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

SEPTEMBER 2023

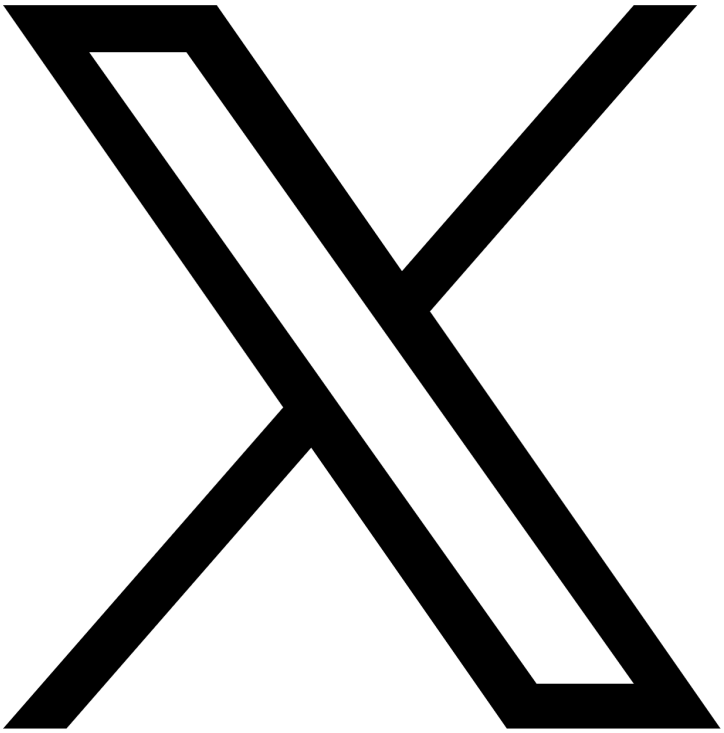


TABLE OF CONTENTS

I. Executive Summary

Today, platforms such as X form part of a multi-platform risk environment. Reports suggest that social media users typically use an average of 6-7 platforms each month. Within this ecosystem, X strives to be the town square of the internet by promoting and protecting freedom of expression. We have always understood that to reach this goal we must give everyone the power to create and share ideas and information instantly, without barriers.

With more than 45M monthly active users in the EU, X's main establishment in the EU was designated on 26 April 2023 as a very large online platform (VLOP) under the EU Digital Services Act (Regulation 2022/2065; the DSA)[1]. In compliance with DSA Article 34, we assessed how the systemic risks identified in the DSA may stem from the design, functioning, or use made of our services. Our risk assessment reflected X services at or around 31 July 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 of X. In accordance with DSA Article 34, our risk assessment process covers the four categories of systemic risks in 15 individual assessment areas.

For each identified risk area, we assessed how our platform's design, functioning, use, or potential misuse, could contribute to an inherent risk, mapped our existing controls and remediations against these inherent risks, and assessed the residual risk that remains on our platform. Following our assessment, we found that our existing controls bring down the level of risk for most areas to a low to medium level. Acknowledging that these systemic risks are continuously evolving and can be impacted by intentional coordinated exploitation, we remain committed to continuing to monitor and mitigate these risk areas.

We welcome the opportunity to continue enhancing our current mitigation system and introducing new mitigation measures, in line with Article 35. Our measures are tailored to address the Article 34 systemic risks and are proportional to the economic capacity of X and the need to avoid unnecessary restrictions on the use of our service – with special consideration given to the impact on freedom of expression. Our planned mitigations include improvements to our policies, content moderation systems (including detection and enforcement measures), and awareness raising measures. We are also committed to continually enhancing our internal data extraction processes for future risk assessments.

[1] Where we refer to 'X' in our report, we are referring to the X organisation as a whole, or the X platform. Twitter International Unlimited Company constitutes X's main establishment in the European Union and a very large online platform within the meaning of Article 33(1) of the DSA.

We have conducted this first 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, Trust & Safety, Product Engineering, Legal, Privacy & Data Protection, Compliance, and Government Affairs, along with external resources, were relied on to leverage industry knowledge and set the blueprint for future assessments. As this is a first assessment, in what will be an undertaking at least once every year from here on, this represents an inaugural review in an evolving and iterative process as envisaged in the DSA. We expect to review and refine our methodology and assessment process in the upcoming risk assessment cycles and welcome constructive feedback from stakeholders.

II. Introduction

With more than 45M monthly active users in the EU interacting on our services, X's main establishment in the EU was designated on 26 April 2023 as a very large online platform (VLOP) under the EU Digital Services Act (Regulation 2022/2065; the DSA). The DSA sets out risk assessment and risk mitigation obligations under which VLOPs must assess any systemic risks stemming from the design or functioning of their service and its related systems, and from the use made of their services. VLOPs must put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks, with particular consideration to the impacts of such measures on fundamental rights.

X's mission is to promote and protect the public conversation, serving as a trusted digital town square. The popularity of our services may attract individuals who, whether intentionally or not, may exploit our services causing societal concerns and negatively impacting the safety and wellbeing of our community. Our global teams joined forces over the last months leading up to the DSA effective date to develop and refine a risk assessment methodology that meets the DSA's requirements, and undertake an assessment across the four systemic risk categories identified in DSA Article 34.

We assessed the systemic risks stemming from the design, functioning, and use – as well as the potential misuse – by any users of X in the EU, in accordance with Article 34. The identification of risks and the assessment took into account the specific nature of our services in relation to these systemic risks. In determining the significance of the impact, we considered the severity and probability of such systemic risks, including how frequently they could occur, if they could negatively impact a large number of persons, their scope of harm, as well as their remediability. The report describes this methodology in detail, followed by the results of the individual risk assessments.

As part of our risk assessment, we analysed our existing controls to reduce the inherent risk and considered the deployment of further measures to mitigate the systemic risks analysed in the risk assessment, in accordance with Article 35. In assessing the controls in place and further mitigation measures, we took into account the residual risks identified, our economic capacity, and any impact on fundamental rights, in particular freedom of expression. These measures are described at the end of the report in the Mitigation Roadmap.

Documents and data that support the preparation of the assessment were preserved to allow for subsequent risk assessments to build on each other and observe the evolution of the risks identified in accordance with Article 34(3). The review reflected the design and functioning of X services at or around 31 July 2023 to allow time for assessment completion and review.

We have conducted this first DSA systemic risk assessment utilising our knowledge, resources, and understanding of DSA requirements, as well as considering established and emerging cross-industry standards. For the first DSA risk assessment and report, both internal and external resources were leveraged to incorporate industry knowledge and set the blueprint for future assessments. As this is a first assessment, in what will be an undertaking at least once every year from here on, this represents an inaugural review in an evolving and iterative process as envisaged in recital 85 of the DSA. We expect to review and refine our methodology and assessment process in the upcoming risk assessment cycles and welcome constructive feedback from stakeholders. In doing so, we will take into account new standards and in particular any DSA regulatory guidance where it may become available.

III. X Risk Environment and Controls

With an estimated 4.9B social media users worldwide, each of whom are reported to spread their digital footprint across an average of 6-7 platforms each month, the risks that manifest on each platform are representative of a multi-platform risk ecosystem. Many risks have societal and/or cultural roots, appearing online as extensions of often already rapidly evolving offline risks and interacting in complex and sometimes novel ways across the online platform ecosystem.

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. 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 Premium. Posts is our most popular feature that allows users to share any message which may contain photos, videos, links, and text. Spaces allows users to create live audio conversations that any logged in user can join, listen, and speak in. Communities was created to give people a dedicated place to connect, share, and get closer to the discussions they care about most. Finally, our subscription service X Premium allows users access to additional features, such as prioritised rankings in replies just like our legacy verified users, fewer ads, and longer posts, with the goal of elevating quality conversations on the platform.

To provide users with the most relevant content, we depend on our recommendation algorithm to distil the roughly 500M posts published daily down to a handful of top posts that ultimately show up across several areas of the app — e.g. For You feed, Search, Explore, Ads, and Notifications. Our users can choose between a For You or a Following timeline, allowing them to modify and personalise their content consumption. When surfacing content from outside our users' individual network, we strive to show content that each user would be most interested in and that contributes to the conversation in a meaningful way; this includes content that is relevant, credible, and safe. We have recently open-sourced our recommendation algorithm, aligned with our efforts to enter a new era of transparency and trust with our users, customers, and the general public. We hope to benefit from the collective intelligence and expertise of the global community in helping us identify issues and suggest improvements, ultimately leading to a better platform for everyone.

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. We monitor trends in online behaviour that may pose novel risks to our community's safety, and leverage internal subject matter experts to craft clear, nuanced, and scalable policies for the X community. Given the dynamic landscape of harms and trends across our platform, we recognize the need to be adaptable in our policy development and enforcement iterations to respond to emerging risks.

We are constantly taking feedback internally and externally to continuously improve our policies, including through cooperation and partnerships with various organisations such as regulators, law enforcement, nonprofits and other relevant stakeholders.

We employ a range of enforcement options, either on a specific piece of content (e.g., an individual post or Direct Message) or on an account, to enforce our policies on the platform. In determining what enforcement option to apply, we carefully consider that activity on X is largely reflective of real offline conversations, events, and social movements that may include perspectives that could be perceived as offensive, controversial, and/or bigoted by our users. In line with our mission to promote open conversation, we encourage a variety of perspectives on the platform. This is central to our *Freedom of Speech, Not Reach* (FoSnR) labelling that moves us away from a binary, absolutist take down/leave up moderation framework for certain policy areas, to a more reasonable, proportionate and effective moderation process. Such restricted posts receive 81% less reach or impressions on average and we proactively seek to prevent ads from appearing adjacent to content that we label. Our community has also 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. Nevertheless, we recognise that certain behaviours are simply unacceptable. We have policies in place to take strong enforcement action against illegal content, including child sexual abuse material (CSAM), violent hate speech, and terrorism content. Involvement with such behaviours will result in suspension from the platform following the first offence.

Along with our Terms of Service and policies, to empower our users to interact with the features of X safely, we have implemented a suite of product-level safety features. For example, X Premium and Verified Organizations come with defined controls in place including eligibility processes, and temporary loss of the verification checkmark as a result of certain actions or behaviours. Any violation, such as platform manipulation or circumvention of enforcement actions, can result in the loss of the checkmark or suspension. In doing so, these product features help defend against impersonation and aid in the reduction of inauthentic accounts and spam on X. Similarly, some of our monetization products - such as our newly launched Creator Ads Revenue Sharing - are only available for X Premium or Verified Organizations and are therefore subject to pre-screening prior to approval. This helps reduce the risk of monetized products being misused by bad actors.

Our diverse product-level safety features allow users to modify their experience and engagement on X to ensure each user is able to participate on the platform in a safe and meaningful way. For example, users can set conversation or interaction controls to filter who has the ability to comment or respond to their posts or to limit their interactions with some users. This feature has been extended to Direct Messages (DMs), where users may choose to limit their DM inbox to only

verified users and people they follow. Following the introduction of this feature, we have seen a 70% reduction in spam in DMs.

Beyond individual settings on the platform, we have recently introduced Community Notes to empower our users to create a better informed community. X users can collaboratively add context to potentially misleading posts or advertisements, and where enough contributors with different points of view rate that note as helpful, the note will be publicly shown on the post. For example, over 400 unique Community Notes have appeared on posts related to the Ukraine conflict, addressing a wide variety of topics. An even larger set — over 3.5K unique notes — have been proposed on posts related to Ukraine, demonstrating the potential for even greater scale of added context. These notes are in numerous languages — including English, French, Spanish, and Japanese — and are written specifically for local audiences. Community Notes exemplifies our platform's commitment to transition towards an enhanced community-based content moderation model that puts our users first always.

As X is a real-time global information service, a high proportion of users access the platform without logging into an X account. Permitting users to access X content without logging into an X account is fundamental to X's mission to serve the public conversation and help ensure the freedom of expression and access to information of its users. By default, X sets high privacy, safety and security settings for these logged-out users to help ensure a safe user experience. Per ourTwitter's February 2023 report on Average Monthly Active Recipients of Service (AMARS) in the EU, measured over a 45 day period, an average of 41.1 million Twitter users had this logged-out experience, while 59.8 million Twitter users logged into an account to access Twitter content. Our February to July numbers are 60.9 million for Logged In Users with 51.3 million Logged Out Guests.

X is often a key platform for information in times of world crises. People use our platform to access and share information, raise awareness about the situations they are in on the ground, and openly and freely exchange ideas on a wide array of topics. Because the decisions platforms and users make can have real-world consequences, we are committed to protecting the people we serve around the world. This means that in response to such real-world crises, we have processes that enable us to make risk-informed decisions, allocate resources, and apply timely and appropriate remediation measures. In order to fulfil this responsibility, we monitor potential crisis hotspots that can activate our crisis response protocol, which includes processes for monitoring, detection, issuance of sweep guidance and activation of control measures.

Our work to address the ever evolving risks in the online space is continuously ongoing. This annual risk assessment process coincides with our increased investments in people, policy and product that will further ensure our communities have access to an open, accurate and safe space for discourse on X.

IV. X DSA Systemic Risk Governance Framework

As part of DSA preparation, we established an EU Digital Services Act Compliance Governance Charter. Under this Charter, a DSA Leadership team has been created to define, oversee and drive accountability for the implementation of our DSA compliance governance arrangements in a manner that addresses the sound management of systemic risks identified by X pursuant to Article 34.

As part of further DSA effective date readiness, we are building our DSA Systemic Risk Governance Framework to frame our ongoing approach. This overall framework confirms our leadership's understanding and commitment to meeting its Article 41 management body obligations with respect to governance arrangements and managing, monitoring and mitigating risks identified pursuant to Article 34.

Further, the DSA Systemic Risk Governance Framework 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 risks identified pursuant to Article 34.

Acknowledging that the Commission can adopt a decision requiring 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 also sets out a process for responding to requirements under the crisis response mechanism.

V. Methodology

We welcome the DSA's approach to conducting the risk assessment with a focus on how our platform may be contributing to systemic risks that can cause harm to our users and the general EU population, both on and off our platform.

Pending specific regulatory guidance on the DSA risk assessments, our risk assessment methodology was developed by referencing a number of existing frameworks, including, but not limited to, the UN Guiding Principles on Business and Human Rights, the DTSP Safe Framework, the Human rights impact assessment guidance and toolbox from the Danish Institute for Human Rights, and the human rights impact assessment reports by BSR. We have also consulted internal and external subject matter experts to ensure that this first DSA risk assessment works as a robust blueprint for the annual risk assessments to come.

Our risk assessment reflected X's services at or around 31 July. As per Article 34(1) DSA, the risk assessment process covered the four categories of systemic risks[2] and the following recitals complementing Article 34 were also considered: 12, 79, 80, 81, 82, 83, 84, 85, 89 and 90.

A. Walkthrough

1. Assessment structure

Fig.1: Four phase process to risk assessment.

[2]

a) *the dissemination of illegal content through their services;*
b) *any actual or foreseeable negative effects for the exercise of fundamental rights, in particular the fundamental rights to human dignity enshrined in Article 1 of the Charter, to respect for private and family life enshrined in Article 7 of the Charter, to the protection of personal data enshrined in Article 8 of the Charter, to freedom of expression and information, including the freedom and pluralism of the media, enshrined in Article 11 of the Charter, to non-discrimination enshrined in Article 21 of the Charter, to respect for the rights of the child enshrined in Article 24 of the Charter and to a high-level of consumer protection enshrined in Article 38 of the Charter;*
c) *any actual or foreseeable negative effects on civic discourse and electoral processes, and public security;*
d) *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 person's physical and mental well-being.*

We adopted a four phase process in approaching the risk assessment exercise:

Phase I: identification of systemic risks

To begin the risk assessment process, we created a DSA Risk Registry that deconstructed Article 34 into its sub articles and related recitals. From these, we extracted individual risks mentioned in the DSA that would need to be assessed.

Recognising that certain risks do not exist in isolation, we grouped some of the risk subcategories together. This reduced duplication, especially in the case of similarities in the way the risk manifests or in the way the risk is mitigated. This decision was also informed by the nature of our platform, and our existing frameworks for risk identification. Irrespective of how they were grouped, each area of potential harm was fully assessed in its respective risk assessment.

The following table provides an overview of how we organised the risk subcategories into 15 independent risk assessments:

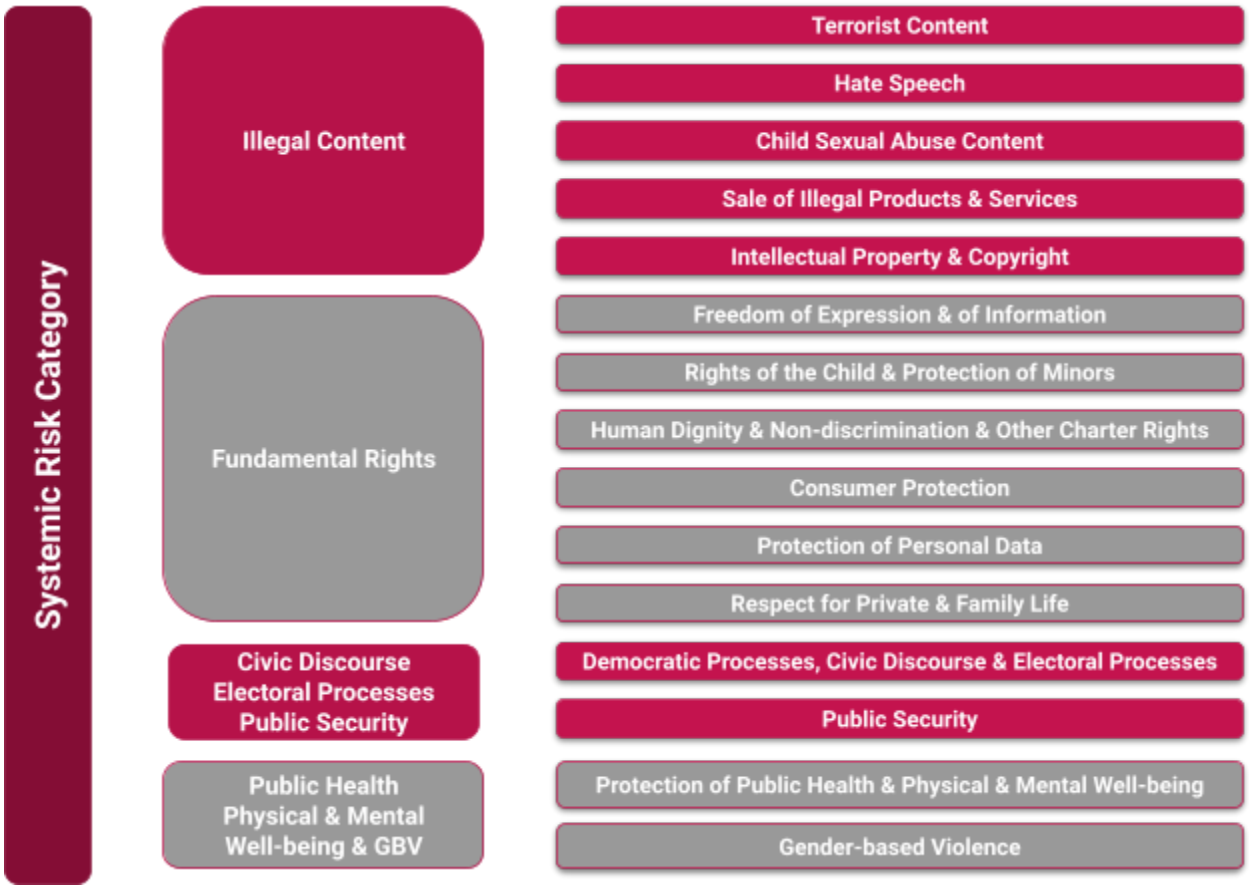


Fig. 2: Risk categories and subcategories.

With regards to **illegal content,** we identified five core subcategories of systemic risks. These assessments analyse how the platform can be used and misused to disseminate illegal content – including hate speech, CSAM, terrorist content, copyrighted content, and illegal products and services – and what policies and controls we have in place to mitigate these risks.

The DSA places a strong emphasis on the protection of **fundamental rights**, which is consistent with the principles that underpin X's policies. While Article 34 highlights some fundamental rights, its scope includes the entire EU Charter. Following analysis, we grouped the charter rights into six separate assessments to be able to thoroughly assess the actual or foreseeable **negative effects for the exercise of fundamental rights**. Given that rights of the child are intrinsically linked with the protection of minors, specified in Article 34(1)(d), we pooled these assessments. For this specific, first-time assessment, we determined that it was most logical to combine our analysis of human dignity, non-discrimination and other Charter rights.

In mapping the third systemic risk group—**civic discourse, electoral processes, and public security**—to our platform's specific risk environment, we split the risk into two subcategories. **Public security** required a standalone analysis as it can transcend into other areas beyond those identified within the scope of our risk analysis for **democratic processes, civic discourse, and electoral processes** on X.

The fourth systemic risk group deals with two subcategories: one on the actual or foreseeable negative effects in relation to the protection of **public health** and serious negative consequences to **physical and mental well-being**, and the other in relation to **gender-based violence** where we also deemed it appropriate to address illegal pornographic content[3] because of the similarity of the harms posed by these two categories on our platform.

Risk assessment templates

Based on our phased approach (Figure 1) and our reading of the DSA, we produced a risk assessment template that was used to assess each risk subcategory.

The template was used to frame the review of inherent and residual risks in each subcategory and across the four systemic risk categories overall. It also supported mapping of control environments for this specific assessment to inform *mitigation measures* review at the end of each exercise.

[3] Recital 87 of the DSA refers to 'illegal pornographic content' and defines it as: "Providers of very large online platforms, in particular those primarily used for the dissemination to the public of pornographic content, should diligently meet all their obligations under this Regulation in respect of illegal content constituting cyber violence, including illegal pornographic content, especially with regard to ensuring that victims can effectively exercise their rights in relation to content representing non-consensual sharing of intimate or manipulated material through the rapid processing of notices and removal of such content without undue delay."

Per Article 34(2), as well as recitals 79-85 and 89-90, we took into consideration if and to what extent the following factors influence each risk subcategory:

- The use and potential misuse of our platform;
- Our policies and terms and conditions;
- The design and functionalities of our platform;
- The design of our recommender systems and any other relevant parameters;
- Our content moderation systems;
- Our systems for advertisements; and
- And any other applicable measures.

Our data-related practices were comprehensively addressed in a risk assessment dedicated to data protection to avoid duplication.

In addition, we analysed if, and to what extent, the systemic risks identified could be influenced by "intentional manipulation", "inauthentic use" - as defined in recital 84 of the DSA -, automated means, and/or the amplification of potentially illegal or otherwise violative content or behaviour that is against our terms and conditions. Additionally, where possible we considered regional and linguistic dimensions, where data points and examples were available as they related to Member States. As we operate and apply our policies globally, many of the data points below represent global numbers, unless explicitly mentioned otherwise.

As part of post completion of this first assessment, we will assess opportunities for future enhancements of our risk assessment methodology. We are also committed to engaging with any feedback from the Commission, other regulators and stakeholders, as well as assessing emerging standards to inform future enhancements to our approach.

2. Risk assessment framework

Taking note of other industries' risk assessment methodologies, the various human rights impact assessments that have been conducted on online platforms, as well as relying on the guidance provided in recital 79 of the DSA with regards to probability and severity, we quantified the risks of the platform and the existing controls through:

- Inherent risk matrix: A 5 x 5 matrix to quantify the systemic inherent risks in the EU stemming from the design, functioning, or use made of our service. This matrix maps the probability against the severity of the identified risks.
- Control strength: A 1 - 5 scale that qualifies the existing controls and mitigating measures applied to the identified inherent risks.

- Residual risk matrix: A 5 x 5 matrix to quantify the residual risk remaining on the platform, following an assessment of our existing controls. This matrix maps our control strength against the inherent risk.

Phase II: Assessment of inherent risk

We assessed the inherent risk as a function of probability and severity of the risk, where severity can be understood through the scope, scale, and remediability of the harm.

Probability: For this assessment, we considered probability as the likelihood of a risk manifesting on our platform. Our 5-level scale is adjusted relative to the unique real-time nature of the X platform, where thousands of posts are created every second. Thus, we consider *very unlikely* those rare events that may occur yearly, while *almost certain* those that occur daily. (Refer to Annex 1).

Severity: For this assessment, we considered severity as a function of three variables: scope, scale, and remediability. This was informed both by the United Nations Guiding Principles for businesses and human rights, as well as our own harm taxonomy - a shared internal taxonomy that sets X's approach to harm and a common framework to identify and quantify propensity for harm by utilising specific and defined criteria.

Scope: Borrowing from the work done in developing our internal harm taxonomy, we consider physical, psychological, informational, economic, and societal harms as part of the scope of the risk. These variables give us an indication of the *gravity* of the harm to the population impacted by the risk. Scope varies from very low, where there is very low gravity of any harm, to very high, where there is a very high gravity of any harm - especially physical and/or psychological harm.

Scale: Scale considers how widespread the impact of the risk can be on users. As such, it varies from very low, where it impacts a negligible amount of people, to very high, where it impacts most users of the platform, as well as the general public.

Remediability: Remediability considers the ease with which the situation could be restored to what it was before the impact of the risk, or rather, how easy it is to reverse the impact of the risk. This ranges from remediable, where any remedy provided will fully restore the person/situation to the state before the impact, to not remediable, where the state before the impact cannot be returned to at all.

With reference to the UN Guiding Principles, we weigh scope as having the biggest impact, with scale and remediability as having secondary and tertiary impacts on severity. This also comes from an understanding that, if the risk of physical and/or psychological harm is high or very high,

we consider these types of harm to have more of a negative effect than other types of potential harm. Finally, remediability is given a lesser weight as it is an added component to our assessment of severity, rather than a central variable.

Mapping inherent risk on a matrix

To map probability and severity on a 5 x 5 matrix, we multiplied the corresponding values to provide us with a grading of inherent risk, as seen in the figure below, subject to certain modifications for crisis events as discussed further below.

		Severity				
		Very low severity (1)	**Low severity (2)**	**Moderate severity (3)**	**High severity (4)**	**Very high severity (5)**
Probability	**Almost certain (5)**	Low (5)	Medium (10)	High (15)	Critical (20)	Critical (25)
	Likely (4)	Negligible (4)	Low (8)	Medium (12)	High (16)	Critical (20)
	Possible (3)	Negligible (3)	Low (6)	Low (9)	Medium (12)	High (15)
	Unlikely (2)	Negligible (2)	Negligible (4)	Low (6)	Low (8)	High (10)
	Very unlikely (1)	Negligible (1)	Negligible (2)	Negligible (3)	Negligible (4)	High (5)

Fig.3: Inherent risk matrix.

Crisis events, with potential to lead to real-world consequences, sit at the intersection of very high severity and very unlikely probability. However, it is very difficult to foresee what such a crisis will be and when it will occur. This results in the potential for crisis events to pose a higher risk than this framework can provide, and action reflecting high to critical risks may be needed from our teams in such cases. We accordingly increased the risk level to high for very high severity situations, approaching this inherent risk with an appropriately tailored response that is reasonable, proportionate, and effective with particular consideration to the impacts on fundamental rights.

The scorecard below explains what the various inherent risk levels mean. This 1-25 scale is also translated into a 1-5 scale, in line with the control strengths, to facilitate the residual risk calculation.

1-25 scale	Inherent risk score card	1-5 scale
20-25 Critical	Implies a critical risk, expected to have a very high scope of harm on the most number of people, with irreversibility, or a very high difficulty to remedy and restore the situation prevailing prior to the potential impact, without controls.	5 Critical
15 - 19 High	Implies a high risk, expected to have a high scope of harm on a large number of people, with potential irreversibility, or difficulty to remedy and restore the situation prevailing prior to the potential impact, without controls.	4 High
10-14 Medium	Implies a medium risk, expected to have a moderate scope of harm on a moderate number of people, with possible reversibility or possibility to remedy and restore the situation prevailing prior to the potential impact, without controls.	3 Medium
5-9 Low	Implies a low risk, expected to have a low scope of harm on a minimal/low number of people, with likely reversibility or likely way to remedy the risk and restore the situation prevailing prior to the potential impact, without controls.	2 Low
1 - 4 Negligible	Implies a negligible risk or no foreseeable risk. If there is any foreseeable risk, it has very low impact on a very low number of people, and is reversible or remedied without difficulty, without controls.	1 Negligible

Fig. 4: Inherent risk scorecard

Phase III: Assessment of mitigation measures and safety environment

As a platform that strives to protect its community, which includes respecting the right to free speech and expression, we have a number of controls in place that mitigate systemic risks on our platform. We have developed a scale to assess the strength of these controls, which considers the completeness/operationality of our control, its effectiveness, and whether it has an established process for improvement. Recognising that systemic risks are constantly evolving, our optimised strength qualification takes into account the focus on continuous improvement to maximise effectiveness.

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 do not have 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, documented, formalised, and repeatable. Processes are proactive, well characterised and understood across the organisation.

2	**Managed**	Mitigation measures are sufficiently defined, documented and regularly managed. There is a set process for integrating feedback to mitigate process deficiencies.
1	**Optimised**	Mitigation measures are comprehensively defined and operating at the highest quality. There are operationally effective controls in place, based on 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. 5: Control strength scale

Phase IV: Identification of residual risk

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 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. Moreover, 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.

For this assessment, the residual risk is derived from multiplying the inherent risk scores and the strength of our controls scores (Residual risk = inherent risk score x control strength score), as shown in the residual risk matrix below:

		Inherent Risk				
Control strength		Negligible (1)	Low (2)	Medium (3)	High (4)	Critical (5)
	Weak (5)	Low (5)	Medium (10)	High (15)	Critical (20)	Critical (25)
	Ad-hoc (4)	Negligible (4)	Low (8)	Medium (12)	High (16)	Critical (20)
	Defined (3)	Negligible (3)	Low (6)	Low (9)	Medium (12)	High (15)
	Managed (2)	Negligible (2)	Negligible (4)	Low (6)	Low (8)	Medium (10)
	Optimised (1)	Negligible (1)	Negligible (2)	Negligible (3)	Negligible (4)	Low (5)

Fig. 6: Residual risk matrix

The scorecard below explains what the various residual risk levels mean.

Residual risk score card	
20-25 Critical	Implies a critical risk, expected to have a very high scope of harm on the most number of people, with irreversibility, or a very high difficulty to remedy and restore the situation prevailing prior to the potential impact, despite controls.
15 - 19 High	Implies a high risk, expected to have a high scope of harm on a large number of people, with potential irreversibility, or difficulty to remedy and restore the situation prevailing prior to the potential impact, despite controls.
10-14 Medium	Implies a medium risk, expected to have a moderate scope of harm on a moderate number of people, with possible reversibility or possibility to remedy and restore the situation prevailing prior to the potential impact, despite controls.
5-9 Low	Implies a low risk, expected to have a low scope of harm on a minimal/low number of people, with likely reversibility or likely way to remedy the risk and restore the situation prevailing prior to the potential impact, despite controls.
1 - 4 Negligible	Implies a negligible risk or no foreseeable risk. If there is any foreseeable risk, it has very low impact on a very low number of people, and is reversible or remedied without difficulty.

Fig. 7: Residual risk scorecard

B. Mitigation measures

Our 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. Our mitigations range from product features like Mute that empower our users to protect themselves from content and accounts that they would consider harmful to enforcement mechanisms such as FoSnR, which mitigate potentially unnecessary restrictions on a user's freedom of expression and right to information on X.

Following our risk assessment across all four systemic risk categories, we identified that certain mitigation measures address all systemic risks horizontally. We have planned mitigation measures, distinguishing between horizontal measures that address cross-cutting risks, and those measures that can more precisely target a specific systemic risk. In accordance with Article 35(1), many of the measures we've already built or plan to build are already aligned with our general compliance efforts towards the DSA. We will monitor these measures to see how they work to proportionately address the residual risks identified in this exercise and build on these results in our next assessment.

C. Stakeholder engagement and consultation

We have consulted with external and internal experts and drawn from their advice and expertise to inform our assessment. As a first of its kind risk assessment, we relied on subject matter experts for a comprehensive assessment of the risks and our policy and cross-functional teams for input on a proportionate and adequate set of recommendations to mitigate them, in line with the requirements set out in the DSA.

Internal awareness sharing, training, consultations and reviews have been conducted throughout the process across globally based teams and with our leadership, including Trust & Safety, Product Engineering, Legal, Privacy & Data Protection, Compliance, and Government Affairs. X management was consistently engaged, reviewed and approved our assessment strategy, and was actively involved in the decisions related to the risk management assessment.

Additionally, members of our teams attended stakeholder, industry and DSA specific events such as the Digital Services Act Stakeholder Event: Shaping the Future of Digital Services of June 27 2023. There, we took the opportunity to exchange with civil society organisations (CSO), academia, the industry and various other participants on the enforcement of the text, and attended the workshop titled ‘‘*Conducting DSA Risk Assessment - Algorithms in the Spotlight.*’’

Between June 21 and June 22 2023, and following multiple earlier consultations, X conducted the first VLOP readiness check with the European Commission DGConnect in its global headquarters in San Francisco. We presented our preparations for the DSA, with particular focus on our work on algorithmic transparency, countering illegal hate speech and child sexual exploitation, disinformation, and interference in electoral processes. These technical deep-dives were followed with a high-level executive summary meeting, with the participation of EC Commissioner Thierry Breton and X leadership.

We continually engage with CSOs to discuss our work on harmful content and engage in meaningful dialogue with them. Below are some non-exhaustive examples:

- X is part of the Online Hate Observatory in France created under the Avia Law, and has a regular dialogue with CSOs who are members of the Observatory. In the run up to the DSA, the Observatory meetings were used to discuss DSA preparation, and some key aspects of the text, for instance collaboration with Trusted Flaggers.
- X is regularly in contact with the Délégation Interministérielle à la Lutte Contre le Racisme, l'Antisémitisme et la Haine anti-LGBT (DILCRAH) in France. Our work with DILCRAH

focuses on various forms of online hate on our service and ways to address them. Additionally, we hold regular meetings with Law enforcement in France on operational cooperation.

- X also participates at the annual Conference of the INACH network, gathering online hate CSOs from all over Europe.
- X is a founding member of the Global Internet Forum to Counter Terrorism, which convenes through a multi-stakeholder process a range of civil society organisations and partnership with academics a range of insights and working group outputs to enhance work in the terrorist and violent extremist space.

This methodology has been developed for the sole purpose of the DSA's first risk assessment process highlighted in Article 34. Due to the wide scope and complexity of the risks, there are inherent limitations to our assessment. Indeed, the difficulty to consider and measure all the factors and their impact(s) on the different systemic risks may lead to a non-exhaustive list of the parameters integrated in our assessment. The results and conclusions from the risk assessment exercise included in this report were drafted for the limited and specific purposes of the DSA and should not be used for any other purpose, including for other regulatory or litigation purposes either within or outside of the EU. Further, the inherent and residual risk scores within this report should be considered and understood in the context of the entirety of the relevant risk assessment and not in isolation.

VI. Summary of risk assessments

Fifteen individual risk assessments were conducted and categorised under the four systemic risk categories as identified by the DSA. These individual assessments provided a framework to assess how our features and functionalities could contribute to systemic risks, and surfaced the potential residual risks on our platform for mitigation review. Below, we set out the key results from each assessment.

A. Dissemination of illegal content

We do not allow the use of X for any unlawful behaviour or to further illegal activities including violent hate speech and terrorist accounts and have a zero tolerance policy towards the dissemination of child sexual abuse material (CSAM). As we build our enforcement approaches, we pay due regard to their proportionality and effectiveness to address these violations and provide an effective appeals process for users to contest our decisions.

For this systemic risk, the inherent risk score across the content subcategories ranges from Medium (e.g. IP Rights) to Critical (e.g. CSAM), offset by a control strength range of Defined to Managed. As a result, the residual risk for this area varies from Low to High.

Inherent risks

There is always an inherent risk that bad actors misuse platforms like ours to disseminate illegal content. An example is *Operation Shapeshift*, which was a global cross-platform attack where threat actors pretended inauthentically to be members of a protected community to harass others and create a backlash against the protected community. The use of slurs or new hateful terms with regard to hate speech, or the use of slang to share CSAM, are examples of areas where language and related code words can stay one step ahead of platform policies and integrity efforts. While our systems do not intentionally promote this content, we also recognise that recommendation systems and algorithms are not invulnerable to manipulation and could potentially promote or amplify violating content before we can detect it and enforce against it.

Controls

X prioritises developing and implementing robust policies and protocols to address the dissemination of illegal content. This includes targeted policies against terrorist content, hate speech and unlawful discriminatory content, CSAM, and illegal products and services. These policies are enforced using a wide range of measures, from content removal to restricting reach and visibility of posts, and a proportional and defined set of sanctions against violating users, from account restrictions to account removal in the most severe cases. Further, our monetization

features are only available for X Premium users, and where subscribers violate our policies their status may be revoked or their account suspended.

To mitigate against illegal content in advertisements, at the ad creation time, our system proactively seeks to ensure that advertisers comply with our Ad Policies. Ads on X are also expected to adhere to our Terms of Service, including policies that prohibit illegal content. We additionally employ machine learning models and business logics such as denylist terms restricting content from appearing on promoted posts. [redacted]

CSAM:
Eliminating CSAM on X is one of our key goals. Our robust child sexual exploitation policy forms the foundation for our enforcement against this illegal behaviour, and was expanded in early 2023 to more effectively combat potential CSAM-related behaviours. We have product features designed to protect minors, including automatically setting known minor accounts to "protected", restricting other users from direct messaging a known minor account unless they follow them, and an age lock that prevents an account from altering their date of birth once one is entered that indicates the user is under 18 years of age.

However, we know these features alone cannot prevent CSAM on X, so we also have automated systems to detect and enforce against CSAM. These include machine learning models, heuristic-based rules, media hashing, and a media classification scorer. This incorporates automated review and actioning on hash matches shared by the National Center for Missing & Exploited Children and other industry partners. In rare cases, our automated systems may fail to detect violations due to technical or operational reasons, or because of evasions by bad actors - as highlighted by a Stanford report. In these cases, we immediately take action when we become aware of performance issues. These automated detections are supplemented by a robust manual content review process, where we review 100% of all CSAM reports that we receive using [redacted] specialised agents based in the EU region with relevant language expertise. Our content moderators working in child safety are provided access to wellness services to help protect their mental wellbeing.

As a result of these robust control mechanisms, we have actioned [redacted] accounts globally this year that created, distributed, or engaged with CSAM (as compared to [redacted] accounts actioned in H1, 2022). Our heuristic-based rules alone suspended [redacted] users for CSAM-related violations in July, 2023. Of our [redacted] total suspensions, about [redacted] of those were suspended when they created a new violating account or when they attempted to upload known CSAM, and about [redacted] of the accounts signed up via an EU IP address. The remainder were users who engaged

with known content, recidivist accounts, or users detected through reports and trends analysis. Approximately [redacted] of the suspensions in 2023 came from automating existing policies forbidding user engagement with known CSAM, which was previously enforced manually by agents. We also trained an additional [redacted] agents on NCMEC reporting–[redacted]–and invested in automated reports for media hash matches with known CSAM. This has allowed us to submit around [redacted] NCMEC reports in H1 2023 compared to [redacted] in H1 2022. Of the approximately [redacted] CSE appeals reviewed in H1 2023, only [redacted] resulted in remediation reversal. In parallel, we also detect and enforce against tactics used by bad actors in this risk area, including inauthentic accounts and platform manipulation.

Terrorist content:
We take a similar approach to terrorist content, based on our Violent and Hateful Entities and Violent Speech policies. At the product level, our monetized products include detection for sanctioned entities, and we block keywords associated with terrorist organisations from Search Autocomplete and Trending topics. [redacted] Finally, we leverage manual reviews based on proactive investigations and user reports where our automated detections fail or lack the precision to automatically enforce. Additionally, X has crisis protocols in place to act swiftly and scale our response to adapt to the rapid dissemination of violating content, and we actively collaborate with others in the industry via the GIFCT Crisis Incident Protocol, the EUIF Crisis Protocol and the Christchurch Call to contain the dissemination of such content across platforms. As a result of these control systems, we suspend an average of [redacted] accounts globally each month.

Illegal hate speech:
X protects against any kind of content that threatens or incites violence on the platform through the aforementioned Violent Speech policy, as well as strict policies against hateful conduct as well as abuse and harassment. Our FoSnR principle is imbued in our product through restricting reach and visibility of hateful speech. Along with this, features such as account filters and controlling replies work to further restrict reach while also protecting our users. Complementing these features, we have automated detection of violations of policies in this area via a combination of models and heuristic-based rules, as well as media matching (i.e. comparison of hashes extracted from media for the purpose of comparison with other media uploaded to the platform). Finally, we have dedicated operational teams responsible for responding to user reports in an accurate and timely manner. With regard to hate speech in particular, the operational teams encompass a diverse array of geopolitical and language expertise to ensure context and cultural nuance are taken into consideration as part of enforcement. For slurs and tropes, for instance, X uses glossaries specific to EU languages.

Furthermore, X is a Member of the EU Code of Conduct on countering Illegal Hate Speech online since its creation and actively engaged in this exercise and in the revision of the Code. During the 7th Monitoring Round, X was found to address notices between 24 hours in 54.3% of the cases and within 48 hours in 28.9% of the cases. X is also part of the Online Hate Observatory in France created under the Avia Law, and has a regular dialogue with CSOs who are members of the Observatory. In the recent Arcom report on Combating the dissemination of hateful content online[4], X is praised for effective communication with law enforcement through the Legos portal, which Arcom saw as a "particularly useful in tracing and guaranteeing the authenticity of requests and requesters".

As a result of our controls, our data has shown that more than 99.99% of post impressions are on content that is deemed "healthy". Less than 0.01% of post impressions contain hateful language.

Residual risk

Overall, the potential residual risk of dissemination of illegal content through our service after applying control measures is mostly medium (up to high for terrorist content) due to the high to critical inherent risk nature, balanced against strong control measures. Terrorist content and CSAM are deemed to have a higher inherent risk given that without controls, they have very high scale and scope and very low remediability. While the control measures are robust and proactive, the nature of the risk itself requires vigilance. This type of illegal content and behaviour can be dependent on new coded language trends that develop extremely rapidly and are created with the specific purpose to circumvent the rules. In contrast, the potential for residual risk is low for illegal content infringing IP rights, in part due to low scale and scope and the fact that it is highly remediable, together with a well-established and managed set of control measures leading to a lower level of residual risk.

We will continue to evaluate these risks and our controls as they may continue to evolve; for example, as generative AI tools improve rapidly and become more widely available. Such tools may facilitate the production of AI generated illegal content or make it more complex for the protection of a user's intellectual property rights. Our efforts to continue to address residual risk are detailed in our mitigation roadmap.

4 https://www.arcom.fr/nos-ressources/etudes-et-donnees/mediatheque/lutte-contre-la-diffusion-de-contenus-haineux-en-ligne-bilan-des-moyens-mis-en-oeuvre-par-les-plateformes-en-ligne-en-2022-et-perspectives

Risk assessment: Child Sexual Abuse Material

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, disseminates child sexual abuse content through the service in the EU.

Inherent risk	Controls	Residual risk
• There is a risk that the **availability of child sexual abuse material** may lead to revictimization of victims of sexual abuse. • There is a risk of **inaccurate distinction** between an adult and minor, as consensual adult nudity and depiction of sexual activity is allowed on the platform. • There is a risk that **groomers or predators** may consume depictions of minor nudity that may not be sexualised. • There is a risk that bad actors may misuse the **anonymity features** to operate pseudonymous accounts to groom minors. • There is a risk of potential **misuse/abuse of features** such as Direct Messages to share CSAM, and Spaces where there is only reactive detection of CSAM.	• **CSAM policy:** Zero tolerance policy towards any material that features or promotes child sexual exploitation. We immediately suspend accounts that violate this policy, including detected recidivist accounts. • **Reporting mechanisms:** Every piece of content is reportable on our platform. Users can report CSAM through our reporting channels and we also have a dedicated help centre page on this topic. • **Exhaustive content moderation:** Our content moderators provide services 24/7 and we take action both proactively and reactively ([redacted] accounts removed in H1, 2023). • **Automatic detection:** We have [redacted] heuristic-based rules specifically designed to detect CSAM-related spam that suspended [redacted] users last month (July, 2023). • **Exhaustive reporting:** We report known and unknown CSAM regardless of whether the depicted minor could	• There remains a residual risk due to the **ever-evolving nature of the space** and cross-industry ecosystem (for e.g. file sharing services, merchants and payment providers' involvement). • There is a risk that our **content moderation systems** may not be able to detect 100% of recidivist CSAM-related accounts and activity due to the rapidly evolving nature of the threat and tactics. For example, our automated text detection may not be sufficient to detect all cases of CSAM. • As Generative AI tools continue to improve and evolve quickly, a residual risk may manifest from bad actors seeking to leverage such technology in relation to CSAM. We will continue to work to understand and detect use of such tools to evade our enforcement.

	be at risk or not. We reported [redacted] instances in H1, 2023. • **Restricted high-risk terms:** X maintains a list of related keywords and phrases that are blocked from Trending. • **Hash-sharing (NCMEC and industry partners):** We leverage a combination of technology solutions to detect violating accounts, including PhotoDNA and internal proprietary tools. For videos we use a proprietary hashing algorithm produced by Thorn. • **External engagements:** Active collaboration and partnerships with NCMEC, The Internet Watch Foundation, The Tech Coalition, Point de Contact, Child Protection Lab, e-Enfance. • **Law enforcement engagement:** Cooperation with Law Enforcement via our dedicated LEGOS online portal for information request submission.	
Inherent risk score: 25, Critical Risk Probability: 5, Almost Certain Severity: 5, Very High Severity *Scope: 4; Scale: 5; Remediability: 5.*	Control score: 2.5, Defined	Residual risk score: 12.5, Medium Risk

Risk assessment: Terrorist content

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, disseminates illegal terrorism content through the service in the EU.

Inherent risk	Controls	Residual risk
• There is a risk that the **ease of account creation** can encourage terrorist organisations (TOs) and extremists to generate accounts, facilitate coordinated efforts for radicalization and recruit users to join entities that violate our Violent and Hateful Entities policy (VHE). • There is a risk that the **anonymity and pseudonymity features** are misused by individuals to engage in abusive or harmful behaviour, often without facing direct consequences in real life. • There is a risk that TOs may exploit the **hashtag function** gap to take over trending topics with harmful and violent content, gaining visibility.		• There remains a residual risk that extremist groups may **circumvent controls** by frequently employing coded language and updating their symbols to bypass moderation efforts, as well as obfuscating keywords and manipulating their images by cropping or adjusting colours to circumvent the automated systems that rely on such cues. • There remains a residual risk that TOs and extremists may **gain familiarity with our detection methods**. The recurrent suspension of certain groups signals to them which keywords to avoid.

• There is a risk that TOs and extremists may use **post replies** to spread violent propaganda under an otherwise healthy post. We have observed this in replies to posts by users with a large follower base due to the high visibility of their content.	Organisations, who need to meet all eligibility criteria, which mitigates against TOs' ability to leverage this feature. • **Threat intelligence:** We perform automated signal collection to better understand how TOs and violent entities are using our platform. • **Perpetrator of Violent Attacks crisis protocol**: X activates Perpetrators of Violent Attacks enforcement guidance immediately after real world events that constitute a violent attack. This enables swift, proactive removal of violative content, as well as suspension of the attacker's account, and scaled removal of manifestos. • **Automated detection & lead generation:** [redacted] • **External engagements:** X is a participant of the EU Internet Forum, is a founding member of the GIFCT, is a signatory of the Christchurch Call, and works with Tech Against Terrorism.	• There remains a residual risk stemming from **"known unknowns",** where TOs and extremists might refrain from overtly promoting their ties to a violative entity on X but could maintain associations with such groups. This risk becomes more pronounced when a user cultivates a significant follower base, recruits through direct messages, or chooses to reveal their affiliation at a later point.

Inherent risk score: 25, Critical Risk Probability: 5, Almost Certain Severity: 5, Very High Severity *Scope: 5; Scale: 5; Remediability: 4.*	Control score: 3, Defined	Residual risk score: 15, High Risk

Risk assessment: Illegal hate speech

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, disseminates illegal hate speech content through the service in the EU.

Inherent risk	Controls	Residual risk
• There is a risk that **commenting** under a post leads to purposeful exposure to hateful commentary, as **@mentioning** the author of the original post will notify the author. • There is a risk that, in cases where our models and filters are unable to catch violating content, **Out-of-Network injections** could potentially be a vector for amplification of hateful rhetoric. • There is a risk that **inauthentic accounts and activity** can be used to drive hateful content, amplify it, or target and harass individuals.	• **Freedom of Speech Not Reach:** We restrict the reach and visibility of hateful speech under the FoSNR approach, and trigger users to remove posts containing hateful speech or targeted harassment. For content that does not violate our terms of service, but may violate local laws, we would restrict access to such content in the country in accordance with applicable law. • **Account filters:** This feature allows users to mute notifications from certain categories of users, such as those with accounts who have not confirmed their phone number or email address, new accounts, accounts who have a default profile photo, accounts that the user does not follow or accounts that do not follow the user. • **Reply controls:** The default position is that everyone can reply but options are available to turn off all replies or only allow the accounts mentioned in	• There remains a residual risk that we may still **miss cases of harmful content** on the platform or users may upload that content again. • There remains a residual risk that our **platform policies could be applied unequally or in a subjective manner**, for example due to moderator bias, or language specialisation (or lack thereof). • False negatives that could result from automated content moderation tools (e.g. due to precision or recall issues) as well as manual moderation decisions.

	the post to reply. A user can also change who can reply to their posts, or turn off replies, after the post has been posted. • **Excluding harmful content from recommender systems:** Recommender systems are designed to exclude harmful and violating content by integrating with Visibility Filtering systems (e.g. FoSNR) and others. It uses content health prediction models to prevent harmful and violative content ranking higher. • **Law enforcement reporting channel:** X has a dedicated communication channel for law enforcement requests	
Inherent risk score: 15, High Risk Probability: 5, Almost Certain Severity: 3, Moderate Severity *Scope: 3; Scale: 4; Remediability: 2*.	Control score: 3, Defined	Residual risk score: 12, Medium Risk

Risk assessment: Sale of illegal products and services

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, disseminates content related to the sale of products or services prohibited in the EU or used for criminal offences in the EU.

Inherent risk	Controls	Residual risk
• There is a risk that, if a piece of content is not explicitly violating, it may be **unintentionally amplified**, facilitating the sale of illegal products or services. • There is a risk that **high follower counts and engagement metrics** could lend an air of legitimacy to accounts promoting illegal products or services. • There is a risk that **inauthentic accounts promote specific hashtags or trends** related to illegal products and services. This could cause these topics to trend and gain visibility, potentially leading users to engage with such content unintentionally. • There is a risk that **inauthentic and fake accounts are used to target specific users** who might be interested in or vulnerable to illegal products and services. These accounts could initiate **interactions, Direct Messages, and mentions** that lead	• **Law enforcement reporting channels:** We have defined processes for law enforcement and government representatives to report content that is found to violate local laws. • **Temporary enforcement guidelines for high-risk scenarios:** In heightened risk situations, we may design temporary enforcement guidelines. • **Content moderation and enforcement:** In the first half of 2023, [redacted] pieces of content related to [redacted] users were removed for violating our drugs policy. Over [redacted] users were also suspended for activity potentially related to illegal goods and services. In addition, almost 1K pieces of content were actioned for violations related to endangered species across [redacted] different users. Lastly, [redacted] pieces of content were actioned across [redacted] accounts related to the sale of illegal weapons. Reactively, X globally actions on reports of content	• There remains a residual risk that bad actors may constantly shift and **change their behaviour to escape enforcement.**

users to harmful content. ● There is a risk that **inauthentic accounts could be used to promote various scams**, including related to illegal products and activities. These accounts could impersonate legitimate entities to gain trust and deceive users.	related to other illegal products upon internal escalation. ● **Updated guidelines:** Our references for drugs and sexual services enforcements were updated in March 2023. For sexual services related content, the error rate in actioning went down from [redacted] ● **External engagements:** Trusted partners have access to global government affairs teams and can pass resources and updated information on signals, trends. For example, reports are received about illegal money games and casinos from France's ANJ (the National Gambling Authority).	
Inherent risk score: 20. Critical Risk Probability: 5, Almost Certain Severity: 4, High Severity *Scope: 5; Scale: 4; Remediability: 3*.	Control score: 2. Managed	Residual risk score: 10. Medium Risk

Risk assessment: Intellectual property & copyright

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, disseminates content infringing EU IP rights through the service in the EU.

Inherent risk	Controls	Residual risk
• There is a risk that users **post and repost** content, media, links, images that could potentially violate someone's copyright. • There is a risk that **additional accounts** may be created by a user to disseminate copyrighted content. Other users may also repost or quote the content. • There is a risk that X may receive **inadvertent takedown reports** from rights holders, their subsidiaries, and their vendors to take down their own content. • There is a risk that bad actors may attempt to use automation to **abuse the notice-and-takedown process** with a view to getting targeted accounts suspended.	• **Diligent enforcement:** We ensure diligent and consistent enforcement of Copyright and Trademark Policies that apply to content on the platform. • **Expert consultations:** X has copyright and trademark policy experts responsible for identifying abusers and making recommendations regarding trends of reports. • **Notice-and-takedown process:** X has a notice-and-takedown process for copyright issues that is actively enforced.	• There remains a risk that bad actors may come up with **innovating means to surpass our systems**.
Inherent risk score: 10, Medium Risk Probability: 5, Almost Certain Severity: 2, Low Severity *Scope: 2; Scale: 2; Remediability: 2*.	Control score: 2, Managed	Residual risk score: 6, Low Risk

B. Exercise of fundamental rights

We believe X users have the right to express their opinions and ideas without fear of censorship. We also believe it is our shared responsibility to keep users on our platform safe from content violating our Rules. This means that in crafting and enforcing our policies we strive to avoid unnecessary restrictions to our services and carefully balance both addressing illegal content and conduct and protecting fundamental rights, in particular freedom of expression.

For this systemic risk, the inherent risk score across the subcategories ranges from Medium (e.g. Consumer protection) to High (e.g. Freedom of expression), offset by a control strength range of Defined to Managed. As a result, the residual risk for this area ranges from Low to Medium.

Inherent risks

The right to **freedom of expression and of information** is interdependent with all fundamental rights. It is underpinned by bodily security, protection from non-discrimination, equality, human dignity and privacy. The most obvious risk to this right is through restriction of content posted by users on X, specifically the cases where content or accounts were considered incorrectly as violating our policies and terms of service.

There are additional risks to the right to access to information in the context of **media freedom and pluralism**. The inability to access information from transparent and pluralistic sources could occur through different means, including the spread of false information and the inability to discern legitimate sources, for example. In addition, our subscription services, including X Premium, provide more features and filter options to users compared to those who are not subscribed. These products may impact the experience of X Premium and non-X Premium accounts. Although research on this is inconclusive, **personalisation of recommended content** could in some circumstances also contribute to information bubbles, limiting users' access to pluralistic sources of information. While this could happen within our platform, it is also a persisting risk in the wider ecosystem of social media platforms.

There is also an inherent risk that a user's private information is posted on the platform, such as contact information or non-consensual nudity, infringing on their right to respect for their private and family life, as well as their right to the protection of personal data.

Our **direct messaging services** as well as **other engagement features** such as mentions and quote posts could be leveraged for harassment, contributing to a risk to human dignity, non-discrimination, and the respect for private and family life. This may result in users being silenced or exiting the platform.

X as a service is not targeted to children. It was estimated based on January 2023 data that ~2% of EU users were minors, likely as a result of the implementation of mandatory age gating. Still, there are potential negative effects the use of our platform could have on **minors** or in contributing to undermining the **rights of the child**. For example, minor exposure to harmful, inappropriate or shocking content can impact the viewer psychologically and contribute to aggressive or problematic behaviour, including self-harm. For minors that are victims of sexual abuse or bullying, for example, there is a risk of revictimisation by the availability of the media depicting the abuse. In addition, anonymity is allowed on X as an inherent part of the right to freedom of expression, but it may also pose risks for minors if bad actors use pseudonymous accounts for the purposes of grooming.

Social media platforms like X can also be used in detriment of **consumers and their rights** to facilitate transactions involving counterfeit goods or disseminate information on illicit services. Additionally, content can be created and posted to facilitate financial scams or other unacceptable business practices. Another avenue of risk for consumers comes from advertising using fraudulent or deceptive business practices such as financial and healthcare scams (e.g. miracle cures).

Controls

We recognise that first and foremost our commitment is to people's safety and fundamental rights. This is why understanding the most risk prone areas is paramount to ensuring we have the right safeguards in place to protect people and continue to enable public conversations.

To ensure that privacy and data protection is embedded throughout the organisation, X conducts both a legal and privacy review on all new projects that involve the collection and/or use of personal data. In the instances where a project is deemed high-risk to the rights and freedoms of individuals, X conducts a Data Protection Impact Assessment (DPIA); in those instances, its completion and sign off from the Global Data Protection Officer (DPO) are requirements prior to launch. 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, and that advertisements are not presented based on profiling using special categories of data. Further, we conduct risk assessments and biannual external audits on our privacy and data protection related control environment.

Our terms of service define what content and behaviour is and isn't allowed on the platform. We also have additional guidelines and processes in place for assessing content that is reported to us by relevant government authorities as potentially illegal. While developing these guidelines and processes, we aim to preserve the space for political speech and speech that is considered

newsworthy and to further public discourse on topics of public interest. If we receive a valid and properly scoped request from an authorised entity or from affected individuals, it may be necessary to withhold access to certain content in a particular country from time to time. Similarly, we have procedures in place to review law enforcement requests for information to ensure they are consistent with internationally recognised standards on human rights, including due process, and the rule of law.

We also have specific protections in place for minors. X is rated as 'suitable' for those at least 17 years of age in app stores. We prohibit content that jeopardises minors' personal safety, including sexually exploitative content, sexual solicitation, sexualisation and physical child abuse, as well as promotion or encouragement of suicide and self-harm. We also have content labels and interstitials to minimise exposure to sensitive content. We also have age-gating mechanisms in place, and dedicated channels for reporting underage users to us.

Due to the potential negative effects that a binary system to content moderation can have on this systemic risk area, we have developed a set of enforcement strategies with the objective of avoiding disproportionate restrictions on freedom of expression. Earlier this year, we implemented the "Freedom of Speech Not Freedom of Reach" (FoSnR) project bringing a new level of transparency to our enforcement actions by displaying which policy content potentially violates to both the author and other users on the platform. Posts with these labels are made less discoverable and are ineligible for monetization or amplification. Since its launch, we have seen how this type of restricted content receives [redacted] less reach or impressions than unrestricted content globally. We have also observed that [redacted] of authors proactively choose to delete the content after they are informed that its reach has been restricted. Between [redacted] of FoSnR labels were appealed globally, with between [redacted] of decisions being overturned.

Given the added complexity that specific regional or linguistic aspects can bring when assessing potential negative impacts on fundamental rights, particularly when it comes to the different Member States, we have translation resources that our specialist teams can leverage to make better informed decisions.

We produce transparency re[redacted]e range of metrics, but we focus primarily on two types of data. The first type is data related to the actions we take on violating content and accounts. The second type is related to the various legal requests that we receive from governments and different law enforcement agencies. We do this so that our stakeholders can understand how our commitment to their safety has evolved over time and to shine a light on the areas where different governmental agencies may be infringing on fundamental rights such as freedom of expression.

At X, and aligned to the DSA, we value diligent, objective, proportionate and reasonable procedures and the right to remedy. We also appreciate that even if someone gets an objectively fair result stemming from a content review decision, it might not be perceived as such unless they have a clear understanding of the underlying processes. To that end, users can appeal decisions regarding suspension of their account or removal or visibility filtering of content. This is part of our policy principles on procedural fairness whereby our users are provided an avenue to appeal decisions made by our content reviewers or proactive tools (including visibility filtering).

We also have an amnesty policy whereby, on occasion, we have granted amnesty to accounts that had been suspended for low severity violations. This effort is in line with the principle of rehabilitation and providing these accounts with new opportunities to engage in public conversations. The reinstatement is part of balancing the safety of our users and their freedom to express themselves, however these accounts are not exempt from any future violations and will be suspended should they engage in violative behaviour that warrants such an enforcement. This approach mirrors DSA's focus on avoiding unnecessary restrictions on the use of the service and in doing so gives particular consideration to the impact on freedom of expression and of information.

Residual risk

Overall, the potential negative effects our platform or its use can have on fundamental rights after applying control measures range from low to medium risk, expected to have a low to moderate scope of harm on a moderate number of people with possible reversibility in most cases and restoration following controls. Special considerations must be made where there is potentially higher severity of impact such as on the protection of rights of the child and safety of minors.

X has well-developed policies grounded on the respect and balance of fundamental rights, as well as robust safety mechanisms implemented across all functionalities and features of the platform designed to ensure policies are enforced consistently, fairly and equitably. There is a certain degree of risk in the balancing act of respecting fundamental rights and keeping people from harm, which is why we continue to work on mechanisms of redress, harm prevention, transparency, proportionality and procedural fairness.

Risk assessment: Freedom of expression & of information

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, impacts the right to freedom of expression, including the freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers, as well as the right of the freedom and pluralism of the media to be respected.

Inherent risk	Controls	Residual risk
• There is a risk that X **policies** place restrictions on the type of content that users may post on the platform. For example, there may be an outsized impact on people's ability to further political messaging in the EU due to the ban on political ads. • There is a risk that **Direct Messaging** may provide an avenue for abuse through different vectors. This can lead to a silencing effect on the platform. • There is a risk of **Communities** impacting free expression by spreading harmful or harassing content more directly to smaller groups and creating a sense of fear and intimidation. • There is a risk that the use or misuse of X may indirectly impact individuals' right to freely express and receive information when users experiencing abuse or harassment on the platform	• **Freedom of Speech not Reach project:** where we adopted a new approach of restricting the reach of Posts (visibility filtering), to move beyond the binary approach to content moderation. • **Transparency:** Being transparent about our rules and processes allows us to ensure that our mitigation measures are effective and accountable. • **Procedural fairness:** We prioritise fairness and impartiality throughout our moderation processes whereby users can contest enforcement decisions via appeals mechanisms. This includes the amnesty policy whereby we may reinstate accounts suspended due to low severity violations. • **Quality controls and process reviews:** We conduct regular reviews of our policies and processes to	• There remains a residual risk that our enforcement processes **over-enforce for removal of content** that may not violate our rules. • There remains a residual risk that despite our controls, **certain users may still feel unsafe and unwelcome** on our platform due to attempts of abuse and harassment by other users.

do not feel safe to express themselves in certain ways (aka self-censorship), or even exit the platform altogether.	ensure necessary and proportionate outcomes are built into our content moderation systems and processes, including necessary interventions and deviations in crisis situations. • **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. Our measurements have shown that Community Notes significantly reduces sharing of potentially misleading posts. • **Default privacy settings**: All new EU users signing up to the service for the first time have, by default, personalisation turned off for adverts, inferred identity, and places you've been. • **Open-sourcing our algorithm**: We recently open-sourced our recommendation algorithm, to build more transparency and trust.	
Inherent risk score: 15, High Risk Probability: 5, Almost Certain Severity: 3, Moderate Severity *Scope: 3; Scale: 3; Remediability: 3*.	Control score: 3, Defined	Residual risk score: 12, Medium Risk

Risk assessment: Rights of the child and protection of minors

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems or the use made of X services, has a foreseeable or actual negative impact on the rights of the child and protection of minors.

Inherent risk	Controls	Residual risk
• There is a risk that minors may lie about their age as X's **age assurance process** relies on self-declaration to collect the user's date of birth. • There is a risk that content shared on the platform may violate the **image privacy rights** of minors. • There is a risk that **minors may be targeted** with abuse, and other forms of bullying and harassment, which can have a severe impact on their mental health and well-being. • There is a risk that minor **exposure to potentially inappropriate or shocking content** could impact them negatively, and exposure to self-harm content may lead to significant harm, up to and including threats to life. • There is a risk that our platform allowing **anonymity and pseudonymous accounts** could make identifying groomers difficult for potential minor victims. • There is a risk that user **visibility of engagement metrics** could lead to	• **Defined target audience:** X as a service is not targeted at younger users (in app stores, X is recommended for 17+). Based on data from January 2023, it is estimated that *~2% of EU* users were minors, and, as a result of mandatory age gates, the proportion of EU users without an age attributed to their account was ~3%. • **Comprehensive abuse policies:** Our abuse policies apply to all our users - irrespective of their age. We ask users to remove the content when we receive a report from the target. Additionally, our new updated enforcement under FoSNR restricts the reach of abusive content when we receive a bystander report or when the remediation is applied through our proactive models. Further, we have a dedicated policy on right to privacy, which allows users to report posts that contain their images that have been taken without their consent. • **Default settings for logged-out users:**	• There remains a residual risk that a **lack of product solutions to decrease excessive usage time** of the platform may lead to negative impact on minors' wellbeing.

unhealthy comparisons and anxiety, impacting minors' mental health. • There is a risk that **scrolling design and long threads** can encourage excessive use and can cause cognitive fatigue, affecting minors' mental well-being.	By default, X sets high privacy, safety and security settings for users who access X without logging into an account to help ensure that the experience is appropriate for all users who fulfil X's minimum age requirement of 13 years of age. Also, for logged out users, known sensitive media is not shown and advertising must be tagged as being "family safe" to be shown. • **Default security settings:** All new EU users signing up to the service for the first time have, by default, personalisation turned off (personalisation of adverts, personalisation based on inferred identity, personalisation based on places you've been). All users also have Direct Messages defaulted to closed. • **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. • **Sensitive Media:** We restrict views and searches of specific forms of	

	sensitive media such as adult content for known minors or viewers who do not include a birth date on their profile with interstitials, under our age restricted content policy. We also obscures sensitive media behind notices and interstationals. • **Restricted recommendations:** X implements eligibility requirements before it recommends content and accounts (e.g. on the "For you" home timeline). Neither the Following tab nor the For You tab permits sensitive content or inappropriate advertising to be surfaced for accounts of known minors. • **Limits to targeted advertisement:** Advertising presented to EU users who are known minors is not based on profiling. Advertisements containing age-inappropriate content will be tagged as "not family safe" and will also be restricted to minors. • **Age inference:** For user accounts without an assigned age, age is inferred to help prevent minors seeing inappropriate ads. • **Suicide and Self-harm policy:** This policy prohibits users from promoting or encouraging suicide or self-harm.	

	When someone searches for terms associated with suicide or self harm, the top search result is a notification encouraging them to reach out for help. This policy does allow space for sharing personal stories and experiences related to self-harm or suicide when its shared without detailed information about specific strategies or methods related to self-harm, as this could inadvertently encourage this behavior.	
Inherent risk score: 12. Medium Risk Probability: 3, Possible Severity: 4, High Severity *Scope: 4; Scale: 3; Remediability: 5*.	Control score: 2. Managed	Residual risk score: 6. Low Risk

Risk assessment: Human dignity, non-discrimination, and other charter rights

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, impacts the inviolable right to human dignity in the EU, the prohibition against discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation, nationality, as well as other charter rights.

Inherent risk	Controls	Residual risk
• There is a risk that the **use of the platform** could expose users to media encouraging self harm, which could indirectly jeopardise a users' right to life**.** • There is a risk that **misuse of the platform to promote hate** may incite hostility, discrimination, or violence. • There is a risk that **content shared** on the platform may encourage, exacerbate, or facilitate discrimination against people. • There is a risk that enforcement may impact or result in **disproportionate action** against members of certain identifiable groups. • There is a risk that **enforcement of X policy** could directly impact individuals' right to express their language and culture.	• **Freedom of Speech Not Reach:** Restricting the reach of posts (visibility filtering), to move beyond the binary approach to content moderation. • **Transparency:** We promote transparency on our rules and processes to ensure that our mitigation measures are effective and accountable. • **Procedural fairness:** We prioritise fairness and impartiality throughout our moderation processes whereby users can contest enforcement decisions via appeals mechanisms, such as our amnesty policy. • **Quality controls and process reviews:** We conduct regular reviews of our policies and processes to ensure necessary and proportionate outcomes are built into our content moderation systems and processes, including necessary interventions and deviations in crisis situations.	• There remains a residual risk from our efforts to **balance other charter rights**, as we always put the risk to the right to life and safety first.

	• **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. • **Default privacy settings**: All new EU users signing up to the service for the first time have, by default, personalisation turned off for adverts, inferred identity, and places you've been. • **Open-sourcing our algorithm**: We recently open-sourced our recommendation algorithm, to build more transparency and trust. • **Transparency reporting:** We publish transparency reports on data related to actions taken on ToS violating content and accounts, and legal requests that we receive from governments and different law enforcement agencies.	
Inherent risk score: 15, High Risk Probability: 5, Almost Certain Severity: 4, High Severity *Scope: 3; Scale: 3; Remediability: 3*.	Control score: 3, Defined	Residual risk score: 12, Medium Risk

Risk assessment: Protection of personal data

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, impacts the right to the protection of personal data, the right that personal data must be processed fairly for specified purposes and on the basis of consent or some other legitimate basis laid down by law, as well as the right of access to data which has been collected concerning an individual, and the right to have it rectified.

Inherent risk	Controls	Residual risk
• There is a risk that user privacy rights are not met due to a failure to maintain **products, tools, and processes** that promote user privacy and enable users to exercise their privacy rights. • There is a risk that privacy considerations are not honoured through **systems and processes** due to failure to identify, design, and implement privacy considerations throughout the product development lifecycle. • There is a risk that **recommender algorithms** for both content and advertisements may profile or process data in a way that is not lawful, fair, or transparent. • There is a risk that **content that harms the user's right to privacy,** such as contact information or non-consensual nudity, is posted on the platform. • There is a risk that **personal data processing** is carried out in a manner	• **System privacy reviews:** X conducts privacy reviews for any new system developed or purchased, or if there are any relevant changes to a system that might pose a material risk. • **Privacy program:** X's privacy program is intended to ensure appropriate consideration of EU privacy laws in relation to the selection and presentation of advertisements, including, but not limited to, ensuring that advertisements are not presented to X users, both adults and minors, based on profiling using special categories of data (as required by Article 26(3) DSA and 28(2) DSA) • All new EU users signing up to the service for the first time have, by default, have personalisation turned off (personalisation of adverts, personalisation based on inferred identity, personalisation based on places you've been).	• There remains a residual risk with respect to personal data protection that X should be cautious of and improve upon, especially due to the **evolving nature of privacy and data protection and technologies that rely on the processing of personal data to operate**. These areas include but are not limited to: data management practices, regulatory management, ensuring lawful processing of personal data in any new systems/processes, and maintaining procedures to respect and uphold user privacy rights.

that does not ensure appropriate security and confidentiality, leading to data loss and/or a data breach. • There is a risk of **personal data being shared unlawfully** and inappropriately with third parties due to inadequacies in X's third party management processes. • There is a risk of personal data not being managed properly due to inadequate **data lifecycle and data management processes.**	• X provides privacy tools that are designed to help users control what others can see about them, including discoverability controls, private posts, permitting users to choose to accept or decline follow requests, photo tagging controls, sharing location when posting. • **Diverse security and privacy controls:** X has incorporated a suite of security and privacy controls to prevent data breaches and leakage of personal information, including, but not limited to, employee security and privacy training and monitoring endpoints for malware infections. • **Private information and media policy:** X has a strict policy against sharing private information. X seeks to proactively guard against infringement of a user's privacy by not allowing users to share, without the permission of the person to whom it belongs, home address or physical location information, identity documents, including government-issued IDs and social security, financial account information, including bank account and credit card details, etc. of another person. If users violate our policy, we	

	temporarily lock them out of their account and require them to remove this content before they can post again. ● **Non-consensual nudity**: Sharing explicit sexual images or videos of someone online without their consent is a severe violation of X's rules. X will immediately and permanently suspend any account that we identify as the original poster of intimate media that was created or shared without consent.	
Inherent risk score: 16, High Risk Probability: 4, Likely Severity: 4, High Severity *Scope: 4; Scale: 3; Remediability: 3*.	Control score: 2, Managed	Residual risk score: 8, Low Risk

Risk assessment: Respect for private and family life

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, impacts the right to respect for private and family life, home and communications in the EU.

Inherent risk	Controls	Residual risk
• There is a risk that exposure of **private content** could impact an individual's physical safety, emotional well-being, psychological health, and financial security. • There is a risk that users might misuse the **Tipping feature** to share nonconsensual content or coordinate the sharing of such content as consensual media, using it as a means to distribute explicit material without consent. • There is a risk that our systems may not identify instances where users misuse the **Subscription feature** to share nonconsensually explicit content posed as consensual media. This could involve users sharing or coordinating the distribution of explicit material without proper consent. • There is a risk that **spam accounts** potentially disseminate private media or personal information, while also contributing to the distribution of deepfake pornography or the	• **Diligent enforcement:** Throughout our ongoing efforts, we have taken actions against accounts and removed content that shared private information. For example, in a six month period in H1 2023, these actions totaled [redacted] accounts and [redacted] pieces of content, as well as against those sharing private intimate images, resulting in [redacted] suspended accounts and [redacted] removed pieces of content. • **Sensitive content notices:** X has introduced sensitive media interstitials over the content to give notice to other users that it contains sensitive content. • **Visibility filtering and rate limiting:** Visibility Filtering for content and recommendations, for example through downranking. • **Denylisting:** Denylisting refers to removing keywords, posts, third party links or an account from appearing on a product surface. We can denylist by adding the account or post to a specific list that is cross-referenced by the product to then drop it/filter it from	• There remains a residual risk related to the **complexities and nuances within this policy domain.** • There remains a residual risk related to our level of automation and proactive detection, as there are **technological and linguistic challenges to enforcement** in this area.

continuation of harassment campaigns. ● There is a risk that users abuse the **reporting process,** leading to the removal of consensually shared content for non-consensual content (NCN).	appearing. ● **Proactive enforcement mechanisms:** [redacted]	
Inherent risk score: 16, High Risk Probability: 4, Likely Severity: 4, High Severity *Scope: 5; Scale: 2; Remediability: 3*.	Control score: 3, Defined.	Residual risk score: 12, Medium Risk

Risk assessment: Consumer protection

This section provides a summarised assessment of the risk that X's TOS and/or platform policies would be considered illegal or not enforceable in a certain EU market, as well as the risk that the design or functioning of X's services and its related systems, including algorithmic systems, or the use made of X services, disseminates content through the service in breach of consumer protection laws in the EU.

Inherent risk	Controls	Residual risk
• There is a risk that content that facilitates or promotes counterfeit goods, scams or sexual services can benefit from **amplification**. • There is a risk that **indirect sharing of violative content** using coded references or directions may inhibit enforcement detection. • There is a risk that X could serve as a vector for off-platform abuse, fraud and scams, through **third party links and connections** that are made and shared on X.	• **Comprehensive policies:** We have implemented comprehensive content and revenue policies with robust feedback loops to improve enforcement quality, such as X's Counterfeit and Financial Scams policies. • **Market specific language resources for enforcements:** For language related issues that come up during response to reported content, content moderators have guidelines they can follow to provide answers in line with linguistic and cultural standards. • **Consumer protection features:** X has features that aim to protect users from harm such as authenticity challenges. • **Visibility filtering, 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 access and exposure to that content.	• There remains a residual risk that, given the fact that most actors that engage in this type of behaviour have intent to conduct such illegal or deceitful behaviour for profit, they will likely continue to attempt to gain profit by **diversifying their tactics to evade enforcement.**

	• **Country-withheld content:** If we receive a valid and properly scoped request from an authorised entity and the relevant organisation is not liable for ToS action, the account may be withheld in a specific country. • **Interdepartmental cooperations:** Trust & Safety has established a cooperation with the partnership department (X team that act as consultants for major publishers on the platform) to initiate Trust & Safety tickets when high profile events that will likely include scaled digital counterfeit campaigns are coming up. • **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.	
Inherent risk score: 12, Medium Risk Probability: 4, Likely Severity: 3, Moderate Severity *Scope: 2; Scale: 3; Remediability: 3.*	Control score: 2, Managed	Residual risk score: 6, Low Risk

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

The ability to communicate with anyone globally, and especially with the capabilities of X as a real-time platform, has opened doors to new forms of potential risks to democratic processes and public security. The scope of harm here can range from negative effects to fundamental rights such as freedom of expression and the right to vote, as well as indirectly extend to physical harm, notably in emergencies and crises. These risks exist in a complex environment where X serves as both a deterrent and potential conduit for their manifestation.

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. We may find a correlation where if there is an increase in hate speech or polarised sentiment in a certain region, we could predict instability and violence or vice versa. In response we aim to ensure our platform is safe for our users, limit the misuse and inauthentic manipulation of the platform, and capture the content before it goes viral and ensure we take the appropriate remediations.

For this systemic risk, the inherent risk score across the content subcategories is High, offset by a control strength range of Ad-Hoc to Defined. As a result, the residual risk for this area ranges from Medium to High.

Inherent risks

Given the upcoming elections in EU Member states, and EU Parliamentary elections in 2024, the likelihood of an inherent risk manifesting regarding democratic processes, civic integrity, and electoral processes is almost certain. There is always an inherent risk that bad actors may use and misuse our platform to intentionally spread disinformation or conduct coordinated attacks that target public security. Previous research has also shown that in certain circumstances our **recommender systems** could lead to accounts from specific ideological leanings to be amplified over others. However, while there is a risk of bias in these systems, the research highlighted that there are no clear, singular factors in this effect and that in different circumstances the same algorithm produced different impacts on political content. Of the seven countries assessed, the results were not fully consistent, evidencing the difficulty of determining a definitive causal effect of recommender systems and political bias on social media platforms without considering a wider range of intervening variables.

Keeping in mind the balance of fundamental rights, we also acknowledge that there could be a risk from our **terms and conditions**, as our misinformation policies, violent speech policy, as well as abuse and harassment policy could infringe on user's freedom of expression, and potentially limit civic engagement. **Enforcement** of these policies may also create a risk of alienating users who are suspended from the platform for posting violent or harmful content.

Controls

We have robust policies with dedicated teams to prohibit harmful behaviours, and policy updates are communicated through various channels including the Help Center and the @Safety account on X.

A key policy that addresses this systemic risk is our Synthetic and Manipulated Media Policy (SAMM) that prohibits the sharing of synthetic, manipulated, or out-of-context media that may advance misleading claims that lead to harm, helping to ensure media authenticity and mitigate disinformation. This applies to all our users, but acknowledging that the harm of this type of content is exacerbated by wider reach, X Premium accounts that may, under certain circumstances, have additional visibility on our platform need to conform to a number of requirements such as not changing the profile picture, display name or user name. Further, in February 2023, we developed a new policy that consolidates all types of content that promotes or condones violence, the Violent Speech policy. This policy consolidation ensures that we have covered all major gaps in policy and that the rules are scalable across most scenarios. This includes enforcement criteria for indirect incitement to violence as well as permanent suspensions for most cases at first offence.

A key **feature** to tackle misleading content (including misinformation and, to an extent, disinformation) and empower users to meaningfully engage with content is **Community Notes,** which is available in all EU states. Our research shows that this feature dramatically **reduces virality** - entirely organically - by leveraging community knowledge and allowing people to make their own decisions. We find that most notes score highly on helpfulness across the political spectrum. Leveraging an approach that is widely-used in the social sciences, we measure how helpful people found a given note, based on estimates of their political viewpoint. In countries in which we have run this analysis (for example, EU countries like Spain) we found that the vast majority of notes score highly on helpfulness across the local political spectrum.

Understanding that elections can cause particular spikes in inauthentic behaviour, we have clear **escalation processes** and a **24/7 team**. During the 2022 French elections, [redacted] accounts were suspended for various platform integrity violations through proactive sweeps, and [redacted] accounts were actioned for impersonation of key political figures. We have taken major steps in 2023 to prevent bad actors from creating accounts on our platform. In addition, we've also mitigated

spam, platform manipulation, and disinformation risks by manual and automated actions against posts and accounts that make it onto the platform in the EU, as shown in the table below[5]:

EU action rates by language			
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

At the time of assessment, X **prohibits political ads** in the EU. In cases where some ads bypass our proactive detection, we rely on **user reports and human review.** From 2019-2023, around [redacted] ads a month were labelled as political and [redacted] of them were removed in the EU (since the beginning of 2023 we allow political ads in the US). Should our business model shift with respect to this restriction, it will come within scope of our tiered systemic risk assessment framework aligned to Article 34.

Finally, mitigating crisis situations remains one of the critical areas of work for Trust & Safety. We have a cohesive, consistent process that enables us to make risk-informed decisions, allocate resources and apply timely and appropriate remediation measures. Our end state is to proactively prevent risks from each crisis, thereby increasing our overall crisis preparedness. Examples where we have applied crisis-related enforcements in 2023 include: [redacted]

[5] Date ranges from Jan 1, 2023 through to June 16th, 2023. The number of non-EU languages relate to international spam networks that use EU infrastructure as a basis for their attempts at manipulation.

[redacted]

Residual risk

Overall, the potential negative effects our platform or its use can have on democratic processes, civic integrity, electoral processes, and public security, after applying control measures range from medium to high risk, expected to have a moderate to high risk scope of potential harm on a large number of people with difficulty to remedy and restore the situation prevailing prior to the potential impact, following controls.

Our control measures take into consideration that the identified risks to democratic processes and public security are present in an adversarial space where bad actors can constantly shift tactics and behaviours. There is also a residual risk from generative AI content in text, media, and audio forms as such tools continue to quickly evolve and - like all tools that can be misused by bad actors - require us to constantly look to adapt in response. In preparation for the upcoming EU and EU member state elections, we are expanding our Threat Disruption resources to include dedicated election integrity analysts. This includes software engineering, senior specialists on information operations, a civic integrity/elections team lead, and elections analysts.

We have also proactively met with the Slovakian Government in Bratislava to discuss the landscape ahead of their election, ensuring lines of escalation and communication are open. We intend to hire a dedicated role to engage with election regulators and political parties across the EU to further this engagement as part of our wider risk mitigation work. Our efforts to address this residual risk are detailed further in our mitigation roadmap.

Risk assessment: Actual or foreseeable negative effects on democratic processes, civic discourse and electoral processes
This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, disseminates misleading or deceptive content, including disinformation, through the service in the EU that negatively impacts democratic processes, civic discourse, or electoral processes.

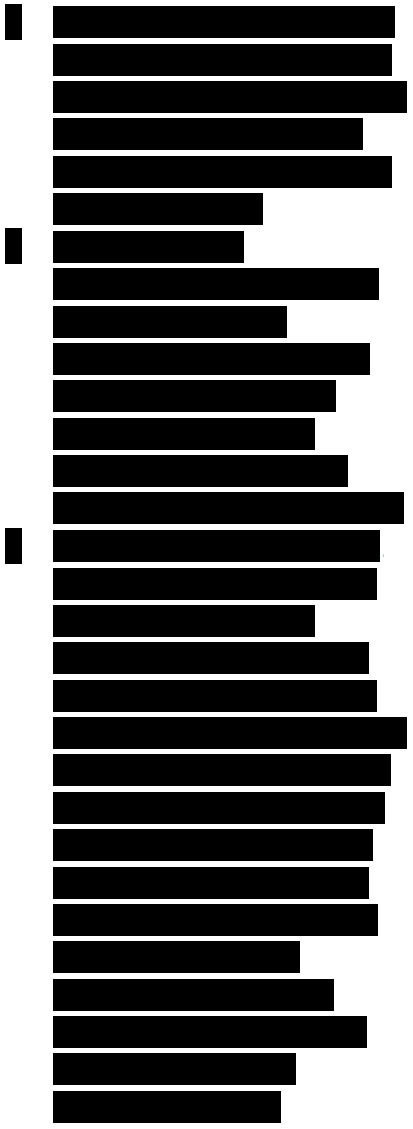
Inherent risk	Controls	Residual risk
• There is a risk that our **content policies**, including our misinformation policies, violent speech policy, and abuse and harassment policy may compromise freedom of speech and consequently limit civic discourse. • There is a risk that X can be **used to share false or misleading information**, including rumours, unverified information or disputed information, about a democratic or electoral process. This could induce citizens to make misinformed decisions throughout their civic participation, limit or reduce civic participation, incite election interference and undermine trust in democratic processes and their results. • There is a risk that **inauthentic use** of the platform could generate manipulative or spammy content. • There is a risk that X's **recommender systems**, including products that bring	• **Comprehensive policies:** Policies are in place that cover content and behaviours that could present risks to democratic processes, including the synthetic and manipulated media policy, civic integrity policy, misleading and deceptive identities policy and platform manipulation and spam policy are applied during electoral processes. • **Exhaustive enforcement:** Processes are in place to remove bad actors on the platform, including disinformation actors and influence operations, at scale, through the enforcement of our platform manipulation and spam policies. In the EU, X takes [redacted] actions a day (based on a six month daily average) under its platform manipulation and spam policy. However, this is only a [redacted] share of the global actions taken. • **Visibility filtering and rate limiting:** These features work to reduce the impact of misleading activity, including	• There remains a residual risk that threats posed by influence operations and disinformation could prevail as **tactics evolve continuously and rapidly.** • As Generative AI tools continue to improve and evolve quickly, a residual risk may manifest from bad actors seeking to leverage such technology to evade existing detections. We will continue to work to understand and detect use of such tools to evade our enforcement. • There remains a heightened residual risk around **elections and the time leading up to it**, where our response could be further enhanced to scale country level and regional efforts swiftly.

amplification benefits like subscription and ads, may inadvertently contribute to the **amplification of misinformation or disinformation.**	malicious URLs, on the platform by reducing impressions and limiting user access to that content. This includes limiting the number of actions an account can take during elections. • **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 contain misinformation or disinformation. • **Profile labels**: Grey checks are granted to government organisations or officials for free, based upon an eligibility criteria, to limit confusion around identities of political figures. • **Ban on political ads in the EU** • We intend to hire a dedicated role to engage with election regulators and political parties across the EU to further this engagement as part of our wider risk mitigation work.	
Inherent risk score: 16, High Risk Probability: 4, Likely Severity: 4, High Severity *Scope: 4; Scale: 4; Remediability: 3.*	Control score: 4, Ad-hoc	Residual risk score: 16, High Risk

Risk assessment: Actual or foreseeable negative effects on public security

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, disseminates misleading or deceptive content, including disinformation, through the service in the EU that negatively impacts public security.

Inherent risk	Controls	Residual risk
• There is a risk of **offline public security incidents** occurring in the EU as often as on a weekly basis, based on recent data of triggers for internal sweeps. • There is a risk of **inauthentic use** of the platform by violent individuals and groups to spread radical ideologies, recruit followers, and even incite violence. • There is a risk that, in the aftermath of highly publicised violent crimes, we may see users **glorifying or condoning violence** or praising the perpetrators. This could promote copy-cat behaviour online and lead to real world harm. • There is a risk that X's **real time nature** could be exploited for coordinating acts of violence. • There is a risk of **coordinated harmful activities** that promotes actions that pose a direct threat to public security or amplifies certain content over	• **Updated policies:** X's Violent Speech policy update derived its recommendations directly from a policy audit derived directly from a policy audit conducted in 2021-2022, subsequently subjected to a secondary revision in 2023 to ensure alignment with the new organisational structure. [redacted]	• There remains a residual risk due to **continuously changing user behaviour, evolving global events, and technological developments**. • As Generative AI tools continue to improve and evolve quickly, a residual risk may manifest from bad actors seeking to leverage such technology to evade existing detections. We will continue to work to understand and detect use of such tools to evade our enforcement.

others. • There is a risk that actors may use **encrypted Direct Messages** to share information to incite offline harm. • There is a risk that X's **policy enforcement** may lead to increased alienation and radicalization of users who are suspended from the platform for posting violent or harmful content. They may alternatively find less popular platforms where there are decreased chances of conversations with opposing views. There is thus a risk of pushing harmful conversations off-platform to more obscure and less regulated spaces. • There is a risk that **data breaches** could have public security implications as compromised user information may be exploited for harmful purposes.		

	[redacted] - **Diligent enforcement:** There is a 24/7 team dedicated to detect and respond to illicit on-platform activities, malicious/harmful user behaviours, and legal/government requests. - **Reporting Mechanisms:** Users can report posts, profiles, lists, spaces, and Direct Messages for containing violent content, harassment, violent speech, violent extremism and misinformation. - **Enforcement teams:** [redacted]	

	• **Community Notes:** Proven to bolster information integrity during public security incidents by helping provide context to users on evolving situations in real time. • **Crisis response protocol:** X's crisis response protocol is based on a tiered approach which assesses harm risk, business risk, 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.	
Inherent risk score: 16, High Risk Probability: 4, Likely Severity: 4, High Severity *Scope: 4; Scale: 4; Remediability: 4.*	Control score: 3, Defined	Residual risk score: 12, Medium Risk

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

The growth of social media usage has raised a strong debate about its impact on public health and the mental and physical well-being of individual users. Insights gained by studies on this subject vary, spanning from a strong link between heavy social media use and increased susceptibility to depression and anxiety, to finding positive effects on users' wellbeing due to platform's ability to foster communities and a sense of belonging.

X's position on the effects of social media on wellbeing remains to be established as its impact will vary depending on pre-existing conditions and personality traits. We acknowledge as with many things, excessive usage is inadvisable and can negatively impact user mental and physical well-being. That said, measurement has suggested that our users exhibit a moderate level of engagement on the platform, rather than excessive usage. Negative interactions and exposure to graphic content can also harm users' psychological state. Misusing the platform to promote dangerous activities or misleading information can be detrimental to public health. The digital gender divide may have also contributed to women and members of the LGBTQ+ community being a target of hate and abuse.

X remains committed to creating a safe and nurturing digital environment for all. We aim to align our pursuit of unregretted user minutes with fostering an environment that promotes positive mental and physical well-being and is free from harm. While X can be misused as a vector for risks, there are notable positive influences on public health, mental and physical well-being as well as rights of vulnerable populations. For example, X serves as an important platform to share valuable information, news updates, and educational content. Users can stay informed on public health incidents, emergency response and other emerging issues related to their safety and security. X also provides the medical community a space to share latest research and facilitate academic discussion. Further, X has been used as a powerful platform to raise awareness on social issues and advocating for change. Cases of domestic abuse, government mishandling and many other forms of hate have been exposed using this platform.

For this systemic risk, the inherent risk score across the content subcategories ranges from Medium to High, offset by a control strength of Defined. As a result, the residual risk for this area ranges from Low to Medium.

Inherent risks

The ramifications at an *individual* level can escalate into systemic public health risks when the impact accumulates across a substantial segment of the population, or when it triggers shifts in social norms and behaviours that attain widespread acceptance (e.g. normalising self-harm).

These risks include: (1) anxiety due to the overwhelming constant stream of information and endless feed of news, opinions, and updates which could be mentally taxing; (2) depressive symptoms due to comparing one's life to the curated highlights shared by others and feelings of inadequacy or low self-esteem; (3) bullying and harassment where such negative interactions could lead to emotional distress and mental health issues; (4) exposure to disturbing images, videos, or discussions could trigger trauma, promote self-harm or exacerbate pre-existing mental health conditions; and (5) excessive usage of X and other platforms in a manner that could negatively affect physical health or attention span.

At a more macro, societal level, bad actors may misuse our platform to facilitate the creation, spread, and amplification of content that can be harmful to public health, including false claims, and divisive narratives that can diminish trust in institutions and proactive measures activated during a health crisis.

Controls

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 content rules and revenue policies govern what can be shared and advertised/promoted on the platform, prohibiting illegal content, and limiting content that could potentially be harmful.

Due to the quick pace at which trends emerge and evolve on the platform, X continuously reviews these policies. A policy audit, originally carried out during the period of 2021-2022, underwent a thorough evaluation in 2023 to ensure its alignment with the updated organisational structure. Recommendations were then put into action through the introduction of new policies or improvements, consolidation efforts, and streamlining of enforcement workflows for increased efficiency and accuracy. Notably, our updated Violent Speech policy accounts for gendered violence, and any form of advocating, glorifying, or threatening sexual violence results in account suspension.

X has also constructed a suite of features to mitigate against potential harms that may manifest on the platform. Community Notes is also an effective measure to tackle health-related misinformation and disinformation. X also applies sensitive media labels for graphic, adult and hateful video and image content, including NSFW labels. Users have the capability to add such labels before uploading the media to ensure automatic labelling at the moment of creation. On top of these product features, we also proactively detect content in this area using heuristic-based rules, which is then manually reviewed.

Furthermore, recognizing that harms contained within this module can be magnified during particular crisis events, X has a robust crisis response protocol that focuses on mitigating harmful effects and protecting the safety of X users, the public, and vulnerable populations. For example,

during the civic disturbances in France in June-July 2023, we activated safety sweeps to swiftly action violating accounts and posts for direct and indirect incitement of violence as well as glorification of violence. This was supported by proactive engagement with French law enforcement and relevant Ministries.

Despite the controls we have in place, we understand that our users may not be aware of all our tools and policies and how to use them. For that reason, we have made sure that all relevant information is available and easily accessible on our help centre pages and that reporting mechanisms are available and intuitive across the platform. We have also developed further features around a range of topics geared towards protecting our users' safety on the platform, for example, our Youth Activist checklist that details all the important and critical points that our users need to keep in mind when it comes to digital safety and protection.

As a result of our controls, from January through June 2023, we conducted the following enforcement on policy areas related to the Physical and Mental wellbeing risk area: abusive behaviours including harassment and bullying [redacted] account suspensions, [redacted] post removals), sensitive media which includes graphic or violent content [redacted] account suspensions, [redacted] post removals) and content encouraging self-harm [redacted] account suspensions, [redacted] post removals). These led to a total of [redacted] post removals and [redacted] account suspensions.

Residual risk

Overall, the potential negative effects our platform or its use can have in relation to public health, physical and mental well-being, and gender-based violence, following control measures, range from medium to low risk, assuming high scope and scale of harm and with different levels of remediability and possibility of restoration to the state prior to the potential impact. We aim to continue developing and strengthening our controls in order to minimise the risk of potential harm to our users, especially as emerging trends present novel risks on the platform. Our efforts to address this residual risk are detailed in our mitigation roadmap.

Risk assessment: Public health & physical and mental well-being

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, disseminates coordinated disinformation campaigns related to public health protection through the service in the EU as well as stimulates behavioural addictions of users of the service.

Inherent risk	Controls	Residual risk
• There is a risk that X may be **used to spread false or misleading information**, which may lead to real-world harm, such as physical harm or diminished trust in institutions responsible for implementing public health emergency response measures. • There is a risk that **inauthentic accounts and activity** may facilitate the creation, spread, speed of amplification and interaction of harmful content. • There is a risk that **heavy usage of social media** may lead to increased risk for depression, anxiety, social isolation, self-harm, and suicidal thoughts. • There is a risk that X's **policies** may not cover all instances of threat to mental and physical wellbeing.	• **Synthetic and Manipulated Media Policy (SAMM):** This policy prohibits sharing synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm ("misleading media"). X takes action on an average of 1M accounts and posts in the EU daily under its platform manipulation and spam policies. • **Comprehensive safety features:** X has introduced a variety of features that aim to protect users from harm, including, but not limited to, NSFW labels, reporting mechanisms, and sensitive content settings. • **Visibility filtering 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. • **Safety features:** NSFW labels on graphic and adult media and sensitive content settings. • **Community notes:** Proven helpful to people from different points of view, and significantly reduces sharing of potentially misleading posts.	• There remains a risk that **public health crisis events may occur unpredictably,** albeit infrequently. • There remains a residual risk that in rare cases our **controls may miss content that can look similar to permitted adult pornographic content,** but is violating**,** such as revenge porn or non-consensual nudity, which could have an impact on physical and mental well-being.

	• **Country-withheld content:** If we receive a valid and properly scoped request from an authorised entity and the relevant organisation is not liable for Term of Service action, the account may be withheld in a specific country. • **Crisis response:** X's protocol is based on a tiered approach that assesses harm risk, business risk, 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. • **Suicide and Self-harm policy:** X has developed a policy prohibiting users from promoting or encouraging suicide or self-harm. When someone searches for terms associated with suicide or self-harm, the top search result is a notification encouraging them to reach out for help. • **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.	
Inherent risk score: 12, Medium Risk Probability: 3, Possible Severity: 4, High Severity (*Scope: 5; Scale: 4; Remediability: 3)*.	Control score: 3, Defined	Residual risk score: 9, Low risk

Risk assessment: Gender-based violence and illegal pornographic content

This section provides a summarised assessment of the risk that the design or functioning of X services and its related systems, including algorithmic systems, or the use made of X services, negatively effect protections against gender-based violence and sexual harassment, disseminates illegal cyber violence content, including illegal pornographic content prohibited in the EU, and contributes to the risk that victims cannot effectively exercise their rights regarding content representing non-consensual sharing of intimate or manipulated material.

Inherent risk	Controls	Residual risk
• There is a risk that we may miss genuine and legitimate cases of violative content where the abuse is not obvious to bystanders and our internal teams, due to X's **tolerance of pornographic content on the platform**. • There is a risk that **automated content moderation** systems may miss violative NCN content due to the difficulty in translating it into automated moderation logic that can detect consensual distribution. • There is a risk that some ads containing abusive content or incitements to harassment or violence, could bypass our **proactive detection**. If a violative ad prevails on the platform, X relies on user reports and human reviews of random ad samples to further catch and remove the violative ad.	• **Comprehensive policies:** We have a robust set of policies that covers all forms of content that could potentially put our users' safety at risk. Our policies cover areas ranging from: abuse and harassment, hateful conduct, NCN, illegal and regulated goods and services (including sexual services) to media policies relating to graphic and adult content. • **Safety features:** Features such as block/mute, account filters, and controlling replies can protect users from gender-based violence (GBV). • **Visibility filtering and rate limiting:** These features work to reduce the impact of harmful activity on the platform by reducing impressions and limiting the number of actions an account can take. • **High privacy settings by default:** All new EU users signing up to the service for the first time will, by default, have personalisation turned	• There remains a residual risk that in rare cases our **controls may miss content that can look similar to permitted adult pornographic content,** but is violative**,** such as revenge porn or non-consensual nudity, which could have an impact on physical and mental well-being. • There remains a residual risk due to the **possibility of other forms of actioned videos or imagery showing up on the platform**, contributing to distress for victims. • There remains a residual risk that the **rapid evolution of trends on the platform, as well as external estimations of lower participation by women on X**, could exacerbate instances of GBV and further contribute to the residual risk.

	off (personalisation of adverts, personalisation based on inferred identity, personalisation based on places you've been), including minors. • **Resources:** We understand that our users may not be aware of all our tools and policies, and for that reason we have made sure that all relevant information is available on our help centre pages. Another resource is the Youth Activist checklist. This details all the important and critical points that our users need to keep in mind when it comes to digital safety and online protection.	
Inherent risk score: 16, High Risk Probability: 4, Likely Severity: 4, High Severity *Scope: 4; Scale: 4; Remediability: 4.*	Control score: 3, Defined	Residual risk score: 12, Medium Risk

VII. Mitigation roadmap

In line with Article 35, this section includes reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to Article 34, with particular consideration to the impacts of such measures on fundamental rights and the nature of the information services provided via the platform.

Given that the inaugural DSA risk assessment pointed to areas of improvement which we had already identified as part of our operations, some measures which serve to mitigate the specific residual risks have already been commenced prior to the end date of the assessment. In some other cases our compliance efforts with the wider DSA obligations contribute to the mitigation of the residual risks identified in this risk assessment.

A. Mitigation measures to address horizontal risks

The mitigation measures detailed below contribute, at a platform wide level, to addressing systemic risks set out in Article 34 of the DSA.

Identified horizontal improvement areas	**Article 35 mitigation measures (pursuant to risk assessment)**
Reinforcing internal processes (Art35(1)(f)): **operational overhaul**	[redacted]

	[redacted]
Adapting design, features, or functioning of services (Art35(1)(a)): **Content moderation functionalities**	[redacted]

Adapting design, features, or functioning of services (Art35(1)(a)): **Community notes**	Community Notes continues to be an agile and dynamic response to the residual risks of misinformation, designed to empower users to participate in the risk mitigation process. This feature exemplifies the transition towards an enhanced community-based content moderation model that relies on user participation rather than solely centralised enforcement. X has Community Notes contributors in all EU member states, and supports all languages in which X is available (including Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slavik, Slovenian, Spanish, and Swedish). We recently launched Notes on Media, which allows a Community Note to appear on all posts that contain a matching image. This is being expanded to videos. These approaches allow notes to scale automatically to multiple posts, with some individual notes already being shown on thousands of distinct posts and growing. We will continue to monitor the effectiveness of Community Notes as the program grows and as events of civic importance happen across the European Union. Our goal is to make the implementation and improvement of Community Notes an open and transparent process. We've published a research paper on Community Notes that provides more detail on how we've been measuring efficacy. All Community Notes contributions are publicly available on the Community Notes site Download Data page so that anyone has free access to analyse the data, identify problems, and identify product enhancement opportunities. Finally, we've made the Community Notes algorithm open source and publicly available on GitHub, along with the data that powers it so anyone can assess, analyse or recommend improvements.
Adapting design, features, or functioning of services (Art35(1)(a)): **Reporting mechanisms**	We are updating our reporting mechanisms to enhance user experience and make them more intuitive for all users. These updates add in-app entry points that bring customers directly to reporting form(s) with important information pre-filled to ensure that all content and profiles on X can be easily reported.

	These changes also reduce the number of clicks required to submit a Terms of Service report, further simplifying the platform's reporting processes. Additional work is planned this year to leverage our reporting flow to provide user education about content that is already labelled (E.g. FoSnR content labels) and the associated content controls customers can use to avoid seeing that type of content in the future.
Taking awareness raising measures (Art35(1)(i)): **Statement of reasons**	We are updating our workflows governing communications sent to the reporter of the content and the user whose content we are moderating to include more detailed data points, improving transparency to users on enforcement decisions we take.
Adapting design, features, or functioning of services (Art35(1)(a)): **Appeals systems**	We are also updating our appeals mechanisms to enhance user experience and make them more intuitive for all users. Furthermore, we will continue to expand our appeal features to enforcement actions such as visibility filtering labelling of sensitive content and FoSnR account labels.
Testing and adapting the algorithmic systems, including recommender systems (Art35(1)(d))**:** **Testing algorithmic systems**	Currently, we do not have conclusive research on whether our proactive models have bias such that it could materially impact a DSA systemic risk, or whether they disproportionately limit speech across different communities. We aim to support further research on bias in recommender systems and content moderation algorithms. This will allow us to train our models better and mitigate against any risk of disproportionate and/or biased enforcement.
Reinforcing internal processes, resources, and supervision of activities as regards detection of systemic	[redacted]

risk (Art35(1)(f)): **data monitoring**	[redacted]
Adapting content moderation processes (Art 35 (1)(c)): **Feedback loops**	User reports provide an invaluable insight into violative content that our product features and automated enforcement systems failed to prevent. We will continue to enhance the feedback mechanisms; for example, escalations to the Strategic Response Team will be analysed for potential efficiencies in product, policy, and operations.
Adapting content moderation processes (Art 35 (1)(c)): **Escalations**	Frontline content moderators should escalate content that is reported but difficult to action, either because the content falls in a grey area of policy (i.e. letter of the policy is not clear enough for agents to enforce), or because the content falls out of scope of our policies. Our in-house escalations team applies additional context, investigation, and stakeholder engagement to implement the most optimal enforcement aligned with terms of service. [redacted]
Adapting advertising systems (Art35(1)(e)) and the online interface to give recipients more information (Art35(1)(i): **Insights from the new Ads Transparency Center**	We have set up an Ads Transparency Center that will show information on every ad served in the EU, including who paid for it, who benefited from it, a depiction of the ad including any media presented, and targeting criteria. We expect public research to follow from analysis of the transparency centre by researchers and the civil society, which we aim to use to inform our next risk assessment cycle.

B. Mitigation measures to address specific systemic risks

In line with Article 35, this section includes reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to Article 34. The measures outlined in Section A above are not included in the systemic risk categories below to reduce repetition, however, the horizontal measures contribute on a cross-cutting basis to mitigate against many of the identified systemic risks listed below.

Dissemination of illegal content	
Identified areas to address	**Residual risk treatment and mitigation measures**
Adapting terms and conditions and enforcement (Art 35 (1)(b)): **Targeted policy improvements**	For the scenarios in which illegal or regulated behaviours are actioned globally through ToS, routine re-evaluation of the effectiveness of operational and policy needs will occur organically based on user ops awareness to new and developing issues. In this area, we will specifically look to review our policies related to terrorist content, sale of illegal products and services, and illegal sexual services and counterfeit, and make any updates deemed necessary to ensure coverage of these risks. More broadly, all ToS policies will continue to undergo regular re-evaluations in response to any changing tactics on the platform.
Adapting content moderation processesses (Art 35 (1)(c)): **Enhancing proactive detection**	Recently, we have increased the number of terrorist and hateful entities that we monitor and remove from the platform. [redacted] We intend to continue increasing proactive detection and enforcement for violative entities via enhanced and/or new models and heuristic-based rules.
Adapting terms and conditions and enforcement (Art 35 (1)(b)): **Measures and**	We have adopted and will continue to build upon our Misuse of Reporting Features policy for abuse of the reporting intake channels and targeting of other users. This will include restricting reports and appeals from users that frequently submit unfounded DSA notices from being reviewed.

protections against misuse	
Enhancing cooperation (Art 35 (1)(g)): **Trusted flaggers**	Trusted flaggers are and will continue to be an important part of our wider response to ensure our users have a safe experience on X. We are expanding our law enforcement portal (LEGOS) and existing mechanisms under copyright and trademark to incorporate Trusted Flaggers reporting. Cases submitted by trusted flaggers will be prioritised once the published list of trusted flaggers is available.
Enhancing cooperation (Art 35 (1)(h)): **Stakeholder engagement**	Given the ever-evolving nature of the space, and cross-industry ecosystem (for e.g. file sharing services, merchants and payment providers may also be involved in the production, discovery, advertisement and distribution of violating content), we will continue to expand our engagement with such industries to improve detection and enforcement as part of the following engagements: ● *CSAM*: Continued partnerships and efforts to invest in proactive detection of CSAM in the EU markets and globally. X is also a member of 'Point de Contact', the French Safer Internet Center's hotline with expertise fighting CSE, online hate and harassment. X also has a partnership with French NGO e-Enfance's child protection hotline. We will also continue our partnerships with non-EU entities to further scale its efforts in this area. ● *Hate Speech*: X has been part of the EU Code of Conduct on Illegal Hate Speech since its creation (2016). X is currently actively engaged in the revision of the CoC to align it with the DSA and invested in its effectiveness. ● *Terrorist content*: X will continue to be an active participant of the EU Internet Forum on fighting terrorism and radical speech online as well as the GIFCT forum and Christchurch Call for Action, contributing relevant material and information on these subjects (shared hashed databases). ● X will continue to take part in the INACH Conference every year, gathering CSOs with expertise fighting hate speech to share experiences and best practices.
Reinforcing supervision of any activities, in particular as regards detection of	We will continue to diligently carry out comprehensive assessments to identify new Violent and Hateful entities, utilising our well-defined objective policy criteria. This systematic approach involves in-depth

systemic risk (Art 35 (1)(f)): **Assessing new entities**	research and analysis of new groups, evaluating whether they fulfil our policy threshold for designation and subsequent removal or visibility filtering. This rigorous process ensures that our decisions are grounded in a thorough understanding of each entity's behaviour and are consistent.

Fundamental rights	
Identified areas to address	**Residual risk treatment and mitigation measures**
Adapting terms and conditions and enforcement (Art 35 (1)(b)): **Targeted policy improvements**	To ensure we have coverage of the risks in this area, we will specifically look to review our policies related to posting private information, right to privacy, non-consensual nudity, harassment, counterfeit, and revenue policies. We will make any updates deemed necessary to ensure coverage of these risks. More broadly, all ToS policies will continue to undergo regular re-evaluations in response to any changing tactics on the platform.
Adapting content moderation processesses (Art 35 (1)(c)): **Enhancing proactive detection**	• [redacted]
Reinforcing internal processes (Art 35 (1)(f)): **Privacy program and risk management**	• Continue to enhance the privacy program based on the annual program and risk assessment that serves as a baseline for strategic plan to discuss and implement measures aimed to reduce identified risks to an acceptable level. • Continue to enhance the current policy management, privacy reporting, and privacy training practices. • Continue to enhance processes to identify, document, and treat privacy risks on an ongoing basis.

	• [redacted]

Testing and adapting the algorithmic systems, including recommender systems (Art35(1)(d)): **Privacy reviews**	• X will continue to conduct privacy reviews to ensure recommender systems remain compliant with personal data requirements. We conduct privacy reviews for any new system developed or purchased, or if there are any relevant changes to a system that might pose a material risk.
	  ○ Continue to ensure that X's Terms of Service clearly explain the main parameters used in its recommender systems, as well as any options for its users to modify or influence those main parameters.
	  ○ Continue to ensure that X users are provided with at least one option for each of their recommender systems which is not based on profiling .
	• Continue to ensure that advertisements are not presented to X users based on profiling using special categories of data; and ensure that advertisements are not presented to X users who are minors based on profiling).

Taking awareness-raising measures (Art 35 (1)(i)): **Enhanced user education**	Develop and launch targeted educational campaigns to raise awareness about X's content moderation policies, mental well-being resources, and how to report harmful content.

Democratic processes, civic discourse, electoral processes, and public security	
Identified areas to address	**Residual risk treatment and mitigation measures**
Adapting terms and conditions and enforcement (Art 35 (1)(b)): **Targeted policy improvements**	We launched our Civic Integrity policy in August that aims to improve the effectiveness of our election integrity efforts, and rebalance our remediations to ensure we are protecting the fundamental right to freedom of expression. • This updated policy aligns with X's new approach to combating misinformation. It entails a shift

	towards policy-based content moderation for only high-severity cases, while leveraging Community Notes to provide helpful context on potentially misleading posts. ● The updated policy will tackle the most severe harms related to civic integrity—mainly voter intimidation and suppression—and will leverage FOSNR labels as the remediation. The policy will no longer evaluate the truthfulness of disputed election-related claims in order to empower users to express their opinions and openly debate during elections in line with our commitment to protect the fundamental right of free speech.
Reinforcing internal processes and resources (Art 35 (1)(f)): **Resource expansion**	● We will be expanding our resources related to civic integrity, including dedicated election integrity analysts that will focus on elections around the world, including the EU. These analysts will review the risk profile of elections globally, and apply the civic integrity policy to minimise harms around civic events. ● We will be expanding our Global Government Affairs team to ensure we have dedicated capacity to cover issues and partnerships related to EU elections.
Adapting design, features, or functioning of services (Art35(1)(a)): **Verification of accounts**	We will scale the option for X Premium users to verify their accounts through identification with a trusted third-party partner.

Gender-based violence, the protection of public health, and serious negative consequences to the person's physical and mental well-being	
Identified areas to address	**Residual risk treatment and mitigation measures**
Adapting terms and conditions and	While we constantly update our policies and enforcement to address abuse and harassment on the platform, we intend to conduct an additional in-depth review of our policies, enforcement and tools to

enforcement (Art 35 (1)(b)): **Targeted Policy improvements**	further enhance our understanding of GBV risks on the platform (cross-functional exercise).
Reinforcing supervision of any activities, in particular as regards detection of systemic risk (Art 35 (1)(f)): **User well-being**	We aim to increase targeted educational campaigns to raise awareness about X's content moderation policies, mental well-being resources, and how to report harmful content.
Cooperation (Art35 (1)(h)): **Stakeholder engagement**	We will explore viable partnerships to enhance our approach to mitigating public health and wellbeing risks on the platform, including: ● Engagement with external organisations and experts on better detection of NCN and possible access to known NCN/revenge porn hashes, including the StopNCII coalition. ● Industry collaboration to address cross-platform harmful activity. ● Collaborations with external experts and organisations to provide support and information.
Reinforcing supervision of any activities, in particular as regards detection of systemic risk (Art 35 (1)(f)): **User well-being**	We will further explore and investigate the impact of dogpiling and understand how to proportionally address such cases through effective proactive and reactive enforcement, and expansion of user-facing safety features.

VIII. Annexes

A. Annex I: Risk Matrices

1. Probability Scale

Scale	Very Unlikely	Unlikely	Possible	Likely	Almost Certain
	1	2	3	4	5
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.8: A probability scale for the DSA risk assessment

2. **Severity scale**

Consideration	Very low severity 1	Low severity 2	Moderate severity 3	High severity 4	Very high severity 5
Scope of impact: The extent to which the harm is physical, psychological, informational, economic, and/or societal. **weighed at 50%**	*Very low* Very low harm on the populations impacted by the risk.	*Low* Low gravity of harm, especially physical and/or psychological harm, on the populations impacted by the risk.	*Moderate* Moderate gravity of any harm, especially physical and/or psychological harm, on the population impacted by the risk.	*High* High gravity of any harm, especially physical and/or psychological harm, on the population impacted by the risk.	*Very high* Very high gravity of any harm, especially physical and/or psychological harm, on the population impacted by the risk.
Scale of impact: Number of individuals affected, on users and non-users, referring to both on-platform as well as societal harms weighed at 40%	*Very low* Impact to a negligible number of users	*Low* Impact to minimal/minor number of users	*Moderate* Impact to moderate number of users	*High* Impact to high number of users	*Very high* Impact to 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.9 A severity scale for the DSA matrix

3. The Residual Risks Scale:

5 Critical	Implies a critical risk, expected to have a very high scope of harm on the most number of people, with irreversibility, or a very high difficulty to remedy and restore the situation prevailing prior to the potential impact, despite controls.
4 High	Implies a high risk, expected to have a high scope of harm on a large number of people, with potential irreversibility, or difficulty to remedy and restore the situation prevailing prior to the potential impact, despite controls.
3 Medium	Implies an medium risk, expected to have a moderate scope of harm on a moderate number of people, with possible reversibility or possibility to remedy and restore the situation prevailing prior to the potential impact, despite controls.
2 Low	Implies a low risk, expected to have a low scope of harm on a minimal/low number of people, with likely reversibility or likely way to remedy the risk and restore the situation prevailing prior to the potential impact, despite controls.
1 Negligible	Implies a negligible risk or no foreseeable risk. If there is any foreseeable risk, it has very low impact on a very low number of people, and is reversible or remedied without difficulty.

Fig. 10: Residual risk scale

4. Control Strength Scale

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 do not have standardised processes in places. 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, documented, formalised, and repeatable. Processes are proactive, well characterised and understood across all organisation verticals.
2	**Managed**	Mitigation measures are sufficiently defined, documented and regularly managed. There is a set process for integrating feedback to mitigate process deficiencies.
1	**Optimised**	Mitigation measures are comprehensively defined and operating at the highest quality. 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. 11: Control strength scale

B. Annex II: Risk scores

Risk assessment	Probability	Severity	Inherent risk	Control	Residual risk
Terrorist content	Almost certain	Very high severity	Critical risk	Defined	High risk
Democratic processes civic discourse, and electoral processes	Likely	High severity	High risk	Ad-hoc	High risk
Illegal Hate speech	Almost certain	Moderate severity	High risk	Defined	Medium risk
Child Sexual Abuse Content	Almost certain	Very high severity	Critical risk	Defined*	Medium risk
Sale of illegal products & services	Almost certain	High severity	Critical risk	Managed	Medium risk
Freedom of expression and of information	Almost certain	Moderate severity	High risk	Defined	Medium risk
Human dignity, non-discrimination, and other charter rights	Almost certain	Moderate severity	High risk	Defined	Medium risk
Respect for private and family life	Likely	High severity	High risk	Defined	Medium risk
Public security	Likely	High severity	High risk	Defined	Medium risk
Gender-based viol. (incl. cyber viol. / illegal porn)	Likely	High severity	High risk	Defined	Medium risk

Intellectual Property & Copyright	Almost certain	Low severity	Medium risk	Managed	Low risk
Rights of the child and protection of minors	Possible	High severity	Medium risk	Managed	Low risk
Protection of personal data	Likely	High severity	High risk	Managed	Low risk
Consumer protection	Likely	Moderate severity	Medium risk	Managed	Low risk
Protection of public health + negative impact to physical and mental well-being	Possible	High severity	Medium risk	Defined	Low risk

Fig. 12: Overall risk scores