may play in disseminating and potentially exacerbating the harms. This is why we continue to invest resources into the DSA risk assessment, an exercise conducted and overseen by a cross-functional team including Safety, Product Engineering, Legal, Privacy & Data Protection, Global Government Affairs (GGA), the Independent Compliance Function, and the TIUC Board.

[2] https://transparency.x.com/en/reports/amars-in-the-eu

[3] https://x.com/XData/status/1769826435576037702

[4] https://blog.x.com/en_us/topics/company/2023/an-update-on-our-work-to-tackle-child-sexual-exploitation-on-x

IV. X Risk Environment: Influencing Factors & Controls.

We are constantly improving our rules, processes, technology, and tools to ensure that all of our users can participate in public conversation freely and safely. X's mission has guided our approach to navigating the multi-platform risk environment in which we exist, aiming to provide a service where all users have the power to create and share ideas and information. Our 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.

Although the factors listed in Article 34(2) were considered in the context of each systemic risk (captured in VII. Summary of risk assessments), many of these factors pose similar risks, and are mitigated by controls, in a horizontal manner - i.e, acting across all systemic risks. As such, they have been explained below, drawing upon the conclusions from the Y1 exercise and providing insights into changes in risk and corresponding controls in Y2.

Risk of misuse and inauthentic use of X

X is situated in a multi-platform risk environment and bad actors can misuse the service in the same way they misuse other social media platforms. Many risks and harms that manifest on X appear as extensions of often already rapidly evolving offline risks. These risks interact in complex and novel ways across the online platform ecosystem. While our controls are constantly working to reduce harm, we recognise that bad actors may stay a step ahead, and our platform is not invulnerable to manipulation.

Between October 2023 to June 2024, almost [redacted] of our total enforcement action[5] for X Rules violations was under our Platform Manipulation and Spam policy, indicating the high volumes of such risk on X, as well as X's efforts to mitigate it. Forms of inauthentic behaviour may include, but are not limited to, financially motivated spam, inauthentic engagements, as well as coordinated activity to artificially amplify hashtags, trends, and other conversations. In April 2024, we initiated additional proactive measures to eliminate accounts that violate our Platform Manipulation and Spam rules to ensure that X remains secure and free of bots.[6] These measures resulted in a significant decline in violative accounts, and we continue to iterate on these measures to continue catching pivoting threats.

Design and functionality

We offer a variety of features for users to engage with on the platform through different mediums and formats, such as posts, Spaces, Communities, and X Live, as well as via subscription through X Premium. To learn more about our suite of product-level safety features as well as user controls that allow users to have a safe and meaningful experience on X, please refer to our Y1 report.

[5] Total enforcement data was calculated by taking the sum of total suspensions, total content removals, and an extrapolated total restricted reach labelled posts for the time period of October 2023 to June 2024. For the extrapolated total restricted reach labels, an estimate for the time period was used, as due to data retention issues, real figures are only available for [redacted]. As such, these values should be understood to be estimates.

[6] https://x.com/Safety/status/1775942160509989256

Over the last year, we have rolled out new updates to our existing features – such as improvements to Community Notes – as well as new features such as making likes private[7], to continue our work in creating a safe experience for our users.

Zoom in: Community Notes

This year, Community Notes has more than 100K contributors across the EU, and has been launched on media and videos as well. Posts that have a note on it are demonetised, ensuring that there is no revenue generated from false or misleading information.

External researchers found that users repost 61% less often after a post gets a Community Note, while another study found around a 50% drop in reposts and 80% increase in post deletions after a post received a Community Note. This aligns with our own research that found a large causal drop in reposts, quotes, and likes on noted posts in an A/B test. This reduction is entirely due to organic user behaviour, since X does not rank posts differently when they are noted. Another recent study found that, across the political spectrum, Community Notes were perceived as significantly more trustworthy than traditional, simple misinformation flags. It also found that Community Notes had a greater effect on improving people's identification of misleading posts. A key driver is believed to be the detailed context that notes provide, right where people can see it.

Speed is important in addressing misleading information — the sooner people see added context, the better. In the past year we've seen that notes can respond quickly at critical times. In the first few days of the Israel-Hamas conflict, notes appeared at a median time of just 5 hours after posts were created. This calculation does not even include notes on images/videos — over 80% of noted posts are showing media notes, which appear instantly on new posts that include previously noted media. It's also common to see Community Notes appearing days faster than traditional fact checks — which is possible because of the collective intelligence of the contributor community. In the past year, we've shaved 3-5 hours off the typical time it takes for notes to be scored, and [redacted] On top of this, people who engage with a post before it receives a note get a notification about it. Updates and improvements to our notes are regularly communicated via our X Community Notes handle[8].

Request a Community Note: As of July 2024, users can request a Community Note on a post they believe would benefit from one. This is both a way for everyone on X to help, and it allows Community Notes contributors to see where help is wanted, potentially helping to accelerate their work in proposing new notes. This feature is in pilot testing, and currently only available on the browser version[9].

[7] https://twitter.com/XEng/status/1800959499932496139

[8] https://x.com/CommunityNotes/status/1788617818784792880

[9] While this feature was only available on the browser version as of the date of conducting the risk assessment, this was expanded to iOS and Android on Aug 20, 2024 .

Prior to deployment, all products go through safety checks to ensure a scaled and monitored approach to launching products. X has incorporated and followed an evaluation process to identify and assess products, features, and functionalities that are likely to have a critical impact on the systemic risks identified under Article 34, in line with the pre-deployment risk assessment duties in Article 34(1).

Beyond products, we strive to give more control to users to control their experience on the platform through features such as block/mute, hide, and unfollow. Likes were also made private in June to better protect our user's privacy[10]. This means that users can no longer see who liked someone else's post. Only a post's author can see who liked their posts. This also protects freedom of expression as public likes may have resulted in self-censorship for fear of reaction from viewers.

Recommender systems (Article 34(2)(d))
Our recommendations are based upon a variety of signals, including, but not limited to, interests you choose during onboarding, accounts & Topics you follow, posts you've liked, reposted, or otherwise engaged with, and content that is popular in your network. Recommendations may amplify content and can unintentionally elevate specific sources and may reduce the reach of pluralistic sources of information. Until our systems have flagged an account or content as violative or potentially violative, they remain eligible for amplification and recommendation by our systems. During that time, such accounts and content may continue to receive engagement, thus contributing to their distribution and reach. In an attempt to create personalised experiences, our systems may also run the risk of limiting pluralistic sources.

To mitigate this risk, recommender system controls include safety models to prevent violative accounts and content from being recommended, implementing eligibility requirements for the recommender system, ensuring that sensitive content or inappropriate advertising is not shown to accounts of known minors, and blocking violative keywords from showing up on search autocomplete and trending. Content that is labelled under relevant policies is ineligible for recommendation, which further reduces the spread of such content. Over the past year, users also have the option for each recommender system to engage with non-profiled content. The content shown to users under these options is typically the most recent or popular content without factoring personalised information, or strictly content from accounts that a user has chosen to follow. Further, user controls tools - such as unfollow, mute, block, report, show less often, and more - are designed to help users control what they see and what others can see about them. Recommender systems are thus influenced by such user choices – for example, recommendations delivered to users will not suggest content that includes their muted words or hashtags.

Our approach to recommender systems, along with the parameters used in these systems and how users can influence them are explained in the following blogs: About our approach to recommendations, Communities Recommendations, Conversations Recommendations, Spaces Recommendations, Trends Recommendations, Search Recommendations, and For You Home Timeline Recommendations.

[10] https://x.com/XEng/status/1800634371906380067

Policies and enforcement (Article 34(2)(c))

Our aim is for our policies and enforcement measures to be consistent, reasonable, proportionate, and effective. To achieve that, we have built a policy development process focused on balancing the safety and freedom of expression of our users. Our operations and policy functions work together to identify limitations and update policies and enforcement guidelines, as part of our incident responses. To learn more about our policy development process, please refer to our Y1 report.

Over the past year, as part of our ongoing commitment to refine our policies and enforcement, we have conducted a comprehensive review of our existing guidelines and workflows. This has led to improvements in X media policies, particularly around consensual adult content and violent media. By separating our Sensitive Media policy to Adult Content and Violent Content[11], we've accomplished the following:

- User transparency with enhanced and distinct Help Center articles, and reporting experience;
- Clearer data on the prevalence of adult versus violent content on our platform. Previously, such content was grouped under the broad category of Sensitive Media, which did not allow for nuanced analysis; and
- Operational efficiency with clearer guidelines and training/onboarding expectations.

We employ a range of enforcement options, either on a specific piece of content (e.g., an individual post or Direct Message) or at an account-level through suspensions. In determining what enforcement option to apply, we carefully consider that activity on X is largely reflective of real life conversations, events, and social movements that may include perspectives that could be perceived as offensive or controversial by our users. To learn more about our approach to enforcement, please refer to our Y1 report. For more information on our approach to restricting reach of content, please refer to B. Exercise of fundamental rights.

Content moderation systems (Article 34(2)(b))

X takes seriously its commitment to being a safe platform for all people who use it in a manner consistent with our Rules, and strives to ensure that our Rules are not implemented in a discriminatory manner with respect to protected characteristics. However, as with all moderation systems, there remain inherent risks of false positives and false negatives, for example due to moderator bias, language specialisation, resource allocation, or potential limitations of automated tools.

Over the last year, we have been moving towards an information-first approach for moderating content, which reduces the risk of moderator bias in decision making. Historically, a decision first approach has been employed – which means that a moderator analyses content against policy criteria, to then decide if it is a violation or not. However, this risks subjectivity, notably if the criteria is inconsistently applied by different people. An information-first approach aims to reduce potential bias and increase enforcement consistency by having moderators get to an enforcement decision by answering a set series of questions, rather than having them immediately make a decision. For more information on our own initiative content moderation

[11] Note that the Sensitive Media policy previously included consensual adult content and violent media within it. As such, allowing consensual adult content on the platform is not an enforcement change, as X has always permitted consensually produced and shared adult content.

efforts as well as on our human resources dedicated to content moderation, please refer to our transparency reports.

Our human review efforts are led by an international, cross-functional team with 24-hour coverage and the ability to cover multiple languages. We provide our reviewers with a robust support system to ensure that they are prepared to perform their duties. Each reviewer goes through extensive training and refreshers, and they are provided with a suite of wellness initiatives. Manual content moderation resourcing requirements can experience fluctuations based on a variety of challenges such as trending issues and product feature changes. To address this, weekly operational capacity review meetings are held that consider incoming volumes, our meet rate against service legal agreements, any case backlog accumulation, and assessment of risk. As a result of this analysis, moderation resources may be reallocated, removed or reserves committed to address emergent crises and opportunities

Automated enforcements for X Rules undergo testing before being applied to the live product to mitigate the above. Both machine learning and heuristic models are trained and/or validated on data points and labels (e.g., violative or non-violative) that are generated by trained human content reviewers. We have feedback loops for our automated detection systems to monitor their performance using the rate at which human content reviews agree with the automated system decision. Reviewers have expertise in the applicable policies and are trained by our policy specialists to ensure the reliability of their decisions. Human review helps us to confirm that these automations achieve a level of precision, and sizing helps us understand what to expect once the automations are launched.

In addition, humans proactively conduct manual content reviews for potential X Rules violations. We conduct proactive sweeps for certain high-priority categories of potentially violative content both periodically and during major events, such as elections. Agents also proactively review content flagged by heuristic and machine learning models for potential violations of other policies, including our Violent Content, Child Sexual Exploitation and Violent and Hateful Entities policies. Once reviewers have confirmed that the detection meets an acceptable standard of precision, we consider the automation to be ready for launch. Once launched, automations are monitored dynamically for ongoing performance and health. If we detect anomalies in performance, our Engineering teams - with support from other functions - revisit the automation to diagnose any potential problems and adjust the automations as appropriate.

Systems for selecting and presenting ads (Article 34(2)(d))
As with all online platforms, there is an inherent risk that violative ads could be posted on our platform. While our moderation systems and human moderators work to identify such ads, they may not catch every violation, potentially leading to missed violations or uneven enforcement. Additionally, advertisers may attempt to target minors based on profiling and using personal data. Users might also face challenges in understanding ad targeting, their privacy options, or the process for reporting ads that violate our policies.

At the ad creation time, our system is set up to proactively detect violative ads by employing machine learning models and business logics such as denylist terms so as to mitigate this risk. Denylist terms restrict content from appearing on promoted posts. When a term is added to the ad review denylist, any promoted content mentioning the term or phrase will automatically put the advertisement into a review hold state, requiring a human review before proceeding. There remains a possibility that some ads may bypass our detection methods. We also leverage human reviews to verify system detections, which can also be initiated due to user reports. Detected ads are halted or restricted per our X Ads policies. As an additional control, Community Notes can be added to X ads, to help ensure the veracity of the advertiser's claims and allow access to more information. Further, since August 2023, X does not present ads to minors in the EU.

Finally, X does not allow political ads in the EU. A recent study by Global Witness on how social media platforms treat election disinformation, notably in ads, showed that X halted all ads and suspended the creation of accounts for violating X Ads policies, indicating a well functioning policy and enforcement mechanism compared to VLOP peers. Finally, in our efforts to protect minors, we have turned off advertisements to minors in the EU.

Data related practices (Article 34(2)(e))

As discussed in the Y1 report, to embed privacy throughout our organisation, X conducts legal and privacy reviews for all new projects involving personal data. Our most recent privacy and security external audit conducted in 2023, for the purpose of assessing the establishment, implementation, and maintenance of X's Privacy and Information Security program, showed that our Privacy and Information Security Program is comprehensive, provides sufficient coverage across all relevant privacy and information security domains, and is in alignment with ISO 27701 and ISO 27001/02 frameworks, upon which the Program is based. The audit findings also stated that our privacy and information security risk management strategies, monitoring, and mitigation approach highlights that we continue to prioritise privacy and information security as foundational within the organisation. Please note, a 2024 privacy and security audit is currently underway. As in Y1, X conducted a dedicated risk assessment for data related practices and protection of personal data, under the systemic risk of negative effects to fundamental rights.

Cooperation with law enforcement

X cooperates with law enforcement authorities in the EU. Law Enforcement can issue X content removal requests, information requests, emergency disclosure requests or data preservation requests. We have dedicated online guidelines and a portal available for law enforcement to use, which our teams monitor 24/7. Requests from governments and law enforcement authorities are reviewed for compliance with international human rights and legal standards. Our DSA transparency reports provide more information around our collaboration with law enforcement in the EU.

Other continuous mitigation measures

At the end of our first DSA risk assessment cycle, our cross-functional risk assessment team considered our risk profile and identified areas where further mitigations could be explored. In our Y1 report, we outlined these measures, in compliance with Article 35(1). Many of these mitigations were described in the III. The DSA & X section above, and others require continuous efforts.

The following are the Article 35 mitigation measures enacted between August 2023 and June 2024:

- Our Civic Integrity policy was launched in mid September 2023, to address voter intimidation and suppression during elections *(Article 35(1)(b));*

- We continued to conduct comprehensive policy reviews, which has led to improvements in our policies. Notably, disassociating consensual Adult Content and Violent Media from the existing Sensitive Media[12] policy has helped with establishing clearer definitions and enforcement guidelines *(Article 35(1)(b));*
- We made changes to our global list of designated violent entities and expanded it, as part of our continuous work to carry our comprehensive assessments. We also increased proactive monitoring and enforcement for violent entities *(Article 35(1)(f))*;
- We built out our Misuse of Reporting Features policy that provides an objective, effective and transparent procedure to mitigate the potential misuse of X's reporting mechanisms from users of the X platform *(Article 35(1)(b)* ;
- Restricted reach labels can now be applied by content moderators to content that users report for violating the X Rules. This allows for more proportionate enforcement action on user reports as well as more consistent application *(Article 35(1)(c))*;
- We continue to take proactive efforts to mitigate online abuse. These measures are tailored to global events and crises, and deployed as needed. Over the last year, this has included the use of heuristic rules for sporting events such as the Euros as well as alerts for additional detection for targeting of politicians during the EU elections. *(Article 35(1)(f));.*
- We updated the reporting flow to ensure users take fewer clicks to report harassment. This eases the burden on the user to ensure a swift and seamless reporting experience *(Article 35(1)(a));*
- We improved our internal workflows to ensure more accurate routing of user reports to the correct teams for reviews – this has resulted in swiftly addressing any instances of harassment *(Article 35(1)(c))*;
- We scaled the option for X Premium users to verify their accounts through identification with a 3rd party partner globally *(Article 35(1)(a))*;
- We expanded Community Notes to Media and weekly updates are rolled out and communicated via our X handle *(Article 35(1)(a))*;
- Designated trusted flaggers in the EU, alongside X Trusted Partners, are able to use our reporting channels and escalate content to us that will be reviewed in a prioritised timely manner *(Article 35(1)(g))*;
- We continued to enhance feedback mechanisms with post-incident reviews and regular syncs to ensure that enforcement aligns with the spirit and purpose of the policies *(Article 35(1)(c))*;
- We have continued to enhance our privacy program with regular updates to leadership, as well as set up the process for privacy reviews on recommender systems *(Article 35(1)(d)&(f))*;
- We have continued and expanded our engagements with civil society organisations. New engagements include involvement with Project Lantern, Jugenschutz, Global project against hate and extremism, INACH and Search for Common ground *(Article 35(1)(f))*;
- We have supported the dissemination of media literacy campaigns that fostered the spread of reliable information on the electoral process. For instance, we supported the EDMO "Be elections smart" campaign and the ERGA campaign to prevent the spread of

[12] Note that the Sensitive Media policy previously included consensual adult content and violent media within it. As such, allowing consensual adult content on the platform is not an enforcement change, as X has always permitted consensually produced and shared adult content.

misleading information on elections. We also supported campaigns to stop violence against women *(Article 35(1)(i))*.

A number of the mitigations also are in progress and require continuous work. These include:

- Our operational overhaul, where continuous work is being done to make our operations measurable and implement built-in feedback loops. So far, completed work includes the streamlining of user reports, improving efficiency of review processes, and updating guidelines to follow an objective, information-first approach. *(Article 35(1)(c))*;
- Reinforcing our internal monitoring and data extraction systems for risk assessments and transparency reports, to showcase trends and regional visualisations. *(Article 35(1)(f))*;
- We continue to expand our Global Government Affairs team and increasing resources allocated to ensuring elections integrity is an ongoing process *(Article 35(1)(f))*.

V. X DSA Systemic Risk Governance Framework

Our risk governance framework, as described in our Y1 report, has been revised and improved at a regular cadence throughout the last year. In accordance with Article 34, we annually report on systemic risks with the involvement of a cross-functional team that comprises Safety, Product Engineering, Legal, Privacy & Data Protection, Global Government Affairs (GGA), the Independent Compliance Function, and the TIUC Board. Our DSA Systemic Risk Governance Framework also foresees, in accordance with Article 34(1), the process for risk assessments prior to deploying functionalities that are likely to have a critical impact on the EU systemic risks.

Furthermore, in line with Article 41 and X's continuous risk management duties, the Independent Compliance Function, the DSA Leadership team, and the TIUC Board work together with X's cross-functional risk assessment team to ensure systemic integrity risks are properly identified, mitigated and managed. These frameworks collectively inform X leadership's understanding and commitment to meeting its Article 41 management body obligations, with respect to governance arrangements and overseeing, monitoring, and mitigating systemic risks under Article 34 and 35.

The Independent Compliance Function Policy outlines the Independent Compliance Function's specific duties. Specifically, the Independent Compliance Function is involved in reviewing the methodology of the risk assessment, ensuring its adequacy and completeness, communicating any updates to the TIUC Board and other relevant leaders, and reviewing the results of the risk assessment. All key stakeholders are involved in ensuring that reasonable, effective and proportionate mitigations are implemented in respect of all systemic risks identified, in observance of fundamental rights.

X acknowledges that the Commission can require VLOPs to take action under Article 36 in cases where extraordinary circumstances lead to a serious threat to public security or public health in the Union or in significant parts of it. Our framework accordingly sets out a process for responding to requirements under the crisis response mechanism. The Independent Compliance Function Policy establishes the Independent Compliance Function's role in monitoring TIUC's compliance with commitments made under the codes of conduct or crisis protocols, when activated.

VI. Methodology

In accordance with DSA Article 34, we have conducted a comprehensive assessment that identifies, analyses and assesses any systemic risks to the Union stemming from the design or functioning of our service, its related systems (including algorithmic systems) and from the use made of our services.

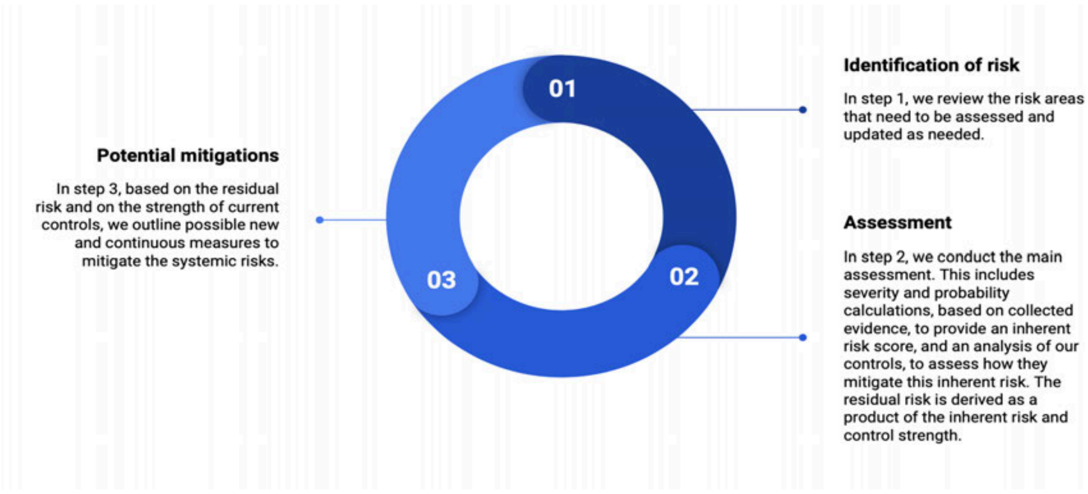
In keeping with our legal obligations under the DSA, we take into consideration the following systemic risks: the dissemination of illegal content through our service; any actual and foreseeable negative effects to the exercise of fundamental rights; any actual or foreseeable negative effects in relation to civic discourse, electoral processes, public security; and any actual or foreseeable negative effects in relation to gender based violence, the protection of public health and minors and serious negative consequences to the physical and mental well-being of individuals. The assessment addresses, in accordance with Article 34(2), our recommender systems, content moderation systems, applicable terms and conditions, systems for the selection and presenting of advertisements and any of X's data related practices. The following recitals complementing Article 34 were also consulted: 12, 79, 80, 81, 82, 83, 84, 85, 89, and 90.

In 2023, we developed our DSA risk assessment methodology with reference to multiple existing frameworks, including, but not limited to, the UN Guiding Principles on Business and Human Rights as well as the DTSP Safe Assessments Framework, and adapted them to the unique environment at X. As part of continuous risk management, our methodology was reviewed and updated to consider any new guidance on the topic, including Ofcom's consultation[13]. This update allowed us to create a more nuanced and evidence-driven assessment of risks.

Our risk assessment reflects X's services at and around 30 June 2024.

A. Walkthrough

To streamline the risk assessment process further, we adopted a three-phased approach to the exercise.



Identification of risk

In step 1, we review the risk areas that need to be assessed and updated as needed.

Assessment

In step 2, we conduct the main assessment. This includes severity and probability calculations, based on collected evidence, to provide an inherent risk score, and an analysis of our controls, to assess how they mitigate this inherent risk. The residual risk is derived as a product of the inherent risk and control strength.

Potential mitigations

In step 3, based on the residual risk and on the strength of current controls, we outline possible new and continuous measures to mitigate the systemic risks.

Fig.1: Three phase process to risk assessment.

[13] 'Protecting people from illegal harm online

Phase I: Identification of systemic risks

The four systemic risks, as defined in Article 34(1), were assessed. We have streamlined the underlying assessments, recognising the overlaps between certain risk areas and in our approach towards mitigating them. As such, the assessment for the risk of sale of illegal goods and services was considered alongside the risks to consumer protection, and the assessment for the risk to the fundamental right of respect for private & family life was considered alongside gender-based violence.

Phase II: Assessment

This assessment of risk analyses (1) the inherent risk, then (2) the control strength and finally (3) the residual risk. The visual below indicates how residual risk acts as a function of inherent risk and control strength; how inherent risk is a function of probability and severity; and finally how severity can be decomposed into scope, scale, and remediability.

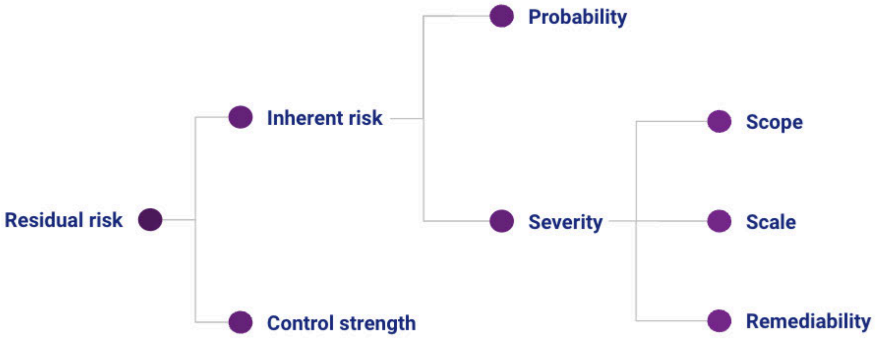


Fig.2: The risk assessment methodology

Inherent risk

Inherent risk is understood as a function of probability and severity, where the assessment of severity considers scope of harm, scale of harm, and remediability of harm.

The definition of ‘scope’ was updated to better reflect the gravity of harm when it impacts vulnerable groups, to reinforce our understanding of severity. Further, our definition of ‘scale’ was standardised to refer to the reach of the harmful content to users in the EU. This definition allowed teams to clearly identify how certain risks were disseminated in the Union, as well as delineate between the inherent risk of certain harms on the platform compared to how users experience them.

[redacted]

Fig.3: User reports under TIUC Terms of Services and Rules

The visual above, depicting volume of user reports between October 2023 to June 2024, can be used as a proxy to understand our users' perceptions of prevalence on the platform. The chart shows that the majority of user reports in the time period were for violations of the Hateful Conduct, Abuse and Harassment, and Violent Speech policies, which overlap with the illegal hate speech risk area. While this is not a perfect measure (e.g., users may not report content violative of different policies at the same rate of impressions), it can indicate that hate speech may reach users more than other risks, such as Child Sexual Exploitation content (overlapping with the risk of Child Sexual Abuse Material, 'CSAM') or violations of Violent and Hateful Entities policy (overlapping with the risk of Terrorist Content).

As such, the inherent risks were recalibrated to align with the standardised data. However, it is important to emphasise that this does not indicate that there was an *increase* in one systemic risk over another on the platform between Y1 and Y2, but rather, our update to the methodology has provided more robust understandings of how such content manifests on the platform and attempts to understand to what extent it reaches our users.

Assessment of controls

As a platform that strives to protect its community, which includes respecting the right to Free Expression and Information, we have a number of controls in place that mitigate systemic risks on our Platform. We evaluate control measures on their operationality, effectiveness, proactivity, and improvement processes. We are continually improving our testing methods and effectiveness of controls.

Identification of residual risk and tiering
Residual risks are calculated by multiplying inherent risk scores by control strength scores. We assessed the residual risk by mapping our existing mitigation measures against the identified inherent risk to showcase how these controls can, and have, already mitigated the assessed risks.

Regardless of the effectiveness of our controls, certain risks will remain, and it is a complex, ongoing and multistakeholder challenge to continuously evolve our control measures and respond to emerging threat patterns. In many of the assessed systemic risks, negligible residual risk level is potentially impossible to reach without unnecessarily restricting the use of our service and infringing on our users' fundamental rights.

Finally, we assigned risks into different tiers according to their residual risk score. We consider critical or high residual risk areas to be Tier 1 risks, medium residual risk areas to be Tier 2 risks, and low or negligible residual risk areas to be Tier 3 risks. These tiers help us prioritise our approach to future mitigations and also provide insights on areas where our current efforts are effective.

For further information on the identification of systemic risks and a detailed methodology, please refer to the Y1 summary report as well as the Annex.

Phase III: Mitigation measures

Similar to our approach in Y1, based on the results of the risk assessment, we considered measures that could be improved on, or new measures that could be implemented to reduce the residual risk of harm. As a first step, our teams took stock, among other factors, of the implementation status of all existing measures, including Y1 Article 35 mitigations and any new controls implemented over the last year, to highlight areas where work has been completed or continuous efforts are ongoing. Then, the teams identified forward-looking mitigations they could explore in order to further reduce or manage the risk areas identified in DSA Year 2. This approach is in line with the core assertions of the DSA that mitigation measures need to be reasonable, proportionate and effective, acknowledge X's economic capacity, and give special consideration to the impact on freedom of expression.

As a platform dedicated to protecting our community while respecting free speech, we have implemented several controls to mitigate systemic risks. It is important to note that we continually update and improve these measures to adapt to our growing user base.

This methodology is specific to the DSA's second risk assessment under Article 34. The results of this assessment should not be used for other regulatory or litigation purposes. Inherent and residual risk scores should be understood in context and not in isolation.

B. Stakeholder engagement and consultation

We regularly engage with stakeholders and partners in the EU as part of our continuous risk mitigation cycle. Leading up to this year's risk assessment, we consulted external and internal experts and sought input from our policy and cross-functional teams to develop a proportionate and adequate assessment, keeping in mind the special consideration to the right to freedom of expression.

Our internal stakeholder engagement included awareness sharing, training, consultations and reviews. Globally based teams involved in this process included Safety, Product Engineering, Legal, Privacy & Data Protection, Global Government Affairs (GGA), the Independent Compliance Function, and the TIUC Board. X management reviewed and approved the assessment strategy, and was actively involved in the decisions related to the risk management.

Our external stakeholder engagement – involving collaboration with governmental organisations, law enforcement authorities (LEAs), NGOs, and civil society organisations (CSOs) – takes multiple forms, including:

- **Training:** Our GGA team provides training sessions for government and non government actors. This includes presentations on the safety features of the platform, targeted training for LEAs on the functionalities and systems available to them, as well as training for NGOs and CSOs on reporting illegal or harmful content;
- **Ads credits**: This is a way for government and non-government bodies to run campaigns on X via ads. X donates a certain number of free ads credits, which can be used by the entity to ensure that their campaign reaches users. This acts as a mitigation for the spread of misinformation by promoting posts by vetted organisations and by supporting the spread of media literacy among our users;
- **Information exchange:** This is useful for notifications about threats, such as LEAs highlighting evolving threats from bad actors and campaigns, notably in the context of elections, as these are societal and multiplatform risks. For example, information received from French and German foreign ministries, following meetings prior to the elections, informed our Safety team's actions;
- **Partnerships and integrations:** Launching formalised partnerships and integrations with CSOs is a key mitigation to target cross-platform harms and improve proactivity;
- **Combating serious crime:** Engagements with EU LEAs (including Europol) have helped combat serious crime.

In response to key societal events over the last year, GGA has worked closely with governments and NGOs to mitigate systemic risks on the platform:

- Following the **October 7th attacks**, and the rise of antisemitic hate speech, GGA participated in meetings organised by the EU Internet Forum to prevent the spread of terrorist content related to the conflict, meetings by the Conseil Représentatif des Institutions juives de France (CRIF), Délégation interministérielle à la lutte contre le racisme, l'antisémitisme et la haine anti-LGBT (DILCRAH), and other NGOs. X provided ad credit grants to CRIF, allowing them to run campaigns on X to combat hate speech and antisemitism in France. X also held two roundtables with members from American Jewish Congress (AJC) and European Jewish Congress (EJC) in November 2023 and January 2024 in Brussels and in the US to establish a cooperation for any content escalations and to exchange information on keywords, behaviours, and patterns that our moderation teams should be aware of;
- X assessed, planned for, and enforced multiple elections in the EU this past year - most notably large scale elections such as the EU elections and France Legislative elections.

    - In preparation for the **EU elections,** GGA proactively engaged information with the European Commission, the European External Action Service, the European Parliament, and key authorities of the 27 EU Member States. X's work on protecting the EU elections was appreciated by the European Parliament's communication service and the EU's External Action Service (EEAS) as communication was effective during the election and escalations were promptly dealt with. X also supported media literacy campaigns with trusted partners and recognised experts in the EU, such as the European Parliament, European Digital Media Observatory (EDMO), the European Regulators Group for European Media Services (ERGA) that aimed at providing reliable information on the EU elections. GGA provided crisis response contact points to DSCs, European Commission, and European Parliament. X also presented its election's approach to Coimisiún na Meán and other DSCs and provided an overview of X's election integrity efforts. Additionally, X gave a safety training to more than 60 EU-based NGOs on how to maximise use of safety tools on the platforms and report hate speech related to elections. X also shipped product interventions in the form of home and search timeline prompts to direct people to key and official resources on how to register to vote and reminders to vote in order to encourage civic participation, as well as election hashmojis.
    - In the context of **France's Legislative elections**, GGA consulted X's NGO partners for updated lists of terms that could be considered racist or antisemitic in France. This was taken into account by internal teams in their moderation work during the elections. Viginum and Quai d'Orsay were also able to submit leads on foreign influence and attempts to impact civic processes to X's Safety team. X also provided ads credits for media literacy campaigns in the context of the elections to Generation Numerique.
    - Ahead of **the 2023 elections in Slovakia and Poland**, X proactively met with the Slovak Government, electoral commission, and law enforcement authorities in Bratislava, as well as the Polish government and electoral commission to discuss the elections.
- Recognising that major sports events have resulted in increases in abuse and harassment on online platforms, during the **2024 UEFA European Football Championship,** X participated in a proactive program with UEFA to monitor, report, and remedy cases of online abuse against players. X also collaborated to expedite key copyright reports throughout the games, and worked with UEFA to address possible violations in the platform. Following a training session with law enforcement bodies in Europe (including Europol, Interpol, Italian, French, Spanish, German, and Irish bodies) where they requested more support during the **Olympics**, X increased its staff to respond to the projected increase in volume of reports during the games. Further, X also cooperated with the International Olympic Committee and e-Enfance to preserve the safety of athletes online. In this context**,** X also provided ads credits for a public health campaign (the "manger-bouger" campaign) to the Red Cross in partnership with the French government to encourage people to practise sport 30 minutes a day to stay in good health.

We also continuously engage with stakeholders to target the following:

- Risk of **illegal content:** In February 2024, X conducted operational meetings with NGOs on how to use X's EU illegal content form. This resulted in the correction of certain technical issues that were flagged by the NGOs. X also participated in the EU Internet Forum Ministerial on the impact of generative AI (GenAI) on terrorism and child sexual exploitation. Further, X took part in the Christchurch Summit as part of the Christchurch Call for Action on Fighting Terrorism on the margins of the Paris Peace Forum;

- Risk of **hate speech:** In May 2024, X provided a training session for over 60 CSOs on online hate speech and violent content, which was attended by DG JUST. X also remains an industry member of the Online Hate Observatory in France. Further, X provides ads credits to INACH and Search for Common Ground for campaigns against hate speech and violence. Finally, X remains an industry member of the EU Code of Conduct on Countering Illegal Hate Speech and has recently signed its membership to the new Code of Conduct +, which is becoming a voluntary code of conduct under DSA Article 45;.
- Risks to **minors:** X is an active participant in the Child Protection Laboratory and attended meetings organised by the Lab in the margins of the Paris Peace Forum. X also provides ads credits to the InSafe Network, which works on the prevention of online child exploitation, and to Point de Contact and e-Enfance, which work in child protection. The partnership with e-Enfance was also for a campaign against harassment in schools. In June 2024, X also provided ads credits to Cybersmile, in the context of Stop Cyberbullying Day;
- Risks of **harassment and gender-based violence**: X provided ad credits to The Sorority for safety of women campaigns in France, as well as to GIP-ACYMA for a campaign on cyberharassment.

For further information on other stakeholders we have continued to work with, please refer to our Y1 report. As we continue to develop our process and risk management cycle, we hope to explore further stakeholder consultations to inform our risk assessment work.

VII. Summary of risk assessments

Our teams referred to EU-specific data that extended from October 1 2023 to June 30 2024, and considered enforcement on TIUC Terms of Service and X Rules violations (from here on ‘X Rules’ or ‘Rules’) [14] as well as on Article 16 DSA notices (referred to as ‘Article 16/DSA user reports’ from here on) to draw consistent conclusions across the risk assessment. Moving forward, as the timing of the risk assessment cycles align with the DSA transparency report, teams will be able to use the transparency report for consistency. The visuals below were built using the October 2023 to June 2024 data, and form the basis of our assessments.

Enforcement actions: Probability
To estimate probability, we looked into total enforcement actions[15], both automated and manual, across policy areas that aligned with the underlying assessments.

[redacted]

Fig.4: Total enforcement for TIUC Terms of Service and Rules violations

This allowed us to understand the volume of violative content and behaviour that existed on the platform and was actioned. As the pie chart shows, almost [redacted] of enforcement action is taken under the Platform Manipulation and Spam policy, indicating high volumes of inauthentic

[14] Note that while Adult Content and Violent Content policies were rolled out prior to the completion of this assessment, there was not sufficient data to be pulled from these enforcement actions. As such, data from enforcement on Sensitive Media and Violent Speech has been used for this assessment.

[15] Total enforcement data was calculated by taking the sum of total suspensions, total content removals, and an extrapolated total restricted reach labelled posts for the time period of October 2023 to June 2024. For restricted reach labelling, an estimate for the time period was used, as due to data retention issues, real figures are only available for an [redacted]. As such, these values should be understood to be estimates.

accounts and spammy activity on X. Accounts suspended under this are primarily inauthentic accounts and this is done to reduce inauthentic use of the platform.

To get a clearer understanding of the relative volume of the other harms, we excluded the Platform Manipulation and Spam figures from the data and looked into the breakdown of both suspensions and post removals).

Fig.5: Comparison of enforcement actions for TIUC Terms of Service and Rules (excluding Platform Manipulation and Spam)

The above chart shows highest content removals and suspensions for Violent Speech and Child Sexual Exploitation, with the strictest enforcement, i.e, suspensions, for Child Sexual Exploitation. This aligns with X's zero-tolerance policy towards this offence (including suspensions for accounts engaging with Child Sexual Exploitation).

A. Dissemination of illegal content

This systemic risk area considers the risk of dissemination of the following: terrorist content, illegal hate speech, child sexual abuse material, and intellectual property and copyright.

We do not allow the use of X for any unlawful behaviour or to further illegal activities, including threats or incitement of violence and terrorist content, and have a zero tolerance policy towards the dissemination of Child Sexual Exploitation. As we build our enforcement approaches, we pay due regard to their proportionality and effectiveness in addressing these violations and provide an effective appeals process for users to contest our decisions.

In comparison to Y1, the inherent risk for some of these areas increased. Simultaneously, the residual risk decreased in Y2 due to improvements in controls. Specifically for the area of IP &

Copyright, the inherent risk and the residual risk remain the same, at a low inherent risk level[16]. The following graph shows the inherent and residual risks for this area in Y2.

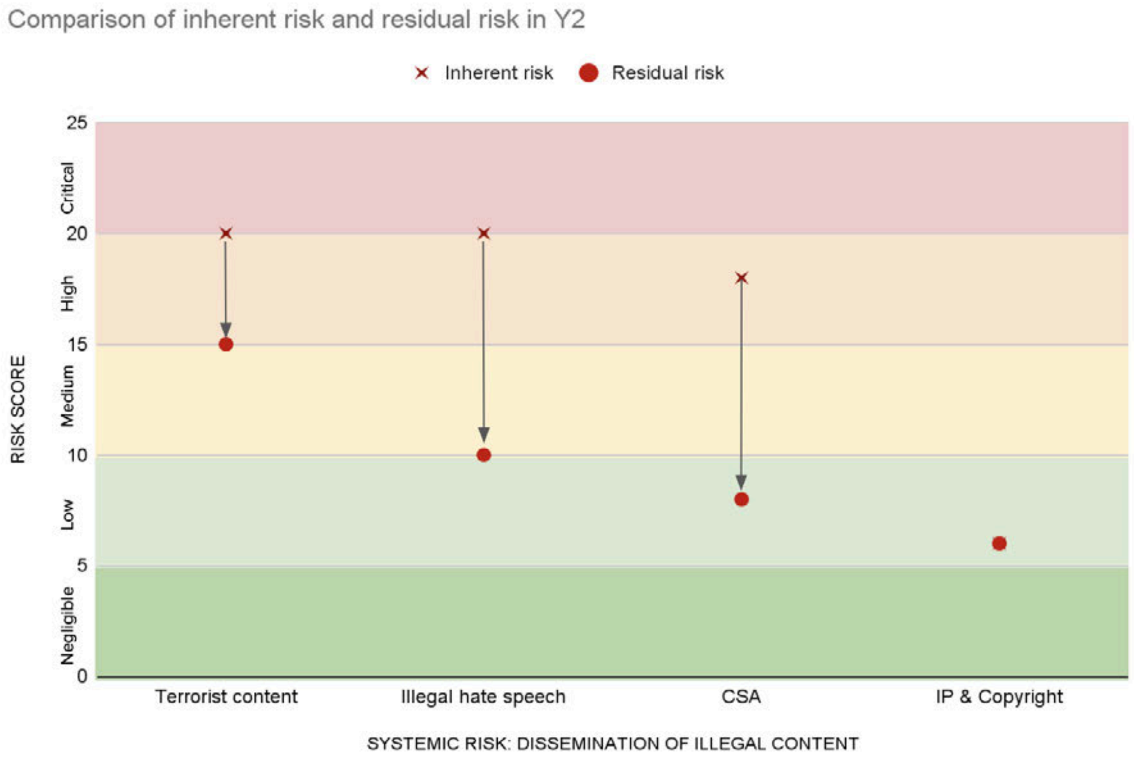


Fig.6: Comparison of inherent and residual risk for dissemination of illegal content

Inherent risks

Over the last year, political, social and cultural events have had an impact on the risk of illegal content being disseminated on X. For instance, the October 7th attacks resulted in an influx of harmful content being disseminated across social media platforms, particularly regarding terrorist content and hate speech. Further, the uptake in use of GenAI has also increased the likelihood of creation and dissemination of AI-generated content.

As discussed in the Y1 Risk Assessment report, there is always an inherent risk of bad actors misusing platforms like X and its functionalities to disseminate illegal content. We recognise that our systems are not immune to manipulation. Furthermore, features such as posting/reposting, tagging, the ability to build anonymous profiles, expanding user networks, and live streaming may be misused by actors to disseminate illegal content.

Controls to mitigate the risk of dissemination of illegal content

Policies and enforcement (Article 35(1)(b))

[16] Note that there is no separate inherent risk and residual risk marking in Figure 6 as the low inherent risk of this area has been mitigated by defined controls, and remains a low inherent risk.

As discussed in the Y1 Risk Assessment report, X continues to develop and implement robust policies and protocols to address the dissemination of harmful and potentially illegal content online. Our controls are anchored on principled policies and leverage diverse interventions to ensure that our actions are reasonable, proportionate and effective.

X Rules

X enforces its own rules to combat terrorist content, hate speech and unlawful discriminatory content, CSAM, and copyrighted materials. These policies are enforced using a wide range of measures, including content removals, account suspensions, and geo blocking of content. X's enforcement of the X Rules operates independently and is supplementary to its process allowing users to report content suspected of being illegal within the EU.

We also launched a Violent Content policy in May 2024, which consolidates two major policies: Violent Speech and Violent Media. Through this policy, X allows users to share graphic media if it is properly labelled, not prominently displayed, and is not excessively gory or depicting sexual violence. Enforcement taken under this policy is proportionate to the harm. For example, violent threats, wish of harm, incitement of violence, glorification of violence, violent sexual conduct, gratuitous gore, beastiality and necrophilia is removed from the platform and further violations may result in the account being suspended or placed on read-only mode. Lower severity harms, such as any minor or non-deliberate instances of violent speech, depictions of physical fights, or bodily fluids, are labelled and consequently have their reach restricted, ensuring that users who do not wish to see it can avoid it and that minors are not exposed to it.

Monetisation Standards

User monetisation features, such as creator ads revenue sharing, are only available to X Premium users. If these users violate our policies, X may take a range of enforcement actions, including demonetisation or account suspension. Furthermore, X conducts sanctions screening on all verified Premium users to ensure that X does not disburse payments to individuals on sanctions lists. If any users are confirmed to be sanctioned, X implements an indefinite restriction on their access to all monetisation features. Posts that have a Community Note on it are demonetised, ensuring that there is no revenue generated from false or misleading information.

Advertising Policies

X Ads policies prohibit the promotion of illegal products and services. We deploy proactive measures to ensure that our Advertisers comply with these policies, as explained in IV. X Risk Environment: Influencing Factors & Controls.

Tooling (Article 35(1)(c))

To target harmful or illegal content, X employs multiple methods to mitigate risks of dissemination of such content on the platform. For instance:

- We have ban evasion detection for accounts that have been suspended for violating X Rules.
- For CSAM and accounts belonging to violent or terrorist entities, we use [redacted] [redacted], and blocking keywords from appearing on X's Search Autocomplete and Trending Topics.
- We employ machine-learning models and business logics such as denylist terms restricting violating content from appearing on ads.

Product-level controls (Article 35(1)(a))

While all social media platforms are vulnerable to being misused for dissemination of illegal content, we recognise that certain product functionalities may pose higher inherent risks. X has a number of standing measures in place to combat this:

- **X Live:** In addition to safety detections such as media-based models for Adult Content and Child Safety (detection of the presence of a minor in live videos), there are a number of product-level protections in place to limit the risk of X Live being abused. These features allow the owner of a live video to block anyone that posts abusive or violent comments, and viewers to report abusive or violent comments allowing a reactive human review to take place.
- **Spaces:** For Spaces, controls include proactive machine learning detections for toxic Space titles, toxic content in transcription text, and Spaces associated with users determined to be high risk, in addition to reports by speakers or listeners. Spaces detected or reported are sent to manual review by content moderators to determine if they contain any violative content. Hosts and co-hosts of Spaces can block or remove abusive speakers from a Space.
- **Communities:** Posts in Communities are subject to our Safety post-level controls. In some cases, these controls are stronger in Communities. For example, Sensitive Media posts are hidden using machine learning if the Community did not correctly label themselves as Adult Content or Violent/Graphic Content. Communities also have admins and moderators who enforce Community rules and use moderator tools to maintain healthy conversations. Furthermore, any X user, whether a member of the Community or not, can report potential violations to X.

Further illegal content controls (Article 35(1)(c)&(g))

Since August 2023, X has also operated its DSA illegal content report form as well as its appeals form.

[redacted]

Fig.7: Enforcement action in the context of DSA user reports

Between October 2023 to June 2024, X has received approximately [redacted] user reports in total, and has actioned [redacted] of them, with most of the actions being geo-blocking content (known as “country withheld content”) and content removals. X assesses all user reports of illegal content against its own X Rules and if there is no violation of the X Rules warranting removal of the content, X then assesses the content for illegality under the law(s) designated by the user in their report. X also continues its work regarding trusted flaggers in the EU, including to receive and action prioritised reports.

In X’s second DSA transparency report, which looked at the period of 21 October 2023 to 31 March 2024, we found that the median time to resolve illegal content reports was 2.7 hours. Furthermore, in the same time period, of a total of 238K illegal content reports received, 115K were found to be violative – approximately 48%. Following this, X received only 667 appeals to its decisions taken on illegal content, and 190 decisions were overturned. As such, the decisions taken on illegal content have around a 0.58% appeal rate, and only 0.17% of the decisions taken are overturned, indicating a high level of accuracy in X’s determinations.

The DSA transparency report also provides insights into removal orders and information requests received by Member States’ authorities. Between 21 October 2023 and 31 March 2024, we received 13 removal orders, from France, Italy, and Spain, for unsafe and/or illegal products and illegal or harmful speech. The median handle time to resolve these orders was 4.1 hours. With regards to information requests, we received 6K requests, with the most requests concerning illegal or harmful speech (from Germany), followed by risks for public security (from France). The median time to resolve these requests was 74 hours.

At X, government, legal, and law enforcement requests are managed through our established guidelines. We provide clear procedures for law enforcement seeking account information and content removal, and may not comply with requests for a variety of reasons. For detailed information and to learn about requests to withhold content, please refer to our Guidelines for law enforcement.

Zoom-in: Israel/Hamas – Crisis Protocol

From the onset of the conflict, X activated its crisis protocol to address the rapidly evolving situation with the highest level of urgency.

This crisis triggered the formation of a cross-functional leadership team that worked around the clock to ensure that the global community has access to real-time information and to safeguard the platform for users and partners. To learn specifics on the enforcement actions taken, please refer to our blog that discusses our escalations, application of interstitials, demonetisation actions, and expansions to our proactive measures, including automatically acting against Hateful Conduct targeted at Jewish and Islamic people.

In addition to our normal defences and as part of our crisis-response protocol, between October 2023 and June 2024, X took additional proactive measures to remove or label hundreds of thousands of pieces of content.

- Across safety policies, such as Violent Speech, Hateful Conduct, and Sensitive Media, we have actioned over [redacted] pieces of violative content.
- Under our policy on Synthetic and Manipulated Media, we have actioned over [redacted] pieces of content that corresponded to Synthetic and Manipulated Media including AI-generated ones (e.g. fake images of Israeli settlers' tents in Gaza).
- In addition, we have enforced over [redacted] accounts for platform manipulation, and suspended over [redacted] accounts related to violent entities in the region - over [redacted] of which were Hamas affiliated accounts.

Moreover, we recognise that false and misleading information is a critical topic that every publisher and platform is tackling. At X, this is mitigated by Community Notes. In the first month of the conflict, notes were viewed well over a hundred million times, addressing topics from out-of-context videos to AI-generated media to claims about specific events.

Following this incident, our GGA team has also established partnerships and contacts with the American Jewish Congress, Conseil Représentatif des Institutions Juives de France (CRIF), Ligue Internationale Contre de Racisme et l'Antisémitisme (LICRA) and SOS Racisme.

The following sections focus on our assessments for each risk area and provide a summary of the results.

Dissemination of Terrorist Content

The inherent risk of dissemination of terrorist content on X arises from the potential for individuals or groups who use the platform to disseminate terrorist and extremist propaganda, recruit followers, facilitate or coordinate violent attacks, solicit funds from sympathisers, and praise, support, or glorify terror attacks.

External events and conflicts, such as the October 7th attacks and ongoing conflict in Gaza, has increased the inherent risk of terrorist content on online platforms.

Probability
Between October 2023 to June 2024, X suspended [redacted] accounts across its Violent and Hateful Entities and Violent Speech policies, and removed [redacted] posts for the same policies. These suspensions amount to only [redacted] of suspensions on the platform. While the number of content removals that violate these policies comes up to [redacted] of the total post removals in the time range, it is worth noting that all Violent Speech removals do not directly correlate to terrorist content. Based on this distinction, the probability of dissemination of terrorist content on the platform has been assessed to be *likely*.

Severity

- ***Scope***: Acts of violence which may have been coordinated via online platforms, alongside the glorification of terror attacks, may result in psychological harm, potentially inducing anxiety, fear, or panic[17]. Inauthentic accounts may rapidly disseminate terrorist and extremist information, and artificially amplify hashtags, trends or messages that align with their narratives. This leads to a *very high* scope of harm;
- ***Scale***: Although the reach of this harm is comparatively lower when considered against violations related to hate speech, user reports for Violent and Hateful Entities and Violent Speech comprised almost [redacted] of user reports between October 2023 - June 2024, indicating that the scale of this harm remains *high*;
- ***Remediability***: Given that a remedy in this situation can rarely restore the individual who experienced the harm to their state before the impact, this risk has been assessed to be *rarely remediable*;
- Based on the assessments above, the dissemination of Terrorist Content on the platform is assessed to have a *very high severity*.

Inherent risk
Based on the probability of terrorist content existing on the platform, along with the high severity, the dissemination of terrorist content on the platform is a *critical inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

[17] Protecting people from illegal harm online, p.27.

Control strength

In addition to the global controls targeting illegal content described above, specific controls targeting this risk include:

- **Article 35(1)(b) - Policies & enforcement:** X's Violent and Hateful Entities, Perpetrators of Violent Attacks, and Violent Speech policies define the enforcement of terrorist content. Our Perpetrators of Violent Attacks policy is implemented following escalations;
- **Article 35(1)(f) - Crisis response:** Our crisis response protocol is led by our Strategic Response Team, which has protocols for operating under a structured incident prioritisation plan and crisis assessment framework;
- **Article 35(1)(f) - Global Internet Forum to Counter Terrorism (GIFCT):** Through GIFCT, X is able to collaborate with industry to identify and resolve challenges, share trends and analysis, hear from civil society about their concerns and engage with experts from academia and governments;
- **Article 35(1)(f) - Christchurch Call:** X is a signatory of the Christchurch Call, and continues to collaborate with governments and civil society to fulfil the commitments made in 2019 and engages directly with the Christchurch Call's crisis protocol.
- **Article 35(1)(f) - Screening prior to monetisation:** X screens all verified Premium users enrolled in the revenue sharing program, against lists of sanctioned entities, to ensure that X does not disburse payments to individuals on sanctions lists. If any users are confirmed to be sanctioned, X implements an indefinite restriction on their access to all monetisation features.

Over the last year, further controls have been implemented and existing controls improved upon, that align with Article 35, to target this risk:

- **Article 35(1)(c) - Reporting of illegal content in the EU:** Users in the EU can report posts through a separate DSA report form accessible to all EU users, not just registered platform users. These reporting channels assist us in combatting content that violates X's Rules or is illegal in the EU;
- **Article 35(1)(b) - Policies & enforcement:** Following a policy audit, we have launched a Violent Content policy that improves upon the existing Violent Speech and Sensitive Media policies to enforce on content that threatens, incites, glorifies, or expresses desire for violence or harm, as well as visual material depicting graphic, violent, or excessively gory content including sexual violence;
- **Article 35(1)(f) - Proactive monitoring:** The number of violent entities that are proactively monitored has increased;
- **Article 35(1)(f) - Crisis response:** Our crisis response was triggered following the October 7th attacks. For more information, please refer to Zoom-in: Israel/Hamas – Crisis Protocol.

Overall, the controls for this risk are assessed to be *defined*. The measures are formalised, documented, and repeatable. Quality assurance frameworks are being implemented and processes tend to be more proactive than reactive. They are well characterised and understood across all organisation verticals.

Tier 1 priority

Due to the *critical inherent risk* of this area, which is mitigated by controls of a *defined* nature, the residual risk of the dissemination of terrorist content remains a *high risk item*, making it a *Tier 1 priority.* While the control measures are robust, the nature of the risk itself requires vigilance. We will continue to evaluate these risks and our controls as they may continue to evolve. Our efforts to continue addressing residual risk are detailed in VIII. Considerations for further mitigations.

Dissemination of Illegal Hate Speech

Given that X is a public platform, we are sensitive to the inherent risks that hate speech can pose both at an individual and a societal level. Hate speech is often targeted towards people based on their protected characteristics, and can manifest on online platforms in multiple ways, including dehumanising speech, calls for discrimination, exclusionary speech, slurs, tropes, and hateful stereotypes, and celebrating or glorifying hate crimes.

Features such as Spaces and Communities, anonymous profiles, direct messaging, and user tagging; as well as external events such as the October 7th attacks, can increase the inherent risk of hate speech on X.

Probability

Between October 2023 to June 2024, X suspended [redacted] accounts across its Abuse and Harassment, Hateful Conduct and Violent Speech policies and removed [redacted] posts for the same. Further, in the same time period, X took [redacted] actions for Illegal or Harmful Speech, following DSA user reports, which is the category with the highest enforcement within the illegal content reporting workflow. As such, we have concluded that the probability of dissemination of illegal hate speech content on the platform is *almost certain*.

Severity

- ***Scope***: Acts of hate speech may lead to targeted abuse, harassment and hate speech based on protected characteristics. While there is some potential for this to result in psychological harm, research shows mixed results when trying to identify the correlation between online hateful language and specific offline crimes.[18] Overall scope is considered to be *moderate;*
- ***Scale***: User reports for Hateful Conduct, Abuse and Harassment, and Violent Speech together resulted in almost [redacted] of user reports between October 2023 - June 2024, indicating the wide reach of this harm. In the same period, X received [redacted] user reports for Illegal or Harmful Speech, which is [redacted] of all DSA reports, and the highest volume of user reports within the DSA categories. Hence, the scale of this harm is *very high;*

[18] Cahill, M, Migacheve, K, Taylor, J, Williams, M, Burnap, P, Javed, A, Liu, H, Lu, H. and Sutherland, A, 2019. Understanding online hate speech as a motivator and predictor of crime

- ***Remediability***: If illegal hate speech is disseminated, the platform's redress mechanisms, such as suspending accounts and removing posts, can curb the dissemination. However, users who witness such illegal hate speech, especially those belonging to the targeted group, may experience some psychological distress. Despite this, platform action may mitigate most of the harm done by reducing the presence of the content. Therefore, remediability is considered to be *likely remediable*.
- Based on the assessments above, the severity of illegal hate speech is *high*.

Inherent risk

Based on the probability and severity of this risk, the dissemination of illegal hate speech on the platform is assessed to be a *critical inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls targeting illegal content described above, specific controls targeting this risk include:

- **Article 35(1)(b) - Policies & enforcement:** X's Abuse and Harassment, Hateful Conduct, and Violent Speech policies are used to enforce on instances of harmful speech on the platform, and illegal hate speech is enforced upon following illegal content EU user reports;
- **Article 35(1)(c) - Proactive moderation for violative speech[19]:** X's automated content detection tools for X Rules violations can act on both text and media, and those detections may or may not overlap with illegal hate speech laws in respective EU member state countries. We use combinations of natural language processing models, image processing models, and other sophisticated machine learning methods, as well as heuristic-based rules, to detect potentially X Rules violating content.
- **Article 35(1)(c) - Training**: We actively provide ongoing training support and mandatory refresher requirements for our frontline moderators to educate them about different types of hate speech and how they may manifest on X;
- **Article 35(1)(c) - Understanding Context:** Due to the fact that "hate speech" is very contextual and language-based, X hires content moderators with a variety of language skills to provide a comprehensive and thorough review of probable hate speech content that is reported from our users. Teams also maintain a live resource of non-English hate speech related terms and slurs in various European languages.

Over the last year, further controls have been implemented, in alignment with Article 35, that target this risk:

- **Article 35(1)(c) - Reporting of illegal content in the EU:** Users in the EU can report posts as illegal hate speech through a separate DSA report form accessible to all EU

[19] Note that automated content moderation tools enforce against our X Rules related to harmful or hateful speech. There can be an overlap with our Rules and the definitions of illegal hate speech.

users, not just registered platform users. These reporting channels assist us in combatting content that violates X's Rules or is illegal in the EU;

- **Article 35(1)(c) - Improving moderation and tooling**: On an ongoing basis, we add new slurs, harmful terms, and phrases to our operational handbook and proactive heuristics to ensure we are capturing the evolving landscape and use of language to target members of protected categories;
- **Article 35(1)(f) - Partnerships**: During the 2024 Euros, X participated in a proactive program with UEFA to monitor, report and remedy cases of online abuse. We were able to effectively review hundreds of posts throughout the tournament and take further action where needed;
- **Article 35(1)(h) - Stakeholder engagement:** X remains an industry member of the EU Code of Conduct on Countering Illegal Hate Speech and just signed its membership to the new Code of Conduct +, which is becoming a voluntary code of conduct under DSA Article 45. X is also an industry member of the Online Hate Observatory in France. Further, X provides ads credits to the INACH, and Search for Common Ground for campaigns against hate speech and violence on the platform.

Overall, the control suite is *managed,* as the control methods are repeatable and are operating effectively. Policies and guidelines are well defined, formalised and regularly managed. We provide clear guidelines to our enforcement teams and are constantly updating our policies and guidelines to reflect changes in trends. Processes are proactive, where possible.

Tier 2 priority

Due to the *critical inherent risk* of this area, which is mitigated by controls of a *managed* nature, the residual risk of the dissemination of illegal hate speech content is a *medium risk* item, making it a *Tier 2 priority*. We continue to evaluate these risks and evolve our controls. Our efforts to address residual risk are detailed in VII. Considerations for further mitigations.

Dissemination of Child Sexual Abuse Material (CSAM)

CSAM is an ever-evolving issue and can manifest in a myriad of ways online. All users, but especially children, may be impacted by the production, distribution and consumption of CSAM, or they may be groomed for sexual exploitation. It is also possible for a minor to be coerced or directed to produce self-generated CSAM or indecent imagery. Features such as anonymous profiles, direct messaging and encrypted messaging can increase the likelihood of this risk manifesting on X. Inauthentic accounts create an additional vector of harm through CSAM spam that either redirects to off-platform content or uses CSAM terms/media to get users to click links or gain followers.

Over the last year, there has been no particular incident or external circumstance that has changed the risk profile for CSAM. X enforces on CSAM under its Child Sexual Exploitation policy, and maintains a zero tolerance policy towards CSAM content, including sexually exploitative content, sexual solicitation, sex trafficking, and sexual child abuse.

Probability

CSAM is a highly adversarial area where bad actors have strong monetary incentives and are constantly probing our defences to try and redirect traffic off-site, or more rarely, posting content directly on X. Between October 2023 to June 2024, X suspended [redacted] accounts violating our Child Sexual Exploitation policy, such as by engaging with such content, and removed [redacted] posts for the same policy. As this area considers both the risk of grooming as well as of child sexual abuse, the probability ranges from *likely to almost certain***.**

Severity

- ***Scope****:* The exploitation of minors coordinated through online platforms can cause severe physical and psychological harm. Additionally, sharing such content and enabling contact between perpetrators and victims can lead to psychological trauma and retraumatisation. This content can also impact adults who view the content. This leads to a *very high* scope of harm from this risk on the platform;
- ***Scale****:* The reach of this harm is comparatively lower when considered against other types of violations, indicated by the number of user reports for Child Sexual Exploitation ([redacted] of all user reports). Therefore, this is assessed to have a *moderate* reach;
- ***Remediability****:* Since it is rarely possible to restore a minor's mental and physical well-being after the harm has taken place, this risk is considered *not remediable*.
- Based on the assessments above, the severity of CSAM content is *high*.

Inherent risk

Based on the probability and severity assessments the dissemination of CSAM on the platform is assessed to be a *high inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls targeting illegal content described above, specific controls targeting this risk include:

- **Article 35(1)(b) - Policies & Enforcement:** X's Child Safety policy captures its enforcement on Child Sexual Exploitation, which may include real media, text, illustrated, or computer-generated media - including GenAI media. In the majority of cases, users are immediately and permanently suspended.
- **Article 35(1)(f) - Hash-sharing**: Content surfacing for human review includes leveraging the hashes provided by NCMEC and industry partners. We scan media uploaded to X for matches to hashes of known CSAM sourced from NGOs, law enforcement and other platforms. Users posting known content are suspended and reported to NCMEC;
- **Article 35(1)(j) - PhotoDNA and internal proprietary tools**: A combination of technology solutions are used to surface accounts violating our Rules on Child Sexual Exploitation (which includes CSAM);

- **Article 35(1)(j) - Reporting to NCMEC:** We continue to report accounts to NCMEC when appropriate;
- **Article 35(1)(j) - Media Risk Scanning**: [redacted] as well as filter false positive hash matches. [redacted] proactively identifies, based on the context of the conversation, possible discussions of child access, child sexual abuse, CSAM, self-generated CSAM, and sextortion. This allows our platform to identify, remove and report child sexual abuse material at scale;
- **Article 35(1)(j) - Language coverage:** Our media detection is language agnostic, which minimises this risk when considering CSA media;
- **Article 35(1)(j) - Restricted high-risk terms:** X maintains a list of related keywords and phrases that are blocked from Trending and/or are blocked entirely from search results. We have since added more than [redacted] CSA keywords and phrases;
- **Article 35(1)(j) - Controls in DMs:** Content moderators are instructed to review DMs whenever there are signs of potential Child Sexual Exploitation violations happening in DMs (such as information from law enforcement or user profile signals) and media shared in DMs is proactively scanned for matches against known CSAM databases;
- **Article 35(1)(a) - Controls in encrypted messaging:** Currently, encrypted DMs are only available to users that have a Premium subscription, and Premium subscriptions are only available to users that have provided payment details. Although encrypted DMs only include text and links, and not media, there is a potential risk of grooming behaviour and sharing links to CSA material via encrypted DMs. Users can report messages for grooming/abuse, where a cryptographically validated excerpt of the text is sent to the agent for review.

Over the last year, further controls have been implemented and existing controls improved upon, in alignment with Article 35, that target this risk:

- **Article 35(1)(c) - Reporting of illegal content in the EU:** Users in the EU can report posts through a separate DSA report form accessible to all EU users, not just registered platform users. These reporting channels assist us in combatting content that violates X's Rules or is illegal in the EU;
- **Article 35(1)(f) - Proactive detection:** Improvements to our hashing detection. We now have our own internal hash list that content moderators can add media to from within our review tools. This allows us to take down content that we've seen immediately without waiting for it to make its way to shared hash libraries provided by NCMEC and industry partners.

Our blog also provides a comprehensive update on the work undertaken to tackle CSA on X. Overall, the controls for this risk are assessed to be *managed*. Our measures are well defined, formalised, and regularly managed, with repeatable quality assurance in place. There is an established process for integrating feedback to mitigate process deficiencies, and processes are proactive, where possible, for all forms of content and behaviour.

Tier 3 priority
Due to the *high inherent risk* of this area, which is mitigated by controls of a *managed* nature, the residual risk of the dissemination of CSAM is a *low risk item*, making it a *Tier 3 priority*. Nevertheless, we continue to improve our controls to protect minors and minimise harm done within the platform, especially since these bad actors are actively adversarial and constantly shift their behaviours. Our efforts to continue to address residual risk are detailed in VII. Considerations for further mitigations.

Dissemination of IP & Copyright infringing content

X's Terms of Service explicitly require that users agree not to post content that is subject to copyright or other proprietary rights unless they have the right holder's permission or are otherwise legally entitled to share the content. However, users may - in violation of our policies - share content on our services without the appropriate legal permissions.

Recently, with the ability of users utilising GenAI to produce content that may resemble existing works, it has become easier for users to post content that may incorporate the intellectual property rights of creators, including, for example,copyright rights.

Probability
Between October 2023 and June 2024, X suspended [redacted] accounts and removed [redacted] posts for intellectual property infringements. Although this is a small in scale compared to other violations, it is important to note that the features of the platforms (posts, long form posts, media sharing, and long video sharing for X premium users), mean that uploading of IP content is a risk that is likely to occur regularly, making the probability *possible.*

Severity
● ***Scope***: Intellectual property infringements result in remediable economic harm and do not necessarily target vulnerable groups, making the scope of such harm *low*; ● ***Scale***: Between October 2023 to June 2024, X received [redacted] reports for intellectual property infringements, which is around [redacted] of the total user reports received in this time. Further, this harm primarily impacts the poster and certain rights owners. As such, the scale is assessed to be *low*; ● ***Remediability***: Since the content can be removed and X can take appropriate actions to restore intellectual property rights to the owners, it is likely that owners' rights can be restored before the infringement expands. Therefore, this risk is considered to be *likely remediable*. Based on the assessments above, the severity of this harm is assessed to be *low*.

Inherent risk

Based on the probability and severity of this harm, the inherent risk of disseminating content infringing on intellectual property rights, including, for example, copyright, is assessed to be a *low inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls targeting illegal content described above, specific controls targeting this risk, :

- **Article 35(1)(b) - Diligent enforcement:** We ensure diligent and consistent enforcement of Copyright and Trademark policies to apply to content on the platform. If an X agent needs additional information when reviewing a case, they will send a message to the report(er) asking for more information, thereby ensuring that the agent has all relevant data points when reviewing the report and committing a final action on the case.
- **Article 35(1)(b) - Repeat Infringer:** The Repeat Infringer sub-policy under X's Copyright policy takes valid retractions and counter reports into account;
- **Article 35(1)(b) - Weekly policy enforcement calibration**: The Copyright agent and Copyright legal teams meet on a weekly basis to review examples of the previous week's cases for noticeable trends, discuss unique cases to ensure a standardised process of review/action, and potential policy updates;
- **Article 35(1)(c) - Notice-and-takedown process:** X has a notice-and-takedown process for copyright issues that is actively enforced for both report(er)s and the report(ed);
- **Article 35(1)(c) - Prioritised reports:** [redacted]
- **Article 35 (1)(c) - Escalations:** X has built out an internal escalation process that is based on specific variables of the user and the content being reported, to enable additional review of content flagged as violative that may warrant more added risk;
- **Article 35 (1)(c) - Preparation for risk events:** X maintains a revolving up-to-date calendar of future popular sporting/TV events to ensure sufficient agent coverage and support when applicable (i.e. additional agents during the peak hours of the event) in anticipation of potential spikes in copyright infringement caseload;
- **Article 35(1)(f) - Expert consultations**: X has copyright and trademark policy experts responsible for identifying abusers and making recommendations regarding trends of content being reported and user behaviour, in addition to having legal guidance and consultations when applicable.

Over the past year, the above controls have been continuously monitored and managed to ensure that the risk continues to be effectively mitigated. Overall, the controls for this risk are assessed to be *defined.* Mitigation measures are sufficiently defined, documented, and regularly managed. There is a set process for integrating feedback to mitigate process deficiencies.

Tier 3 priority
Due to the *low inherent risk* of this area, which is mitigated by controls of a *defined* nature, the residual risk of the dissemination of IP, including, for example, copyright infringements remains a *low risk item*, making it a *Tier 3 priority*. The control measures are robust, however we continue to evaluate and improve them to ensure their continued effectiveness given modern trends, patterns, and user behaviour. Our efforts to continue to address residual risk are detailed in VII. Considerations for further mitigations.

B. Exercise of fundamental rights

This section considers the risk of negative effects to the exercise of the following fundamental rights: freedom of expression, consumer protection, protection of minors, personal data, and other fundamental rights. The assessment of fundamental rights considers the rest of the rights enshrined in the Charter, paying special consideration to the right to life, human dignity, and equality, right to liberty and security of a person, right to non-discrimination, and freedom of peaceful assembly and association.

We believe that X is a platform where users can express their opinions and ideas freely without fear of censorship. Simultaneously, it is our shared responsibility to ensure the safety of our users from content that violates our Rules. Therefore, as we develop our enforcement strategies, we strive to balance the protection and freedom of our users.

The inherent risk for some of these areas increased this year, whereas improvements in our controls resulted in a reduction in the residual risk. The following graph shows the inherent and residual risks for this area in Y2.

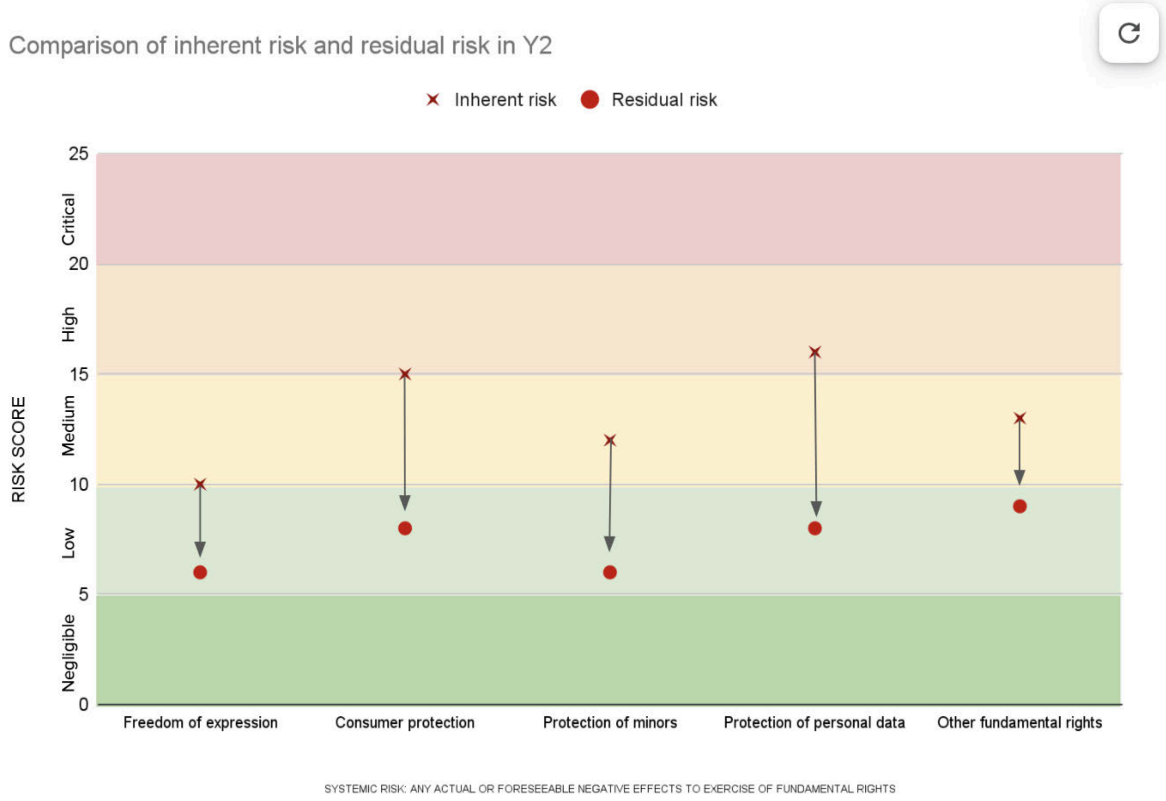


Fig.8: Comparison of inherent and residual risk for fundamental rights

Inherent risks

As a digital public town square, users come to the platform everyday to discuss and engage in conversation. However, there is always an inherent risk, on X as with other platforms, that actors or users can intentionally or unintentionally infringe on other individuals' fundamental rights. Although X as a platform is not directed to minors, minors over the age of 13 are allowed on the service and there remains an inherent risk that they may be exposed to harmful content. Noting that minors are more vulnerable than adults, features such as DMs, user network expansion recommendations, a recommender feed and anonymous profiles may act to exacerbate certain risks. For more information on the inherent risk to fundamental rights, please refer to our Y1 report.

Controls to mitigate the risk to fundamental rights

Policies & enforcement (Article 35(1)(b))

X enforces on a range of violative content, which spans across content that could hinder another user's free expression (such as abuse-related content); harm consumers (such as the selling of drugs or firearms on the platform); suicide or self harm related content; as well as content and conduct that could harm minors. With regards to personal data, X has robust internal policies to ensure that user data is protected, in compliance with the EU GDPR.

These policies are enforced using a wide range of measures, including content labelling, restrictions, removals, and account suspensions for severe violations or repeat infringements.

Aligned with the DSA, we value diligent, objective, proportionate and reasonable procedures, offering users the right to appeal content moderation decisions. Our amnesty policy occasionally reinstates accounts suspended for a specific subset of low-severity violations (e.g., we would never provide amnesty for accounts suspended for Child Sexual Exploitation), balancing user safety with freedom of expression. This aligns with the DSA's focus on avoiding unnecessary service restrictions and considering the impact on freedom of expression and information when making enforcement decisions. Requests from governments and law enforcement authorities are reviewed for compliance with international human rights and legal standards.

Zoom in: Transparent restricted reach labelling

We have invested in developing a broader range of remediations, with a particular focus on education, rehabilitation and deterrence through implementing the freedom of speech not reach approach - our enforcement philosophy which means, where appropriate, restricting the reach of posts that violate our policies by making the content less discoverable - using transparent restricted reach labels.

All content moderation systems are susceptible to certain inherent risks, as outlined in IV. X Risk Environment: Influencing Factors & Controls. As such, false positives and false negatives may occur with restricted reach labelling, which forms a part of our suite of remediations alongside suspensions and content removals. In the case of fundamental rights, false positives - where an action is taken when it should not be - could result in unfair restrictions on non-violating users.

Expanding our enforcement options to include this restricted reach labelling has allowed us to make progress in balancing the safety of users while protecting freedom of speech and being transparent in our enforcement actions. We strive to strike this balance by continuing to remove posts that harass, abuse or share hateful content directed towards specific individuals and protected groups, as we believe such targeted harassment violates individual fundamental freedoms.

Our community has provided valuable feedback to help us make meaningful changes to the accuracy of our label application, such as identifying instances where reach was not appropriately restricted and improving recognition of context in our detection. We proactively seek to prevent ads from appearing adjacent to content that we label. Users are also made aware of any restricted reach implemented against their content and are given the ability to submit an appeal if they disagree with our enforcement decision.

Regular studies conducted over the past year have shown consistent results when looking at impressions on content with restricted reach labels versus healthy posts from the same author. The restricted reach posts have had a [redacted] reduction in impressions and analysis over time has shown the impression reduction consistently stays in this range.

Data from April 2024 to June 2024[20] shows that of the posts that received a restricted reach label, only [redacted] were appealed. Less than half of these appeals were overturned, indicating that approximately [redacted] of these labels were incorrectly applied. We continue to work towards improving the accuracy of our labelling, and communicate to users when such labels are applied for X Rules violations to ensure that they can seek redress effectively.

[redacted]

Fig. 9: Comparison of enforcement action for TIUC Terms of Service and Rules

As seen in the visual above, our restricted reach labelling is primarily used for Hateful Conduct. This is in line with our belief that users have the right to freedom of expression, and we continue to restrict the reach of toxic content to maintain a healthy community online.

Nevertheless, we recognise that certain behaviours are unacceptable and use other enforcement measures in those cases. In instances where content or conduct is considered abuse, harassment, and violence, we remove content or suspend accounts, depending on the severity of the violation. We have policies in place to take strong enforcement action against and remove illegal content, including CSAM and terrorism content. Production and publication of such content results in suspension from the platform following the first offence.

Product-level controls (Article 35(1)(a))

At a product level, X provides a suite of tools designed to help our users control what they see on X and what others can see about them on X, so that they can express themselves on X with confidence. Find out more about how to control your X experience here and our safety and

[20] Due to data retention issues, we are only able to extract data for restricted reach for [redacted]. To show a comparison on real figures, all policies here are compared on the same time frame.

security tools here. X continues to be a leading player in the industry by open-sourcing its recommendation algorithm to allow feedback from the community.

Controls for minors (Article 35(1)(j))
X is rated for ages 17+ in iOS App Store, meaning that children with the correct date of birth in their App Store will not be able to download the X app. We prohibit content jeopardising minors' safety. We use content labels and interstitials to minimise exposure to sensitive content. We have also implemented age-gating mechanisms and age-appropriate reporting channels for underage users.

For further information on our controls for this systemic risk, please refer to our Y1 report. The following sections provide insight into our assessments for each risk area related to fundamental rights and provide a summary of the results.

Freedom of expression

Abuse and harassment, hateful conduct, violent speech and privacy violations can result in risks to freedom of expression, through harms such as censorship resulting from enforcement of platform policies as well as self-censorship from users who experience abuse and harassment on the platform. Further, inauthentic manipulation of information by government and non-state actors with the intention to control the information space, off-platform coordination to boost engagement and manipulate organic trends, as well as instances of mass reporting with the intention to trigger disproportionate enforcement can increase this risk.

Probability
Between October 2023 and June 2024, X suspended [redacted] accounts for violations related to Abuse and Harassment, Hateful Conduct, and Violent Speech policies, accounting for [redacted] of all suspensions. Additionally, X removed [redacted] posts for the same violations, representing [redacted] of all removed posts. Although not all these actions directly relate to freedom of expression, they may be understood as offences that could result in self-censorship or other kinds of suppression of speech. Consequently, the probability of this harm has been deemed *almost certain***.**

Severity

- ***Scope***: The scope is considered *moderate* as there is no clear risk of physical and/or psychological harm. However, this harm may impact vulnerable groups;
- ***Scale***: Over the past year, X has made changes to its enforcement policies to ensure that mitigations are proportionate and that X is not unnecessarily suspending accounts. Between October 2023 to June 2024, excluding Child Sexual Exploitation and Platform Manipulation and Spam related violations[21], account suspensions accounted for

[21] For CSAM, given the severity of the violation and X's zero tolerance policy, suspensions are used. For Platform Manipulation and Spam, given that it is a behaviour related violation rather than a content related violation, suspensions are used. Platform Manipulation and Spam suspensions are mainly directed at inauthentic accounts. As such, these two policies were excluded from the calculation.

approximately [redacted] of all enforcement actions, and post removals accounted for approximately [redacted] of all enforcement actions. The reach of harms associated with freedom of expression, especially stemming from X as a service itself, is thus assessed to be *low,* notably when considering suspensions. This is because in the majority of cases, users can remain on the platform and can thus exercise their right to free expression;

- ***Remediability***: All enforcement actions are appealable, this includes X Rules appeals and a dedicated DSA appeal form for EU users. Further, content, DMs, and ads are reportable in cases where such user-generated content may impact another user's freedom of expression. As such, this risk is considered *remediable*;
- Based on the assessments above, the risk for freedom of expression on the platform is assessed to have a *low* severity.

Inherent risk

Even though the probability of the risk is almost certain, this is offset by a low severity of harm. As such, the inherent risk is assessed to be a *medium risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls to protect fundamental rights described above, specific controls targeting this risk include:

- **Article 35(1)(a) - Anonymity**: X allows anonymity and we consider the right to anonymity not just a part of the right to privacy, but an inherent part of the right to freedom of expression. The ability to create anonymous or pseudonymous accounts provides a sense of safety and security to people who may otherwise be afraid to speak truth to power, or challenge the status quo (e.g., whistle-blowers, dissidents, members of marginalised or at-risk communities, activists etc.);
- **Article 35(1)(c) - Reporting X Rule violations:** If abusive behaviour or harassment happens, X makes it easy for users to report such instances;
- **Article 35(1)(c) - Appeals**. Users can submit an appeal if they believe that an incorrect enforcement action has been taken against their account. If the appeal is successful, we will reverse the decision where possible.

Over the last year, further controls have been implemented and existing controls improved upon, that align with Article 35, to target this risk:

- **Article 35(1)(c) - Reporting illegal content in the EU:** EU users can also report content that violates laws in designated EU countries, or across the entire EU. These reports may lead to an account or an account's content being reviewed by content moderators. There is a risk that users may seek to silence other users by repeatedly reporting their content for violating EU laws, though we do not see this arising in practice. Any harm to a user from mass reporting of their content would also have to coincide with a human making a moderation mistake or displaying personal bias.

- **Article 35(1)(i) - Improved transparency:** We aim to provide meaningful transparency on our enforcement policies and actions, including through notice to our users on our enforcement actions, when and how policies are updated through our Help Centre articles and @Safety handle, how potential violations can be reported and reviewed, when enforcement actions happen, and pathways for user appeals. We produce global transparency reports, alongside biannual DSA transparency reports, that cover a wide range of metrics. We do this so that our stakeholders can understand how X's commitment to safety has evolved over time and to shine a light on the areas where different governmental agencies may be infringing on users rights to free expression.
- **Article 35(1)(a) - Improvements to Community Notes**. We have invested in tools such as Community Notes, which allow people on X to collectively add helpful, informative context to potentially misleading posts. This is an opportunity for our users to provide more information rather than removing the content that may be considered to be making a misleading claim. For information on improvements over the past year, refer to Zoom in: Community Notes.
- **Article 35(1)(c) - Proportionate enforcement:** Restricted reach labels (under our freedom of speech not reach enforcement philosophy) can now be applied by content moderators following user reports. This allows for more proportionate enforcement action on user reports as well as more consistent application. X users have the right to express their opinions and ideas without fear of censorship.

Overall, the controls for this are assessed to be *managed*. Our policies and enforcement protocols have been created in a manner that prioritises protecting physical safety as the most important consideration. We strive to strike an appropriate balance between safeguarding privacy and enabling free expression. The measures are well defined, documented, and regularly managed. There is an established process for integrating feedback to mitigate process deficiencies, and processes are proactive, where possible.

Tier 3 priority

Due to the *medium inherent risk* of this area, which is mitigated by controls of a *managed* nature, the residual risk to freedom of expression is a *low risk item,* making it a *Tier 3 priority.* However, we continuously evaluate the situation to adapt to changing risks.

Consumer protection

Risks to consumer protection can stem from the sale of illegal goods and services, counterfeits and brand impersonations, financial scams and deceptive, misleading or harmful ads. Illegal goods and services can range from sales of drugs and firearms, to sexual solicitation. Certain features such as anonymity, the potential to reach or connect with wide audiences, direct messaging and Communities, can be leveraged by bad actors to increase this risk. Given that bad actors in this space are engaged in this behaviour with intent, tools, tactics, and given that procedures can change at any time, X's external facing policies are potentially susceptible to being intentionally circumvented.

Over the last year, there have been no particular instances that have changed the risk profile for this section; however, the risks stemming from sale of illegal goods and services have also been considered under consumer protection this year, resulting in changes to our assessment.

Probability
Between October 2023 to June 2024, X removed [redacted] posts for violations of its Illegal or Certain Regulated Goods or Services policy. This amounts to [redacted] of total post removals. It also actioned [redacted] illegal content reports of Unsafe and Illegal Products, primarily via geo-blocking, amounting to [redacted] of the total action taken following illegal content reports. Although enforcement actions against counterfeit goods and financial scams are routinely conducted, the tactics used by bad actors on X continuously change. Therefore, we see the probability of this risk arising on the platform as *likely to almost certain.*

Severity
● ***Scope****:* The scope of harm ranges from *low* to *very high* based on the various sub-risks. For example, while fraud and financial service offences may tend to only result in economic harm, the sale of drugs and firearms, if successful, has the potential for physical and/or psychological harm; ● ***Scale****:* Between October 2023 to June 2024, X received [redacted] reports in total for Scams and Fraud and Unsafe and Illegal Products, under its DSA reporting form, which amounts to [redacted] of total reports received. As such, the scale of harm ranges from *low* to *moderate*; ● ***Remediability***: It is possible that enforcement may help to restore the person to the state before the impact, especially if enforcement takes place before a sale/solicitation happens. Compensation in many cases will be capable of restoring the user to the state before the impact, though we assume that paid compensation is not possible in a large proportion of cases. As such, remediability for this harm ranges from *possibly remediable* to *likely remediable*; ● Based on the assessments above, the severity of this ranges from *low* severity to *high* severity.

Inherent risk
Based on probability and the range of severity of this harm, which is averaged, the inherent risk to consumer protection is calculated to be a *high inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength
In addition to the global controls to protect fundamental rights described above, specific controls targeting this risk include: ● **Article 35(1)(e) - Misleading ads:** The X Ads policies prohibit deceptive and fraudulent

content in ads. When advertisers opt to promote their content using X Ads on the platform, their accounts and content undergo a review process to ensure quality and safety standards. We utilise a combination of machine learning algorithms and human reviews to verify that advertisers adhere to our advertising policies;

- **Article 35(1)(c) - Proactive and reactive moderation on ads:** X's Ads policies are enforced both proactively and reactively by human reviewers who conduct proactive sweeps for violative content, review potentially violative content flagged by automated systems, and assess user and Article 16 reports;
- **Article 35(1)(c) - Market-specific language resources for enforcements:** For language-related issues that come up during responses to reported content, content moderators have guidelines they can follow to provide answers in line with linguistic and cultural standards and norms;
- **Article 35(1)(a) - Consumer protection features:** X has features that aim to protect users from harm, such as authenticity challenges;
- **Article 35(1)(c) - Restricted reach, rate limiting and unsafe URL detection:** These features work to reduce the impact of misleading activity, including malicious URLs, on the platform by reducing impressions and limiting user exposure to such content;
- **Article 35(1)(c) - Reporting mechanisms for ads:** Users can report ads for deceptive and fraudulent content and illegal products and services through in-app reporting or X Ads web form;
- **Article 35(1)(c) - Reporting of illegal content in the EU:** Users in the EU can report posts through a separate DSA report form accessible to all EU users, not just registered platform users. These reporting channels assist us in combating content that violates X's Rules or is illegal in the EU;
- **Article 35(1)(c) - Country-withheld content:** Following an DSA user report in the EU, if the report If we receive is not a violation of our Rules but is illegal in a certain jurisdiction, the content may be withheld in the relevant jurisdiction, limiting its reach.

Over the last year, further controls have been implemented and existing controls improved upon, that align with Article 35, to target this risk:

- **Article 35(1)(a) - Community Notes:** Users can help provide context and warnings to other users if they identify misleading information or third-party links that may be unsafe, including those that may attempt to scam users. For information on improvements over the past year, refer to Zoom in: Community Notes.
- **Article 35(1)(f) - Interdepartmental cooperations:** Safety has established a cooperation with the Global Content Partnerships team (X team that acts as consultants for major publishers on the platform) to initiate tickets when high profile events that will likely include digital counterfeit campaigns are coming up;

Overall, the strength of the controls for this risk are assessed to be *managed*. For counterfeit and financial scams violations, there are functioning enforcement capabilities, with well defined and documented policies. Additionally, there are avenues to escalate edge cases and adjust training materials and policies based on those escalations. There is an established process for integrating feedback. Based on operations feedback, how to action the selling of counterfeit currencies was included in training materials as being a likely scenario to take place on the platform.

Tier 3 priority
Due to the *high inherent risk* of this area, which is mitigated by controls of a *managed* nature, the residual risk to consumer protection is a *low risk item*, making it a *Tier 3 priority*. Consumer protection necessitates constant supervision and adaptable measures due to the evolving nature of the offence. Our ongoing efforts to address the residual risk are outlined in VII. Considerations for further mitigations.

Protection of minors

X is not a service that is directed primarily to children and it is listed as an app for 17+ on the iOS App Store, meaning known minors will not be able to download the app. According to our Terms of Service, an individual must be at least 13 years old to create an account, and a date of birth is required to access certain content. For users who are over the age of 13 but under the age of GDPR consent in the member state where they reside, X has built an additional workflow permitting such users to create an account with their parent or guardian's consent.

However, X is a real-time global information service, with some users (including minors) accessing the platform without logging into an account or by circumventing the age gate with false information. For online platforms, there are inherent risks that minors become exposed to harmful and violative content including bullying, harassment, non-sexual abuse, graphic violent and/or sexual content, as well as content about self-harm, eating disorders, and suicide. Over the last year, there has been no particular incident that has changed the risk profile of this harm.

Probability
As of metrics from August 2024, X's internal figures showed that 0.98% of EU-based X account holders were minors. As a result of mandatory age gates, the proportion of EU account holders without an age attributed to their account stands at 6.3%.[22] However, given that this is based on self-declaration, it is possible that the number of minors on the platform are higher. Between October 2023 to June 2024, X actioned [redacted] user reports for 'Protection of Minors' under the DSA illegal content reporting. This comes up to around [redacted] of all actions taken following DSA illegal content reports. Based on this, the probability of minors encountering such content has been assessed as *possible*.

Severity
● ***Scope***: As minors are a vulnerable group, they are more likely to experience any negative or potentially harmful content or behaviour on the platform in a more severe manner. Exposure to content encouraging or promoting self harm, violent or graphic media, and non-sexual abuse may result in physical harm and psychological distress.

[22] Based on logged-in average monthly active recipients of the service ("AMARS"). This is directionally aligned with external figures, which suggest that minors 13-17 represent 2.4% of global account holders.

Self-harm content, even if it is recovery focused content, may be upsetting or triggering. As such, the scope of harm is assessed to be *high*;

- ***Scale:*** Between October 2023 to June 2024, under the DSA illegal content reporting, X received [redacted] reports for Protection of Minors, which constitutes [redacted] of the total reports under Article 16. The reach of this item is comparatively lower as children are not X's primary demographic. Therefore, the scale is assessed to be *moderate*;
- ***Remediability:*** Given that a remedy in this situation typically cannot restore the minor to their previous state, this risk has been assessed as *not remediable.*
- Based on the assessments above, the severity of the risk is *high*.

Inherent risk

Given the probability and severity of this harm, this offence is assessed to have a *medium inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls to protect fundamental rights described above, specific controls targeting this risk include:

- **Article 35(1)(b) - Comprehensive abuse policies**: Our dedicated Child Safety policy covers content and behaviour that impacts minors the most, such as Child Sexual Exploitation, Physical Child Abuse Media, and Media of Minors in Physical Altercations. Policies to protect rights to privacy and prohibitions on content that encourages suicide and self-harm are also applicable to protection of minors.
- **Article 35(1)(a) - Default settings for logged-out users:** Permitting users to access X content without logging into an account is fundamental to X's mission to help ensure freedom of expression and access to information of its users. To mitigate risks stemming from this, X sets high privacy, safety and security settings for users who access X without an account, including the inability to view sensitive media and only displaying ads that have been tagged as "family safe". Attempting to view non-verified accounts or accounts under a threshold level of engagement while logged out redirects users to the login screen. Content that can be accessed is age-gated with a non-dismissable interstitial if it has been labelled as sensitive by the account or our systems;
- **Article 35(1)(a) - Default privacy and security settings:** All new EU users signing up to the service for the first time have, by default, personalisation turned off (including personalisation of ads, personalisation based on inferred identity, personalisation based on places you've been). All users also have direct messages defaulted to protected, meaning that only accounts that follow them can message them;
- **Article 35(1)(a) - Encrypted DMs:** Encrypted DMs are only available to X Premium users, who mainly have a paid subscription, meaning that minors are less likely to access them. Furthermore, Encrypted DMs can only include text and links; media and other attachments are not supported yet, meaning that it is less likely to be used for sextortion or other behaviour that is harmful to minors;

- **Article 35(1)(j) - Security features for minors:** We age-gate sensitive content to limit exposure to minors and allow users to report suspected underage accounts. We also have parental reporting, minimum age, and GDPR consent features that apply to minors;
- **Article 35(1)(d) - Restricted recommendations:** X implements eligibility requirements before it recommends content and accounts. Neither the Following nor the For You Timelines permit sensitive content or inappropriate advertising to be surfaced for accounts of known minors;
- **Article 35(1)(j) - Age inference:** For user accounts without an assigned age, age is inferred to help prevent minors seeing inappropriate ads;
- **Article 35(1)(i) - Support messages:** X prompts safety resources and support messages when users search for content related to self harm and suicide.

Over the last year, further controls have been implemented and existing controls improved upon, that align with Article 35, to target this risk:

- **Article 35(1)(e) - Limits to targeted advertisement:** X does not serve any ads to users under the age of 18 in the EU, as of August 2023. Logged-out users are also not served ads.

Overall, the control strength is assessed to be *managed*. We have sufficiently comprehensive control measures. There are usable reporting mechanisms, enforcement teams and proactive efforts for all X Rules at work here. X's policies and enforcement guidelines are clearly defined and thorough. Policies address key risks that harmful content poses on the platform, and have been drafted after careful deliberation with internal and external stakeholders. We provide clear guidelines to our enforcement teams when it comes to the content review process. This area (similar to all other policies) often requires further clarification from our agents and we are constantly updating our policies and enforcement guidelines to reflect changes in trends.

Tier 3 priority

Due to the *medium inherent risk* of this area, which is mitigated by controls of a *managed* nature, the residual risk of protection of minors is a *low risk item*, making it a *Tier 3 priority*. As with other risks, this risk necessitates constant supervision and adaptable measures. Our ongoing efforts to address the residual risk are outlined in VII. Considerations for further mitigations.

Protection of personal data

X is a platform that aims to foster communication all around the world. As a result, it processes personal data. This may entail potential inherent risks for the protection of personal data and the exercise of the right to privacy. This could include, for example, users' personal data being processed in ways that exceed their expectations, private information being published on the platform without proper authorisation or X being subject to security incidents that could potentially expose users' private information.

A failure to maintain products, tools, and processes that promote user privacy and enable users to exercise their privacy rights could create inherent risks for this fundamental right. Over the last year, there has been no particular incident that has changed the risk profile of this harm.

Probability

Between October 2023 to June 2024, X suspended [redacted] accounts for violations relating to Private Information and Media and removed [redacted] posts for the same policy. This amounts to [redacted] of all removed posts. Additionally, between October 2023 and June 2024, X has conducted [redacted] privacy reviews and [redacted] data protection impact assessments (DPIAs) to ensure privacy and data protection is upheld across the platform. Without any privacy and data protection controls, the probability of this harm is assessed to be *likely*.

Severity

- ***Scope**:* Without effective risk management, data could be processed in a manner that does not ensure appropriate security and confidentiality, leading to data loss and/or a data breach. This would lead to critical privacy risks and have a significant impact on users and their trust in X to handle their personal data, which could result in psychological distress. As such, the scope of the risk is determined to be *high*;
- ***Scale***: Between October 2023 to June 2024, X received [redacted] reports for Data Protection & Privacy Violations through the DSA illegal content reporting channel, and around [redacted] reports for violations of the Private Information and Media policy. These correlate to [redacted] of all DSA reports, and [redacted] of all policy violation reports respectively. As such, the reach of harm is assessed to be *moderate*;
- ***Remediability**:* Given that a remedy in this situation can often restore the individual who experienced the harm to their state before the impact, this has been assessed to be *possibly remediable*.
- Based on the assessments above, the severity of the risk to personal data is *high*.

Inherent risk

Based on the probability and severity assessments, the risk to the protection of personal data has a *high inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls to protect fundamental rights described above, specific controls targeting this risk include:

- **Article 35(1)(b)&(d)) - Compliance with privacy laws**: We uphold user rights in compliance with EU privacy laws and have a comprehensive privacy, data protection and security program. In compliance with both the GDPR and the DSA, our privacy program ensures that recommender system parameters - and how to modify them - are clearly explained to users;

- **Article 35(1)(f) - Data Protection Impact Assessment (DPIA):** In the instances where a project is deemed high-risk to the rights and freedoms of individuals, X conducts a DPIA, which requires the completion and sign off from the Global Data Protection Officer (DPO) prior to its launch;
- **Article 35(1)(f) - Regular privacy audits:** We conduct risk assessments and biannual external audits on our privacy and data protection related control environment;
- **Article 35(1)(e) - Ads**: X does not present ads to users based on profiling using special categories of data;
- **Article 35(1)(d) - Privacy reviews on recommender systems:** We have continued to conduct privacy reviews to ensure that recommender systems remain compliant with personal data requirements.

Over the past year, the above controls have been continuously monitored and managed to ensure that the risk continues to be effectively mitigated. Notably, our 2023 privacy audit found that our Privacy and Information Security Program is comprehensive in that it provides sufficient coverage across all relevant privacy and information security domains and is in alignment with the ISO 27701 and ISO 27001/02 frameworks upon which the Program is based.

Overall, the control strength is assessed to be *managed*. X maintains a comprehensive and effective set of technical, administrative and operational privacy and data protection controls. There is an established process for integrating feedback and processes are proactive, where possible.

Tier 3 priority

Due to the *high inherent risk* of this area, which is mitigated by controls of a *managed* nature, the residual risk to protection of personal data is a *low risk item*, making it a *Tier 3 priority*. Nevertheless, we will continue to evaluate these risks and our controls as they may continue to evolve. Our efforts to continue to address residual risk are detailed in VII. Considerations for further mitigations.

Other fundamental rights

Content moderation on online platforms can inadvertently replicate and amplify offline biases and patterns of discrimination based on protected characteristics. Additionally, exposure to content related to self-harm, violence and its glorification may cause psychological harm, impacting the right to life, human dignity, and equality. Features of the platform can be leveraged to infringe on these rights, including mass reporting accounts to trigger disproportionate enforcement as well as using direct messaging to harass users.

Following the October 7th attacks, there was an increase in antisemitic, Islamophobic, and anti-Arab sentiments worldwide. Such content has the potential to infringe on the right to non-discrimination of users. While all fundamental rights can be considered equal, we are aware that these rights may sometimes be in conflict. In such cases, we prioritise protecting physical safety as the most important consideration and strive to strike an appropriate balance between safeguarding privacy and enabling free expression.

In alignment with the fundamental rights considered, this assessment pays particular consideration to the risks of encouraging or assisting suicide, harms related to unlawful immigration and human trafficking, and harassment, stalking threats, and abuse offences.

Probability

Between October 2023 to June 2024, X suspended [redacted] accounts and removed [redacted] posts for violations related to Abuse and Harassment, Hateful Conduct, Suicide and Self-harm, Violent and Hateful Entities, and Deceased Individuals policies. While these violations also overlap with other risk areas, they may directly or indirectly pose a risk to user's fundamental rights.

Severity

- ***Scope***: The possible harms of the sub-risks included here encompass physical, psychological, and societal harms. For example, advocacy of hatred could incite hostility and violence resulting in coordinating physical or psychological harm on the platform. Content shared on X may exacerbate, encourage or coordinate discrimination against specific individuals, vulnerable groups or businesses. Exposure to such discriminatory content can indirectly harm an individual's physical or psychological safety. As such, the scope of harm ranges from *high* to *very high*;
- ***Scale***: Between October 2023 to June 2024, X received more than [redacted] user reports for Abuse and Harassment ([redacted] of all reports), indicating the high reach of this content. However, of all reports received in this time, only around [redacted] related to Suicide and Self-harm. As such, the scale of harm here ranges from *low* to *high*;
- ***Remediability***: While for certain sub-risks, such as online harassment, the victim may be able to be restored to state before impact; for more serious offences, especially those causing physical or psychological harm, this is not possible. As such, the remediability of this harm ranges from *likely remediable* to *not remediable*;
- Based on the assessments above, the severity of the risk to fundamental rights is *high* severity.

Inherent risk

Based on the probability and severity assessments, the inherent risk for this harm is *medium inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls to protect fundamental rights described above, specific controls targeting this risk include:

- **Article 35(1)(b) - Policies & enforcement:** X has a range of policies that relate to protecting fundamental rights, including but not limited to Abuse and Harassment, Hateful Conduct, and Suicide and Self-Harm. These policy domains are considerably complex, often requiring further clarification from our content moderators. The policy, operations and product functions work together to simplify and train our content moderators to ensure we're taking action accurately and in a consistent manner.
- **Article 35(1)(c) - Doxxing**: X takes proactive measures for doxxing – this includes a heuristic rule that continuously searches for potential instances of doxxing in content, such as addresses and phone numbers, that are shared with abusive intent. The heuristic rule surfaces [redacted] for review and action globally. Our escalations team also proactively searches for violative content on the platform with certain keywords and hashtags within a given period;
- **Article 35(1)(e) - Ads**: X ensures that ads are not presented to users based on profiling using special categories. We also provide transparency about how ads are selected and delivered to users with our "why this ad?" functionality;

Over the last year, further controls have been implemented and existing controls improved upon, that align with Article 35, to target this risk:

- **Article 35(1)(c) - Proactive enforcement:** We continue to take proactive efforts to mitigate online harassment. These measures are tailored to global events and crises, and deployed as needed. Over the last year, this has included the use of heuristic rules for sporting events as well as alerts for additional detection for targeting of politicians during the EU elections.
- **Article 35(1)(f) - Partnerships:** Our collaboration with UEFA during Euro 2024, which was a mitigation for illegal hate speech, also acts as a mitigation to protect other fundamental rights such as the right to non discrimination.
- **Article 35(1)(a) - Streamlined reporting flows:** We have updated the reporting flow to ensure users take fewer clicks to report harassment. This eases the burden on the user to ensure a swift and seamless reporting experience .
- **Article 35(1)(c) - Improved moderation workflows:** We have improved our internal workflows to ensure more accurate routing of user reports to the correct teams for reviews – this has resulted in swiftly addressing any instances of harassment.

Overall, mitigation measures are assessed to be *defined.* Measures are documented, formalised and repeatable. Processes are proactive, well characterised and understood across all organisation verticals. The rights included in this assessment cover a wide range of issues and policy areas. We believe that we have the necessary and proportionate policies and enforcement protocols in place to address the risks and impact.

Tier 3 priority

Due to the *medium inherent risk* of this area, which is mitigated by controls of a *defined* nature, the residual risk to other fundamental rights is a *low risk item*, making it a *Tier 3 priority*. However, we continually monitor the situation and adjust our controls as needed. Our ongoing efforts to address residual risks are detailed in VII. Considerations for further mitigations.

C. Democratic processes, civic discourse, electoral processes, and public security

This systemic risk area considers the risk of negative impact to democratic processes, civic discourse, electoral processes and public security.

X provides opportunities for participation in democratic processes by allowing people to access information, express their views and organise within civil society. X also enables people to directly engage on important topics with their elected representatives, candidates, and fellow citizens. Nonetheless, the influence of social media platforms also means that they may pose risks if they affect public trust in institutions, the ability for people to participate freely in the public square, organise peacefully, or generally exercise their fundamental and political rights. These values and capabilities are the bedrock of any democracy. Broadly defined, the public security risk includes threats that have the potential to undermine social order, disrupt civil harmony, and compromise the safety of individuals and communities. That said, the relationship between harmful messaging on the platform and offline action is complex and causation is difficult to ascertain.

In comparison to Y1, the inherent risk for this area has not changed, however, the residual risk has decreased, as a result of improvements in the control strength. The following graph shows the inherent and residual risks for this area in Y2.

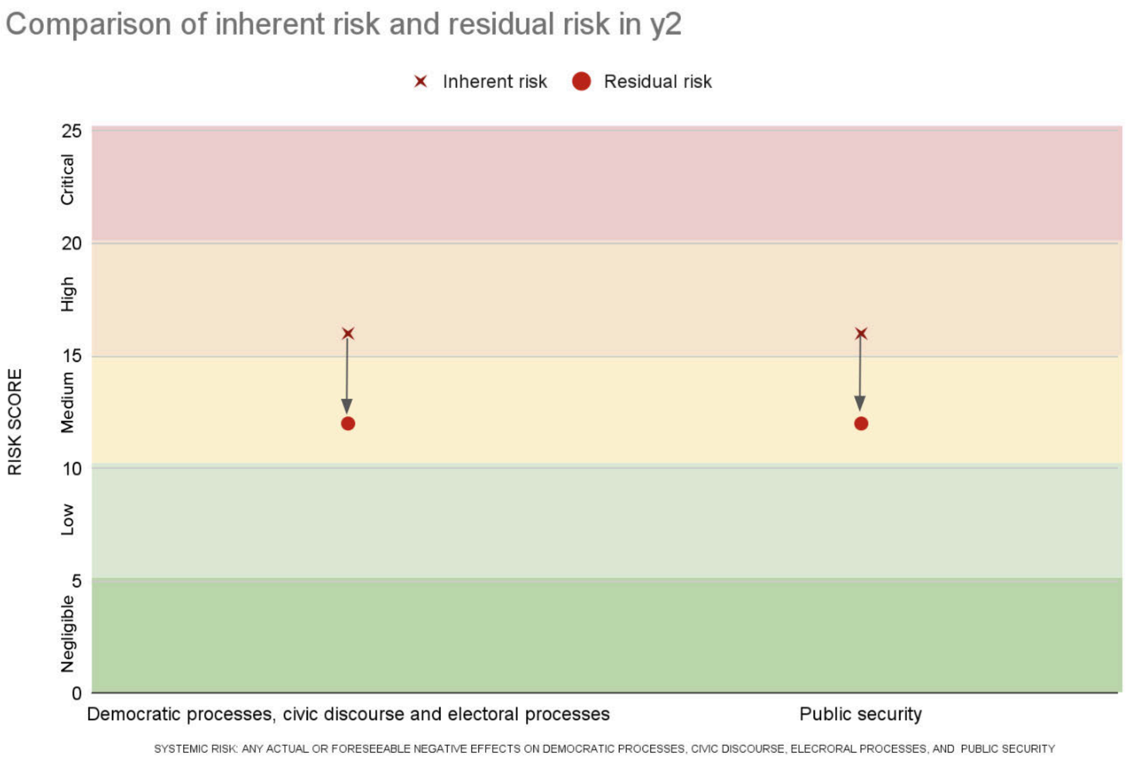


Fig. 10: Comparison of inherent and residual risk for democratic processes, civic discourse, electoral processes, and public security

Inherent risks:
This year has seen key elections in Europe - notably the EU elections along with other national elections such as the France Legislative elections. The ongoing Israel-Hamas conflict following the October 7th attacks has also raised the likelihood of threats to public security in Europe. As discussed in our Y1 report, such external events may result in bad actors misusing X to spread false or misleading information, as well as conduct coordinated attacks to target public security. The risk environment is heightened by the potential for echo chambers to form, where users may be exposed to information that aligns with their existing beliefs, which can reinforce biases and may stifle healthy debate.

Controls to mitigate the risk to democratic processes, civic discourse, electoral processes, and public security

Policies & enforcement (Article 35(1)(b))
As discussed in our Y1 report, we have robust policies with dedicated teams in place to prohibit harmful behaviours. To learn more about how our Synthetic and Manipulated Media and our Violent Speech policies mitigate this risk, please refer to the Y1 report. In August 2023, we launched our updated Civic Integrity policy, which addresses four categories of misleading behaviour and content: (i) misleading information about how to participate in an election or other civic process, (ii) suppression, (iii) intimidation, and (iv) false or misleading affiliation. Posts enforced under this policy will receive a label informing both authors and viewers that the post's visibility has been restricted. This enforcement makes the post less discoverable on X, such as excluding it from search results, trends, recommended notifications, For You and Following timelines, and downranks the post in replies. This policy is activated leading up to, during, and after an election for a certain period of time. Any attempt to undermine the integrity of civic participation undermines our core tenets of freedom of expression, and as a result, we use labels to inform users that the content is misleading.

As mentioned in the section dedicated to our risk environment and controls, we also launched a Violent Content policy in May 2024, which consolidates two major policies: Violent Speech and Violent Media. Through this policy, X allows users to share graphic media if it is properly labelled, not prominently displayed, and is not excessively gory or depicting sexual violence. Enforcement taken under this policy is proportionate to the harm. For example, violent threats, wish of harm, incitement of violence, glorification of violence, violent sexual conduct, gratuitous gore, beastiality and necrophilia is removed from the platform and further violations may result in the account being suspended or placed on read-only mode. Lower severity harms, such as any minor or non-deliberate instances of violent speech, depictions of physical fights, or bodily fluids, are labelled and consequently have their reach restricted, ensuring that users who do not wish to see it can avoid it and that minors are not exposed to it. Any attempt to undermine the integrity of civic participation through violent speech also undermines our core tenets of freedom of expression, and as a result, we action this content.

Product-level controls (Article 35(1)(a))
At a product level, the Community Notes function remains a leading mitigation for the risk of misinformation, relating to both public security and civic integrity. For more information, please refer to the Zoom in: Community Notes. Additionally, we have product interventions to direct people to key resources on how to register to vote and reminders to vote in order to encourage

civic participation. These take the form of election prompts on the home timeline and search timeline, which display official voting information, along with hashmojis on common election hashtags.

At the time of this report, X does not allow political ads in the EU. The effectiveness of this measure has been evidenced by a study conducted by the organisation Global Witness, who submitted test ads containing false information about polling stations, incorrect ways to vote and incitement to violence against immigrant voters to the platform. On X, all ads were halted, and account level action was taken due to repeat offences.[23]

Partnerships (Article 35(1)(f))
As part of a multi-risk environment, we recognise the importance of collaborating with partners and sharing information to take down bad actors and threats to civic integrity. In our Stakeholder engagement and consultation section we have discussed the range of engagements we have undertaken this year. Specifically to mitigate this systemic risk, we cooperated with the German Minister of Foreign Affairs, French Viginum and the French Ministry of Foreign Affairs, as well as with the Polish government's cybercrime centre, exchanging leads and information on investigations into coordinated networks on the service. Cooperation with Germany, France and Poland on this front are ongoing and framed under the "Weimar group" established between the three EU Member States to tackle foreign interference online. X teams were also in contact with the European External Action Service (EEAS) and European Digital Media Observatory (EDMO) during the EU Elections to exchange alerts and relevant information on potential threats to the platforms' integrity during the elections. We also visited Slovakia ahead of their election to meet relevant agencies. At a more global level, X is also in contact with NATO [redacted] to allow the agency to share information related to misleading information and foreign interference, [redacted] We also have escalation paths established between X and the Access Now Digital Helpline and Article19 to provide support as needed to civil society groups.

We have continued to collaborate with our existing partners as well as law enforcement authorities, notably in the context of threats to public security. For more information, please refer to our Y1 report.

Zoom-in: EU elections

As we build the most trusted global town square, we know that the public debate around elections happens on X. We are proud that our platform powers democratic discourse and life around the world, and for us, authenticity, accuracy, and safety are fundamental to how we approach elections. Our consideration of authenticity has two principal dimensions: accounts and conversation. Our Safety team is constantly monitoring the service for action under our

[23] https://www.globalwitness.org/en/campaigns/digital-threats/ticked-tiktok-approves-eu-elections-disinformation-ads-publication-ireland/

policies around Platform Manipulation and Spam. Our teams consistently thwart and disrupt threat campaigns designed to degrade the integrity of the platform.

We strongly believe that freedom of speech and safety must coexist, and the election context brings with it a diverse set of challenges that may include abuse and harassment, violent content, deceptive identities and impersonation, violent hateful entities, hateful conduct, synthetic and manipulated media, political advertising (where applicable), and misleading information about how to participate and vote.

Our EU elections response involved a cross company effort, with multiple teams providing additional monitoring to identify potential violations of X Rules on top of Safety's existing enforcement mechanisms and other mitigation measures, such as 24/7 escalations support. In the months before the elections, we participated in a series of events organised by the European Commission (DG CNECT), such as: stress tests on platforms' preparedness to prevent and tackle threats to elections integrity, and election roundtables to share information on identified potential harms, as well as on platforms' and EU institutions and member states' initiatives to protect civic integrity. We also presented our elections approach and an overview of X's election integrity efforts to Coimisiún na Meán and other Digital Services Coordinators. Ahead of, during, and after the EU elections, we activated a comprehensive set of measures and engagements to protect civic processes, which included:

- ***Proactive engagement:*** We proactively engaged and exchanged information with the European Commission, the European External Action Service (EEAS), the European Parliament, and key authorities of the 27 Member States. As a part of this engagement, we provided crisis response contact points to the European Commission, European Parliament, and DSCs and gave a safety training to more than 60 EU-based NGOs on how to maximise use of safety tools on the platform. X also proactively cooperated with the European Commission and Member States on identifying and disrupting networks of inauthentic profiles that were posing a threat to elections integrity. We are proud that our work on elections was praised by the European Commission, a number of Member States, the European Parliament's communication service and the EEAS, as communication moved smoothly during the election and escalations were dealt with promptly;
- ***Media literacy campaigns:*** To promote civic engagement, we supported media literacy campaigns with trusted partners and recognised experts in the EU, such as the European Parliament, European Digital Media Observatory (EDMO), and the European Regulators Group for European Media Services (ERGA) that aimed at providing reliable information on the EU elections. Specifically, we supported media literacy campaigns via ads credits and received positive feedback from ERGA on the reach obtained by their campaign thanks to the credits. ;
- ***Election enforcement period:*** Leading up to EU elections, we activated our Civic Integrity policy, conducted additional monitoring on top of Safety's existing enforcement

mechanisms to identify potential violations of X Rules, and provided 24/7 escalations support. As a result of this, over [redacted] total enforcement actions occurred during our monitoring and escalations support of the EU elections; and

- ***Product interventions:*** We provided election-related prompts on the Home and Search timelines, which surfaced official information from the European Parliament to users. These had over [redacted] and [redacted] impressions respectively, from X users in the EU. We also launched [redacted] election hashmojis on the platform when users used hashtags such as #UseYourVote, #EUelections2024, and #EU2024 in the EU official languages. These hashmojis resulted in more than [redacted] clicks and [redacted] searches by X users in the EU.

Thanks to this work, X did not see a major disinformation incident in the last critical days ahead of EU-wide voting. Officials, experts, and tech firms across the bloc had flagged in the last months different waves of disinformation from coordinated campaigns such as pro-Russian Doppelganger on X and Meta to homegrown falsehoods. However, no last minute major deepfakes were identified, confirmed [redacted], coordinator of fact-checking of teams around the bloc for EDMO. [redacted] a Commission spokesperson, also confirmed that platforms were well prepared and that no major incidents took place over the election weekend.

Negative effects to democratic processes, civic discourse, and electoral processes

Risks to democratic processes, civic discourse, and electoral processes may arise from false or misleading information, voter intimidation and/or suppression, presence of hateful entities, government requests and surveillance, foreign interference and manipulation (FIMI), as well as other inauthentic behaviour.

X has paid particular attention to civic integrity on its platform, over the last year, during EU elections and other national elections in the Union - by creating election risk assessments, implementing mitigation measures and collaborating with EU stakeholders to protect election integrity in the EU.

Probability
Between October 2023 to June 2024, X actioned [redacted] posts in the EU, following DSA illegal content reports under 'Negative Effects on Civic Discourse or Elections'. Around [redacted] of these items were geoblocked. Actions for this category account for only [redacted] of the total actions taken on DSA user reports. However, our enforcement on Platform Manipulation and Spam comprises more than [redacted] of the X Rules enforcement on the platform – an area which overlaps with risks to civic discourse. Further, the multiple national elections that took place in the EU in the past year, most notably the EU elections in June 2024, increased the likelihood of risks in this area. As such, the probability of this risk has been assessed to be *likely*.

Severity

- ***Scope**:* The amplification of false or misleading information on X, combined with harassment and intimidation of people, notably vulnerable groups, related to electoral processes can have a significant impact on civic participation. As a multi-dimensional harm that also impacts vulnerable groups, this was assessed to have a *high* scope;
- ***Scale**:* DSA illegal content user reports under 'Negative Effects on Civic Discourse or Elections' accounted for less than [redacted] of the total user reports received between October 2023 and June 2024. However, conversations regarding politics are among the top items discussed on X globally and receive significant engagement.[24] As such, this risk was assessed to have a *high* reach;
- ***Remediability**:* Risks related to false and misleading information can be remedied by providing users with additional context, such as a Community Note or Synthetic and Manipulated Media label. As such, this has been assessed to be *possibly remediable*;
- Based on the assessments above, the risks to democratic processes, civic discourse, and electoral processes are assessed to have a *high* severity.

Inherent risk

Based on the probability of risks to democratic processes, civic discourse and electoral processes on the platform, along with the high severity of such a risk, this area has a *high inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls for risks to democratic processes, civic discourse, electoral processes, and public security described above, specific controls targeting this risk include:

- **Article 35(1)(b) - Policies & enforcement**: Our Civic Integrity, Synthetic and Manipulated Media, and Platform Manipulation and Spam policies primarily cover this area and are well-defined. X has effective means of removing bad actors, including actors attempting to inauthentically manipulate user conversations, at scale, through enforcement of the Platform Manipulation and Spam policy.
- **Article 35(1)(f) - Elections playbooks and 'retros':** Election-specific processes to prepare for and during elections are in place and well documented, such as our election playbooks. Following an election, the cross-functional election working group builds a retrospective analysis of the enforcement taken during the relevant time frame. This 'retro' acts as a feedback loop to inform the working group in future efforts.

Over the last year, further controls have been implemented and existing controls have been improved upon, that align with Article 35, to target this risk:

- **Article 35(1)(b) - Policies:** Our Civic Integrity policy was launched mid September 2023, to address voter intimidation and suppression during elections. In preparing for each

[24] https://x.com/XData/status/1764757748707672167

election and the enforcement of the Civic Integrity policy, teams prepare guidelines to ensure reviewers have relevant information and regional and linguistic context of the country in question.

- **Article 35(1)(f) - Election risk assessments:** For each national election, X conducts an assessment to evaluate the election's potential risk to civic discourse and electoral processes on X, which allows us to determine what services or additional mitigations to activate on top of our already existing and comprehensive policies and enforcement processes.
- **Article 35(1)(a) - Community notes**: This feature is now live in 72 countries, including all EU member states, and over 30% of ratings come from EU contributors, indicating interest and engagement from users in the EU. For further data on this feature, please refer back to Zoom in: Community Notes.
- **Article 35(1)(f) - Partnerships**: Over the past year, X has cooperated with both the French VIGINUM Taskforce, [redacted], and, more recently, with the Weimar Triangle (consisting of the French, German and Polish Ministries of Foreign Affairs), exchanging leads and information on investigations into coordinated inauthentic networks on the service. Additionally, and especially during the EU elections, X teams have also actively cooperated with the European External Action Service (EEAS), the European Parliament and the European Commission's Communication teams, as well as other key stakeholders like the European Digital Media Observatory (EDMO) and EUDisinfolab, and key authorities from the 27 Member States.
- **Article 35(1)(f) - Election integrity:** We have a cross-functional working group focused on elections integrity, and increasing resources allocated to ensuring elections integrity is an ongoing process.
- **Article 35(1)(i) - Election product interventions:** In both the EU elections and the France Legislative elections, we launched Home and Search timeline prompts, which surfaced official information from the European Parliament and the French Ministry of the Interior, respectively, to users. This has received positive feedback from the French government, which attributes 45% of the total traffic on their interministerial webpage on the France Legislative elections to X. This activity was cited as a direct consequence of trend takeovers and election day and reminder prompts. Additionally, we launched multiple hashmojis for election-related hashtags for both the European and France Legislative elections.

This control is assessed to be *defined*. Over the past year, we have made efforts to expand and develop measures and policies specific to elections, as outlined above. Robust quality assurance frameworks will be implemented and processes will continue to be improved. Generally, processes tend to be more proactive than reactive, and they are well characterised and understood across all organisation verticals.

Tier 2 priority

Due to the *high inherent risk* of this area, which is mitigated by controls of a *defined* nature, the residual risk for negative effects to democratic processes, civic discourse and electoral processes is a *medium risk item*, making it a *Tier 2 priority*. As our controls have improved, the residual risk in this area is better managed, resulting in this tiering. Nevertheless, we will continue to evaluate these risks and our controls as they may continue to evolve. Our efforts to continue to address residual risk are detailed in VII. Considerations for further mitigations.

Risks to public security

Broadly defined, the risk to public security includes threats that have the potential to undermine social order, disrupt civil harmony, and compromise the safety of individuals and communities. Such risks may manifest on the platform through spread of harmful misinformation, coordination of unlawful activities by bad actors, sale of illegal goods such as psychoactive substances and firearms and perpetrators of violent attacks using the platform to recruit, organise and disseminate violent content.

Over the last year, there has been no particular incident that has changed the risk profile of this harm. There are overlaps between this risk area and the risk of terrorist content on the platform, as well as risks to consumer protection – risks following the October 7th attacks have thus been considered in the terrorist content Risk Assessment.

Probability

Between October 2023 to June 2024, X took a total of [redacted] actions for X Rules violations, including [redacted] content removals and 38K suspensions under its Violent and Hateful Entities, Violent Speech, Perpetrators of Violent Attacks, and Illegal or Regulated Goods or Services policies. Further, almost [redacted] items were actioned following DSA illegal content reports under 'Risk for Public Security'. Although this is only [redacted] of the total actions taken following DSA user reports, when understood alongside the volume X Rules enforcement, the probability of this risk can be assessed as *likely*.

Severity

- ***Scope***: Harms from this area may contribute to a polarised society, erode trust in information sources, influence public opinion and potentially incite real world violence. Such harms could also affect users' psychological well-being. That said, it is difficult to establish a causation between harmful on-platform content resulting in a societal risk to public security and, in most cases, the harmful content appears on the platform once the public security incident has already happened. Nevertheless, the scope of harm here is assessed to be *high*;
- ***Scale***: Almost [redacted] of user reports for X Rules between October 2023 to June 2024 were made under the Violent Speech and Violent and Hateful Entities policies. While there is no certain correlation between, for example, Violent Speech enforcement and public security, it does provide a proxy for the potential reach of such harmful content. As such, this risk category was assessed to have a *high* scale, given that relatively more users seem to have seen and reported these items;
- ***Remediaibility***: The high potential of physical and psychological harm arising from this risk makes the reversibility of this harm unlikely. It has hence been assessed to be

rarely remediable;

- Based on the above, the risk of public security on the platform is assessed to have a *high* severity.

Inherent risk

Based on the probability of risks to public security on the platform, along with the high severity of such a risk, this area has a *high inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls to risks to democratic processes, civic discourse, electoral processes, and public security described above, specific controls targeting this risk include:

- **Article 35(1)(b) - Policies & enforcement**: Public security risks are enforced upon under the Violent Content, Illegal or Certain Regulated Goods or Services, Violent and Hateful Entities, Perpetrators of Violent Attacks, and to a lesser extent Abuse and Harassment and Impersonation policies. The DSA reporting form also has a category dedicated to 'risk for public security';
- **Article 35(1)(c): Consistent moderation:** The above policies are accompanied by cohesive, consistent processes that enable agents to make risk-informed decisions, allocate resources and apply timely and appropriate remediation measures. For the Violent Content, Violent and Hateful Entities, and Abuse and Harassment policies, X employs both automated and manual enforcement mechanisms.

Over the last year, further controls have been implemented and existing controls improved upon, in alignment with Article 35, that target this risk:

- **Article 35(1)(b) - Policies & enforcement**: We have conducted a comprehensive policy review, which has led to improvements in X policies, particularly around Violent Media;
- **Article 35(1)(f) - Violent entities:** We have made changes to our global list of designated violent entities and expanded it, as part of our continuous work to carry our comprehensive assessments. We have also increased proactive monitoring and enforcement for violent entities;
- **Article 35(1)(c) - Incident response and post-incident reviews:** We have continued to enhance feedback mechanisms with post-incident reviews and regular syncs to ensure that enforcement aligns with the spirit and purpose of the policies. We continue to have internal incident response protocols in place when a high-visibility event occurs and virality triggers rapid and widespread proliferation of various content types on the platform. Even if the incident does not reach the 'crisis' level, our escalations team may direct resources toward an immediate response.

The current mechanisms in place are defined, scalable, and operating effectively. X has well-developed policies to moderate content that promotes or celebrates violence or endangers public security across corresponding teams (enforcement/operations, training, engineering, data analytics, and external engagement) and ensures policy development, enforcement and maintenance is up to date. As a result, the control strength is assessed as *defined.*

Tier 2 priority

Due to the *high inherent risk* of this area, which is mitigated by controls of a *defined* nature, the residual risk to public security is a *medium risk item*, making it a *Tier 2 priority*. As our controls have continued to evolve, the residual risk remains managed; nevertheless, we will continue to evaluate these risks and our controls as they may continue to evolve. Our efforts to continue to address residual risk are detailed in VII. Considerations for further mitigations.

D. Public health, physical and mental well-being, and gender-based violence

This systemic risk area considers the risk of negative effects to public health, including harms to physical and mental well-being and gender-based violence (GBV). As discussed in our Y1 report, the discourse around the usage of social media and its impact on health remain varied. While all online platforms may be misused as a vector for risks, there are notable positive influences on public health, mental and physical well-being as well as the rights of vulnerable populations. In comparison to Year 1, the inherent risks and residual risks for this systemic risk area stayed the same, indicating that this continues to be a managed risk area. The following graph shows the inherent and residual risks for this area in Y2.

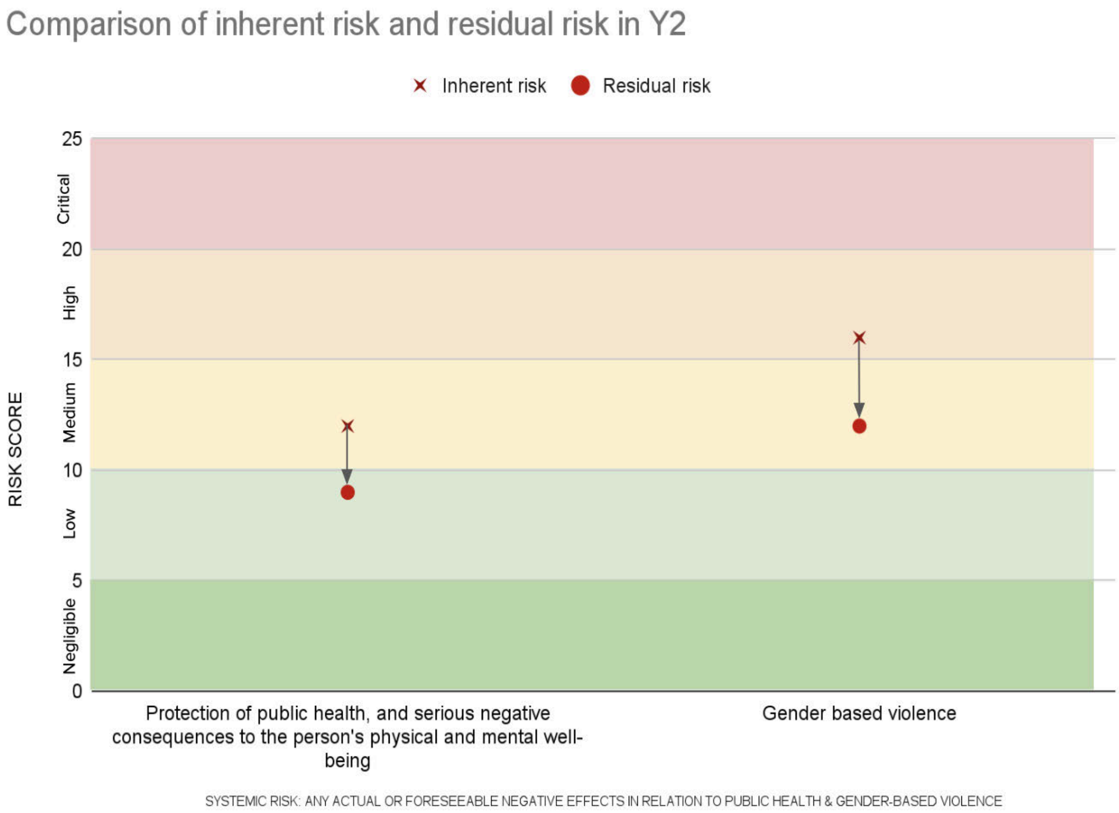


Fig. 13: Comparison of inherent and residual risk for public health and gender-based violence

Inherent risks

Our analysis suggests that, globally, X users spend an average of 30 minutes a day on the platform.[25] The full extent of the effects of negative interactions and exposure to graphic content may harm users' psychological well-being is yet to be determined. Similarly, misuse of the platform to promote dangerous activities or misleading information may be detrimental to public health. Although there has been no public health crisis declared in the EU or globally in the past year, this risk area may still be present at a societal level through users amplifying misleading information related to public health, and at an individual level through users who may share sensitive or harmful media such as self-harm content and discussions promoting eating disorders.

GBV may result in risks to physical safety, especially when it involves non-consensual intimate image sharing or outing of a victim's identity. Such abuse may further result in impacted communities self-censoring their voice. The use of AI tools can exacerbate the risks of dissemination of GBV content, for example, as seen in the sharing of non-consensual nudity (NCN) imagery related to Taylor Swift. Although X allows consensual adult content on the platform, there is a risk of illegal pornographic content being disseminated, this may include CSAM, NCN and intimate imagery either shared or produced without consent of the person depicted in the image.

Controls to mitigate the risk to public health, physical and mental well-being, and gender-based violence

Policies & enforcement (Article 35(1)(b))

In order to mitigate the identified inherent risks, we have developed a comprehensive and targeted set of policies that capture all our services and features. X's Rules and revenue policies govern what can be shared and advertised or promoted on the platform, prohibiting illegal content and limiting content that could potentially be harmful.

X has multiple policies that capture this risk area. For risks to public health, this includes Abuse and Harassment, Platform Manipulation and Spam, Suicide and Self-harm, Child Safety, and Illegal or Certain Regulated Goods or Services, as well as Self-Harm and Unsafe and Illegal Products under DSA reporting categories. For risks of gender based violence, this includes Abuse and Harassment, Sensitive Media, and Non-Consensual Nudity as well as Non-Consensual User Behaviour and Pornography or Sexualised Content under DSA reporting categories. These policies are enforced using a wide range of measures, including content labelling, restrictions, removals, and account suspensions.

Over the past year, as part of our ongoing commitment to refine our policies and enforcement, we have conducted a comprehensive audit of our existing guidelines and workflows. As mentioned in IV. X Risk Environment: Influencing Factors & Controls, this audit led to improvements in X policies, particularly around consensual Adult Content and Violent Media. As before, X takes a nuanced approach to sexual content whereby we allow space for consensual sharing and self-expression, but at the same time, draw a clear line when it comes to non-consensually shared nudity or sexual content. Users are allowed to post Adult Content - which includes adult nudity and sexual behaviour - provided that it is properly labelled with a

[25] https://x.com/XData/status/1769826435576037702

content warning so that users who do not wish to see it can avoid it. However, this content is not allowed on highly visible areas including live videos, profile pictures, header, banners, or Community cover photos. As minors' accounts are defaulted to protected, they are not exposed to such labelled content either.

Product-level controls (Article 35(1)(a))
X has a suite of product-level features to mitigate against potential harms related to public health, physical and mental well-being and GBV that may manifest on the platform, which includes Community Notes and content warning labels. Content warning labels can be proactively added by users or reactively added by our content moderators. User safety features such as block/mute, account filters, and protecting posts/controlling replies, also limit exposure to harmful content.

If a user searches for terms related to self-harm or suicide in certain countries, X guides the user towards resources with expertise in crisis intervention and suicide prevention that the user can contact. Users can also alert the X team focused on handling reports associated with accounts that may be engaging in self-harm or suicidal behaviour. For further information on our controls and enforcement in this area, please refer to our Year 1 report.

Zoom-in: GenAI & Gender-Based Violence – Taylor Swift Deepfake

At the beginning of 2024, X became aware of AI generated Non-Consensual Nudity (NCN) being spread of the singer Taylor Swift. Immediately on being alerted to this trend, X initiated its incident response protocol, allowing it to take prompt and comprehensive steps to stop the spread of these images.

Working around the clock, teams from across the company carried out proactive sweeps to remove violative content and to suspend the accounts of bad actors and repeat offenders. Our sweeps were escalated as the incident progressed and the volume of violative content increased. Ad-hoc guidance was issued and further training provided to our enforcement teams at short notice to respond to the incident. A statement was published on Safety, sending a clear signal regarding our zero-tolerance approach to Non-Consensual Nudity. As a temporary safety measure, searches for "Taylor Swift" were blocked on the platform.

The enforcement numbers from the incident as of Feb 21, 2024, when the sweeps were ceased, are provided below.

Account Suspension	[redacted]
Post removal	[redacted]
Post removal (one-off)	[redacted]
Content warning label	[redacted]

The actions we took are a testament to the flexibility and robustness of our incident response mechanisms, and are in line with our zero-tolerance approach to non-consensual nudity. At the same time, the event proved to be a valuable opportunity for X to improve our products and policies. Efforts include:

- Following a post-incident review, we conducted a policy-mapping exercise and clarified with our operational teams how to enforce our rules on AI-generated deep fakes;
- A tooling exercise was conducted to improve our automated systems and their recognition of various hashes related to Non-Consensual Nudity.

Risks to public health and physical and mental well-being

Unprecedented use of social media can negatively impact users' mental health and, in severe cases, their physical health. On a societal level, risks include the dissemination of harmful or false health information, particularly during public health emergencies, and content that undermines trust in health institutions and professionals. There is also a risk to fundamental rights of free expression when discussing public health topics as we've seen in the past with examples such as the Covid-19 pandemic that there can be significant public discussion on public health measures that evolve over time. On an individual level, users may encounter harmful content such as bullying, harassment and self-harm, or develop issues like addiction and reduced attention span due to the platform's design and functionality. A recent study by Internet Matters has shown that children aged 9-15 and their parents found that active users were more likely to encounter harm online. At the same time, this age group experienced more positives across all the dimensions of wellbeing - developmental, emotional, physical, and social - compared with their less active counterparts.[26] Over the last year, there has been no particular incident that has changed the risk profile of this harm.

Probability

Between October 2023 to June 2024, X has actioned [redacted] posts and accounts for violations of Abuse and Harassment, Suicide and Self-Harm, and Sensitive Media[27]. However, there is no clear correlation between some of the sub-harms that can trigger the enforcement of the listed violations and an impact to public health. For example, enforcement for Abuse and Harassment could be a result of a slur being targeted at a user, however, there is no direct indication that this may have impacted the user's mental health. As such, while we recognise the risks to public health stemming from our platform, the full effects remain unknown, as they are related individual determinants of wellbeing. The probability for this risk is *possible*.

[26] https://www.internetmatters.org/wp-content/uploads/2023/02/Internet-Matters-Childrens-Wellbeing-in-a-Digital-World-Index-report-2023-2.pdf

[27] It should be noted that Sensitive Media included both adult content and violent content. Following the policy update, this policy has now been separated. However, for the purpose of this risk assessment, we are unable to provide data that is specific to adult content. This will be updated in next year's assessment.

Severity

- ***Scope***: Users amplifying false and misleading information about public health related items, or promoting the sale of counterfeit documentation, may result in societal harm and has the potential to cause physical harm. Furthermore, risks to physical and mental health inherently constitute physical and/or psychological harm, and may target vulnerable groups. As such, the scope is assessed to be *very high*;
- ***Scale***: [redacted] of user reports received by X were for X Rules violations that overlapped with this risk area. This indicates that the reach of this type of content on the platform is wide, putting the scale at *high*;
- ***Remediability***: Although mitigation measures could potentially help limit the extent of the harm, the remediability for negative health outcomes that have already occurred is limited, especially when it comes to the impact of public health crises. As such, remediability for this harm is *possibly remediable*;
- Based on the assessments above, the risk to public health on the platform is assessed to have a *high* severity.

Inherent risk

Based on the probability of risks to public health on the platform, along with the high severity of such a risk, the inherent risk of this area is a *medium inherent risk*. That is when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls to risks to public health and negative effects to physical and mental well-being, described above, specific controls targeting this risk include:

- **Article 35(1)(b) - Policies & enforcement:** X has a suite of policies to enforce against risks to public health, as well as negative effects to physical and mental well-being, such as Abuse and Harassment, Sensitive Media, and Suicide and Self-harm policy. The latter prohibits users from promoting or encouraging suicide or self-harm content.
- **Article 35(1)(i) - Mental health prompts:** X has product features in place with suicide and self-harm resources, such as mental health prompts in certain countries that appear when users search for words related to suicide and self-harm.
- **Article 35(1)(c) - Restricted reach and rate limiting:** These features work to reduce the impact of misleading activity on the platform by reducing impressions and limiting the number of actions an account can take;
- **Article 35(1)(a) - Safety features:** X has content warning labels on graphic and adult media and sensitive content settings;
- **Article 35(1)(f) - Crisis response:** X's crisis response protocol is based on a tiered approach that assesses risk of harm, business risks, and urgency. This informs the crisis activation procedure, and assigned ratings allow X to deploy an appropriate response based on the level of risk and prioritisation of each crisis;
- **Article 35(1)(c) - Reporting workflows:** Reporting mechanisms are in place for users to submit reports on rules violations, particularly Suicide and Self-harm, with ability to

appeal if they feel the wrong action was taken;

- **Article 35(1)(i) - Resources:** If a user is thinking about engaging in self-harm or suicidal behaviour, we have resources available that allow people to contact services with expertise in crisis intervention and suicide prevention. Users can also alert the X team focused on handling reports associated with accounts that may be engaging in self-harm or suicidal behaviour if they encounter this type of content on X.

Over the last year, further controls have been implemented and existing controls improved upon, that align with Article 35, to target this risk:

- **Article 35(1)(i) - Partnerships**: X provided ads credits for a public health campaign to the Red Cross in partnership with the French government to encourage people to practise sport 30 minutes a day to stay in good health.
- **Article 35(1)(a) - Community notes:** This feature has proven helpful to people from different points of view, and significantly reduces sharing of potentially misleading posts. For more information on improvements to this feature, please refer Zoom in: Community Notes.

The current mitigation measures are defined, well-documented and repeatable. Additionally, most of our mechanisms are proactive, which allows us to limit the misinformation within the platform. There is an established process for integrating feedback to mitigate process deficiencies. As such, the control strength is *defined*.

Tier 3 priority

Due to the *medium inherent risk* of this area, which is mitigated by controls of a *defined* nature, the residual risk to public health as well as negative consequences to a person's physical and mental-well being is a *low risk item*, making it a *Tier 3 priority*. As our controls evolve and public health conditions change globally, we continuously assess these risks and refine our measures. Notably, there may be product solutions that can support individuals' mental health, such as more curated support for victims of self-harm and cyberbullying. Such considerations are detailed in VII. Considerations for further mitigations.

Risks of gender-based violence

Due to similarities in the harms and controls, this year our assessment for GBV also considered the risk to the fundamental right of 'respect for private and family life'. As such, this risk area includes cyberviolence - such as sexual harassment, violent speech, gendered hate speech - sexual exploitation, non-consensual nudity (NCN), intimate imagery, disclosure of private information, sharing images of one's likeness without their permission and threats to expose individuals' private information or media.

Over the last year, there have been a few incidents that have shown how the use of AI tools may be used to exacerbate the risks of dissemination of content that may constitute GBV. The most known case is the Taylor Swift NCN incident which primarily affected the US (discussed above).

Probability

Between October 2023 to June 2024, following DSA illegal content reports, X actioned [redacted] items under the categories of Non-Consensual Behaviour and Pornography or Sexualised Content, although this is only a small fraction of the total enforcement on illegal content reports. However, X took a total of [redacted] actions for violations of Non-Consensual Nudity, Abuse and Harassment, and Sensitive Media – i.e, approximately [redacted] of all its X Rules enforcement actions at this time (excluding Platform Manipulation and Spam enforcement). Although not all of this may have been gender-specific (for example, Abuse and Harassment violations may go beyond gendered harassment) an overlap nevertheless exists. As such, the probability of this risk on X is assessed to be *likely*.

Severity

- ***Scope***: The scope of the sub-risks within gender-based violence span across physical, psychological, economic, societal, and informational harms, and they impact vulnerable groups. For example, dissemination of non-consensual nudity may pose significant risks to physical safety in countries where women and marginalised groups are disproportionately vulnerable to violence and reputational harm. Exposure of private content may impact an individual's financial security, be a reason for sextortion/blackmail, and result in the loss of further economic opportunities. As such, the scope of harm of this risk is *very high*;
- ***Scale***: The sub-risks within gender-based violence have a range of reach, depending on the nature of the risk. For example, between October 2023 to June 2024, X received approximately [redacted] user reports for the Sensitive Media policy, which accounts for only [redacted] of total user reports.[28] X received only [redacted] user reports for NCN in the same time period. Under DSA illegal content reports, X received [redacted] user reports across the categories of non-consensual behaviour and pornography or sexualised content; however, this also accounts for only [redacted] of the total DSA user reports received between October 2023-June 2024. As such, the reach of this harm ranges from *low* to *moderate*;
- ***Remediability***: When considering GBV, remediation is unlikely to restore the individual to their state prior to the impact. As a result, the sub-risks within this range from *possibly remediable* (e.g. respect for private and family life) to *not remediable* (e.g. gender-based violence and NCN);
- Based on the assessments above, the risk of gender based violence on the platform is assessed to have a *high* severity.

[28] It should be noted that Sensitive Media included both adult content and violent content. Following the policy update, this policy has now been separated. However, for the purpose of this risk assessment, we are unable to provide data that is specific to adult content. This will be updated in next year's assessment.

Inherent risk

Based on the probability of risks of GBV on the platform, along with the high severity of such a risk, this area has a *high inherent risk*, when assessed as a hypothetical scenario without considering the existing controls that reduce the risk.

Control strength

In addition to the global controls for risks to public health, negative effects to physical and mental well-being and gender-based violence, described above, specific controls targeting this risk include:

- **Article 35(1)(b) - Policies & enforcement**: X enforces on GBV via Abuse and Harassment, Hateful Conduct, NCN, Illegal or Certain Regulated Goods or Services (including sexual services) and media policies relating to Violent Content and Adult Content. We provide clear guidelines to our enforcement teams and we regularly update our policies and guidelines to reflect changes in trends;
- **Article 35(1)(c) - Training**: In order to sensitise our enforcement teams, we have also created cultural abuse training to help teams better understand how vulnerable groups tend to be targeted. We have regular meetings with agents to go through edge-cases. We also provide detailed guidance to agents when they're reviewing cases in different languages;
- **Article 35(1)(c) - Moderation**: Both proactive and reactive enforcement is used for this risk area with tight feedback loops; and
- **Article 35(1)(a) - Safety features**: Features such as block/mute, account filters, and controlling replies allow users to protect themselves from potential GBV;

Over the last year, further controls have been implemented and existing controls improved upon, that align with Article 35, to target this risk:

- **Article 35(1)(b) - Policies & enforcement:** We have conducted a comprehensive policy review, which has led to improvements in X policies, particularly around consensual Adult Content. We have also updated our Abuse and Harassment guidelines to account for unwanted sexualisation and objectification using AI-generated content.
- **Article 35(1)(c) - Incident response:** Following the Taylor Swift NCN incident, a post-incident report was created with a number of suggested improvements for the future. For further detail, please refer to Zoom-in: GenAI & Gender-Based Violence – Taylor Swift Deepfake.
- **Article 35(1)(f) - Partnerships**: X has recently partnered with StopNCII to work towards mitigating the risks of NCN. For more information on this, refer to VII. Considerations for further mitigations.

The current mechanisms in place are defined, repeatable and operating effectively. Processes are well characterised and understood. While many of the controls in this area may be considered to be 'managed', there is no proactive enforcement for NCN. As such, the overall control strength is *defined*.

Tier 2 priority
Due to the *high inherent risk* of this area, which is mitigated by controls of a *defined* nature, the residual risk of gender-based violence is a *medium risk item*, making it a *Tier 2 priority*. As our controls have continued to evolve, the residual risk remains managed; nevertheless, we will continue to evaluate these risks and our controls as they may continue to evolve. Our efforts to continue to address residual risk are detailed in VII. Considerations for further mitigations.

VIII. Considerations for further mitigations

Despite an increase in political and societal risks in 2024, over the last year, the residual risks have reduced in several areas in comparison to Y1. Notably, the residual risk has improved across five areas – illegal hate speech, CSAM, freedom of expression, other fundamental rights and democratic processes, electoral processes, and civic discourse. For risks to consumer protection, due to the expansion of this assessment to also consider the risk of sale of illegal goods and services, the residual risk has marginally increased from Y1, while still remaining a low risk.

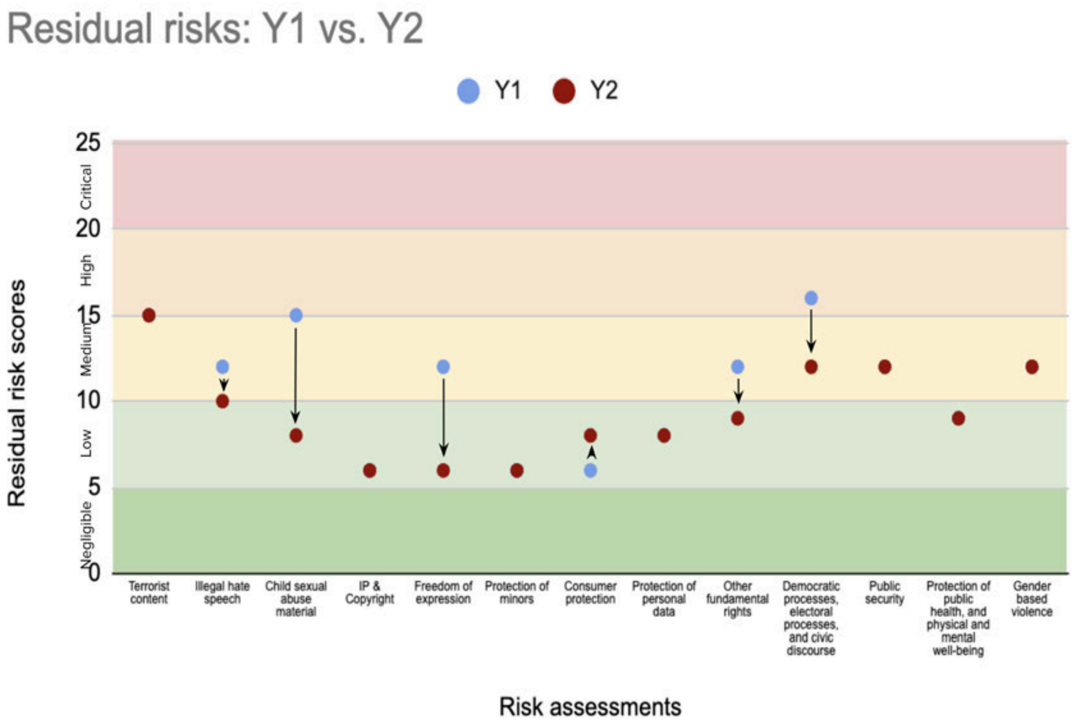


Fig. 12: Comparison of residual risk between Y1 and Y2

This improvement in residual risk comes both as a result of a more refined evaluation of the risks on the platform, based on the more data-driven approach, as well as improvements in our controls over the last year. Notably, for illegal content, several measures put in place to comply with the DSA have increased our suite of controls tackling illegal content in the EU. Similarly, improvements to our restricted reach labelling, the launch of our Civic Integrity policy as well as collaboration with external stakeholders, such as EDMO and other government bodies, has improved our controls and overall reduced the assessed risks to fundamental rights and democratic processes.

The following prioritisation derives from the residual risk calculation, and informs the VII. Considerations for further mitigations in Y2:

[redacted]

Ultimately, we recognise that these systemic risks continue to evolve and as such we remain committed to our vigilance in managing these risks. It is important to note that we diligently continue to monitor and mitigate the risk areas considered as Tier 3 priorities so that they remain at a low residual risk, however, this tiering allows us to prioritise our efforts over the next months to tackle the highest risk areas on our service first.

In line with Article 35, the following table outlines further reasonable, proportionate and effective mitigation measures X plans to explore in Y2, with particular consideration given to the impacts of such measures on fundamental rights. These measures are additional improvements and avenues to consider, stemming as a result of this risk assessment, and will be considered in conjunction with our current suite of controls.

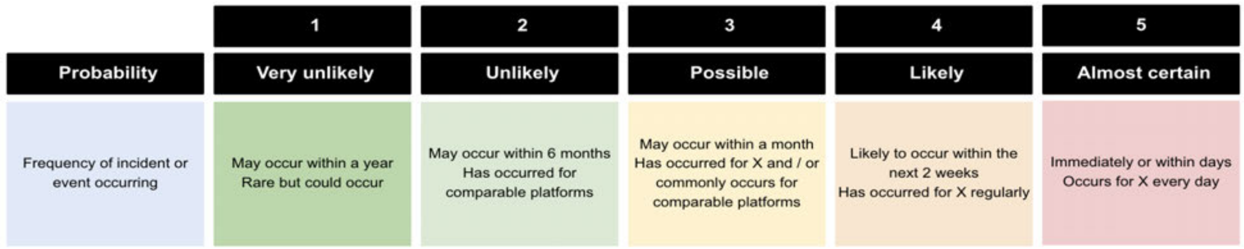
Systemic risk	Considerations for further mitigations
Measures that target systemic risks horizontally	● **Article 35(1)(a):** X will continue to improve on Community Notes. As of July 2024, users can request a Community Note on a post they believe would benefit from one. We also aim to continue making improvements to application speed; ● **Article 35(1)(c):** X will continue efforts to ensure that reporting options are better targeted and more effective across all policy areas; ● **Article 35(1)(b):** X will continue to conduct policy reviews for potential improvements and simplification;

	● **Article 35(1)(c):** X will continue to iterate and improve upon automated moderation techniques for improved detection of violative content before it is reported. [redacted]
Risk of dissemination of illegal content	[redacted]
Risks of negative effects to fundamental rights	[redacted]

	[redacted]
Risks of negative effects to democratic processes, civic discourse, electoral processes, and public security	● **Article 35(1)(b) & (f):** X will continue to evaluate and make any needed improvements to the suite of policy areas overlapping with elections and civic integrity and will remain ready, if necessary, to adapt our internal procedures and resources for ensuring election preparedness and responsiveness; ● **Article 35(1)(f):** X will continue to improve the efficiency and organisation of our election planning and enforcement resources, specifically incorporating input from the cross-functional election working group.
Risks of negative effects to public health, including physical and mental well-being, and gender-based violence	[redacted] [redacted] ● **Article 35(1)(b):** X aims to continue reviewing all policies related to adult sexual exploitation and address any gaps or concerns we may identify.

IX. Annex: Matrices

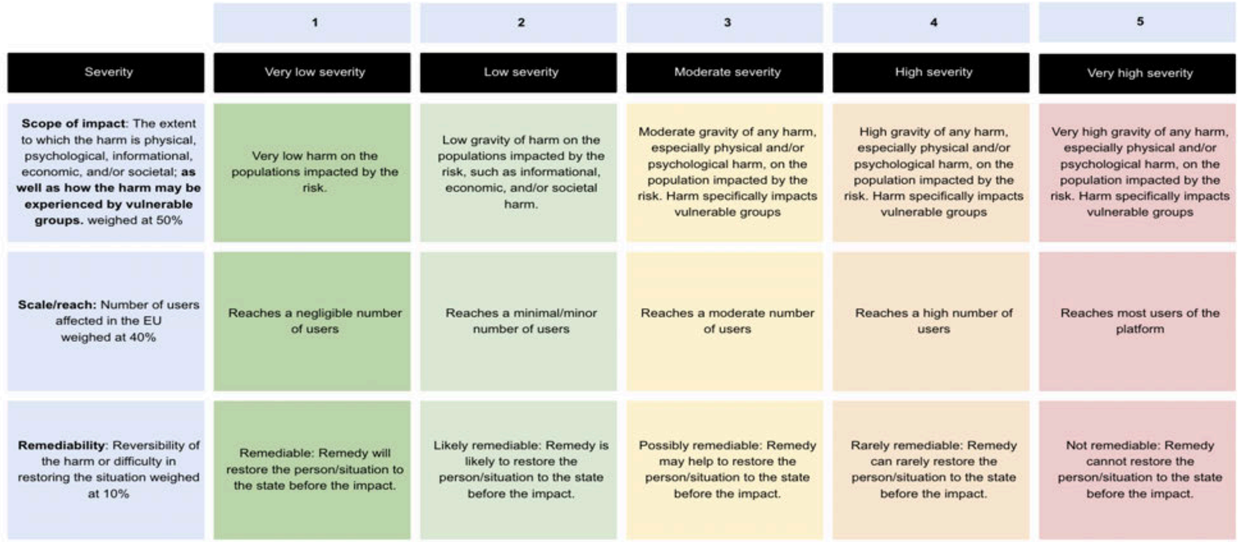
1. Probability matrix



	1	2	3	4	5
Probability	**Very unlikely**	**Unlikely**	**Possible**	**Likely**	**Almost certain**
Frequency of incident or event occurring	May occur within a year Rare but could occur	May occur within 6 months Has occurred for comparable platforms	May occur within a month Has occurred for X and / or commonly occurs for comparable platforms	Likely to occur within the next 2 weeks Has occurred for X regularly	Immediately or within days Occurs for X every day

Fig.13: Probability scale for the purpose of the DSA risk assessment

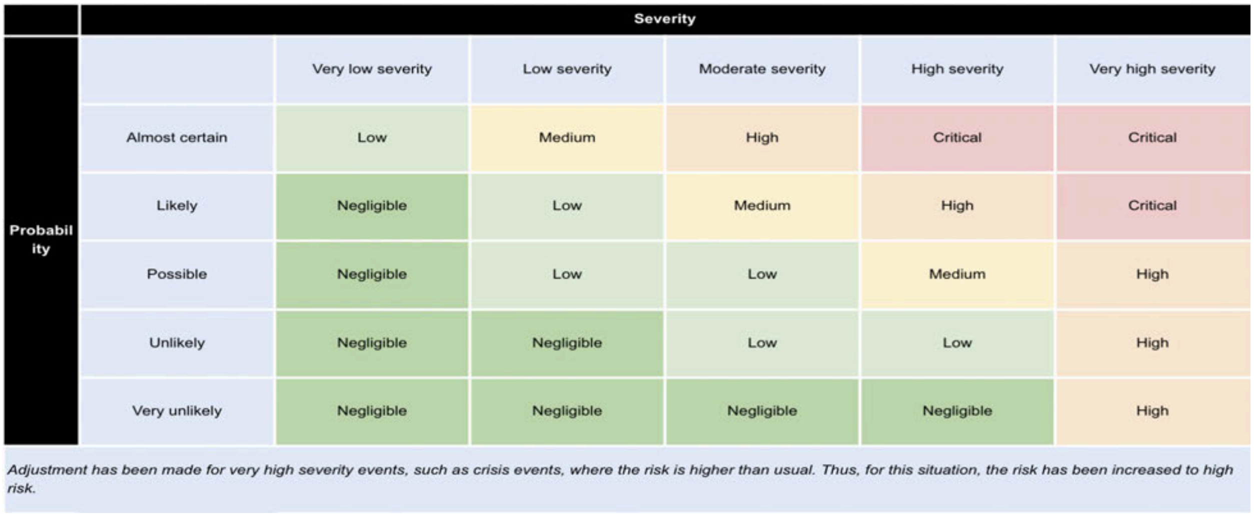
2. Severity matrix



	1	2	3	4	5
Severity	Very low severity	Low severity	Moderate severity	High severity	Very high severity
Scope of impact: The extent to which the harm is physical, psychological, informational, economic, and/or societal; **as well as how the harm may be experienced by vulnerable groups.** weighed at 50%	Very low harm on the populations impacted by the risk.	Low gravity of harm on the populations impacted by the risk, such as informational, economic, and/or societal harm.	Moderate gravity of any harm, especially physical and/or psychological harm, on the population impacted by the risk. Harm specifically impacts vulnerable groups	High gravity of any harm, especially physical and/or psychological harm, on the population impacted by the risk. Harm specifically impacts vulnerable groups	Very high gravity of any harm, especially physical and/or psychological harm, on the population impacted by the risk. Harm specifically impacts vulnerable groups
Scale/reach: Number of users affected in the EU weighed at 40%	Reaches a negligible number of users	Reaches a minimal/minor number of users	Reaches a moderate number of users	Reaches a high number of users	Reaches most users of the platform
Remediability: Reversibility of the harm or difficulty in restoring the situation weighed at 10%	Remediable: Remedy will restore the person/situation to the state before the impact.	Likely remediable: Remedy is likely to restore the person/situation to the state before the impact.	Possibly remediable: Remedy may help to restore the person/situation to the state before the impact.	Rarely remediable: Remedy can rarely restore the person/situation to the state before the impact.	Not remediable: Remedy cannot restore the person/situation to the state before the impact.

Fig.14: Severity scale for the purpose of the DSA risk assessment

3. Inherent risk matrix

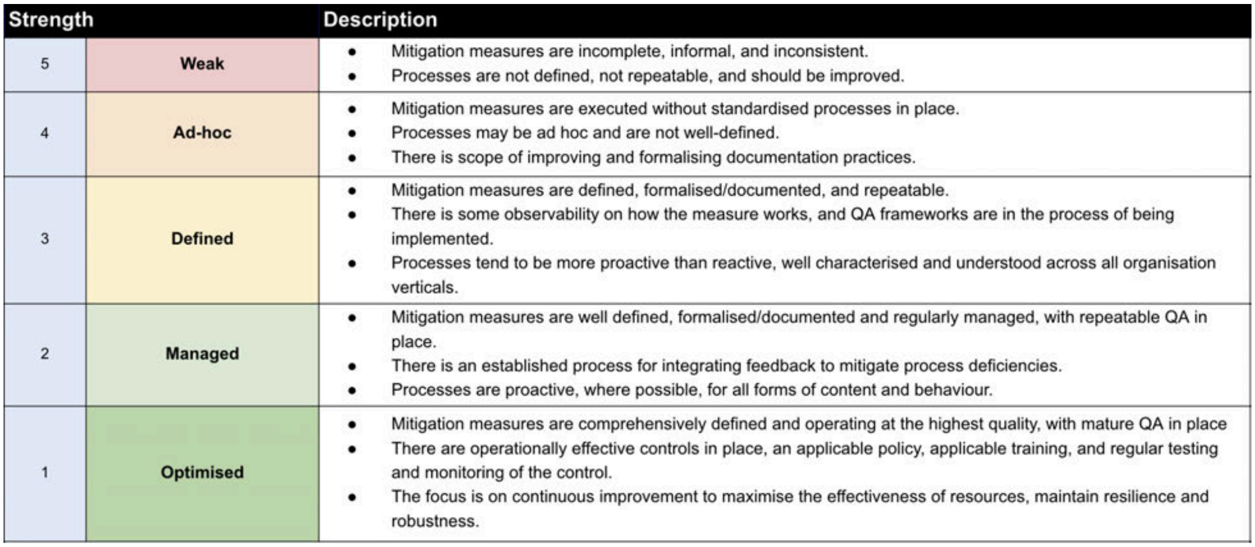


		Severity				
		Very low severity	Low severity	Moderate severity	High severity	Very high severity
Probability	Almost certain	Low	Medium	High	Critical	Critical
	Likely	Negligible	Low	Medium	High	Critical
	Possible	Negligible	Low	Low	Medium	High
	Unlikely	Negligible	Negligible	Low	Low	High
	Very unlikely	Negligible	Negligible	Negligible	Negligible	High

Adjustment has been made for very high severity events, such as crisis events, where the risk is higher than usual. Thus, for this situation, the risk has been increased to high risk.

Fig.15: Residual risk matrix for the purpose of the DSA risk assessment

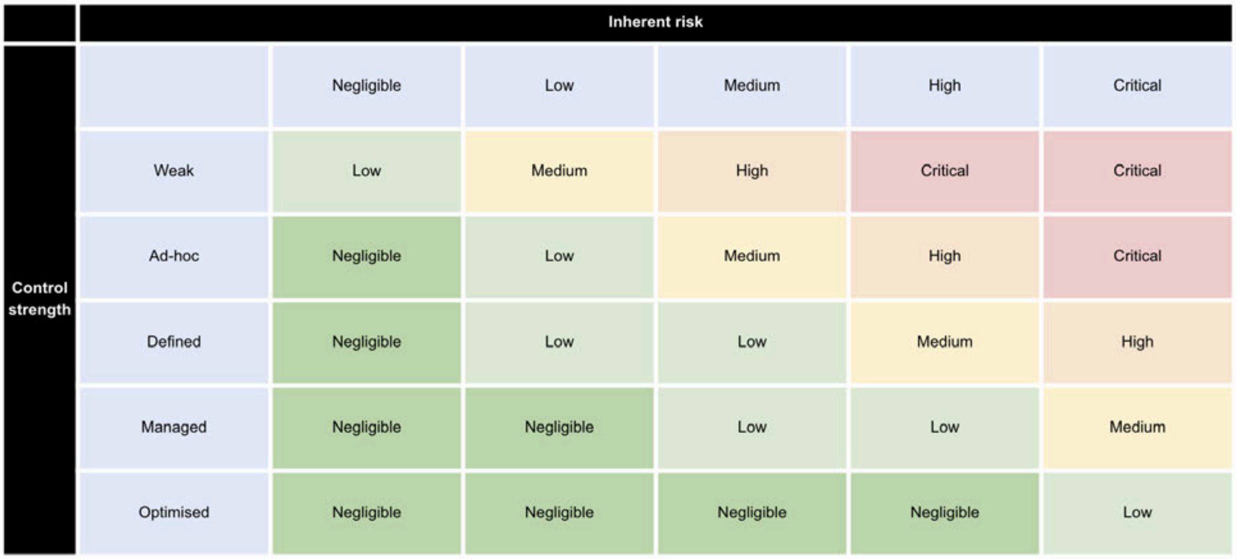
4. Control strength matrix



Strength		Description
5	Weak	• Mitigation measures are incomplete, informal, and inconsistent. • Processes are not defined, not repeatable, and should be improved.
4	Ad-hoc	• Mitigation measures are executed without standardised processes in place. • Processes may be ad hoc and are not well-defined. • There is scope of improving and formalising documentation practices.
3	Defined	• Mitigation measures are defined, formalised/documented, and repeatable. • There is some observability on how the measure works, and QA frameworks are in the process of being implemented. • Processes tend to be more proactive than reactive, well characterised and understood across all organisation verticals.
2	Managed	• Mitigation measures are well defined, formalised/documented and regularly managed, with repeatable QA in place. • There is an established process for integrating feedback to mitigate process deficiencies. • Processes are proactive, where possible, for all forms of content and behaviour.
1	Optimised	• Mitigation measures are comprehensively defined and operating at the highest quality, with mature QA in place • There are operationally effective controls in place, an applicable policy, applicable training, and regular testing and monitoring of the control. • The focus is on continuous improvement to maximise the effectiveness of resources, maintain resilience and robustness.

Fig.16: Control strength scale for the purpose of the DSA risk assessment

5. Residual risk matrix



		Inherent risk				
Control strength		Negligible	Low	Medium	High	Critical
	Weak	Low	Medium	High	Critical	Critical
	Ad-hoc	Negligible	Low	Medium	High	Critical
	Defined	Negligible	Low	Low	Medium	High
	Managed	Negligible	Negligible	Low	Low	Medium
	Optimised	Negligible	Negligible	Negligible	Negligible	Low

Fig.17: Residual risk matrix for the purpose of the DSA risk assessment