RISK ASSESSMENT REPORT

on the results of the Systemic Risk Assessment
Under the Regulation (EU) 2022/2065 Digital Services Act (DSA)

Created on	**April 2025**
Conducted by	**WebGroup Czech Republic, a.s.**

Contents

1. Executive Summary

This document presents the second iteration of WebGroup Czech Republic a.s. (also referred to as "WGCZ")'s Risk Assessment Report for the XVideos.com platform ("XVideos" or "the platform"), as mandated by Article 33(4) of Regulation (EU) 2022/2065 (Digital Services Act or DSA) (the "Report"). It builds on last year's assessment as presented in the first iteration of this report and benefits from (i) our improved understanding of potential risks to users[1], society, and the company's compliance obligations; and (ii) insights obtained through the requests of information sent by the European Commission (the "Commission").

The Report identifies a broader range of risks, with a focus on: (i) illegal and non-consensual content; (ii) minors' exposure; (iii) effects of adult content on public health, focusing on issues such as addiction, body image concerns, and health misinformation; and (iii) threats to fundamental rights. As previously, by adhering to ISO 31000: Risk Management principles, the assessment underpinning the Report evaluates the platform's inherent vulnerabilities and measures the effectiveness of the platform's current controls in addressing/ reducing the identified risks.

Key Changes from Last Year's Assessment:

- The platform's overall risk coverage has expanded substantially from **38 to 87 scenarios**, representing a commensurate broadening of the focus areas. This increase reflects a deeper examination of a broader range of systemic risks, including mental health concerns, gender-based violence, civic discourse manipulation, privacy intricacies surrounding user data, and ethical considerations for commercial content integrity; the deeper scrutiny behind that expansion. (See also Section 4.1: Evolution from the First Risk Assessment);

- To better align with the DSA requirements and the insights received through the Requests for Information received to date, and to provide a clearer risk register structure, WGCZ has **added dedicated categories** such as Gender-Based Violence (GB), Civic Discourse and Electoral Processes (CD), Negative Effects on Public Security (PS), Protection of Public Health (PH), Fundamental Rights (FR), Recommender Systems (RS), User Rights and Platform Policies (UR), and Advertiser and Commercial Content Integrity (AD). These specific categorisations allow for an assessment and management of the relevant risks in a more focused manner. Further reasoning behind the addition of more specific categories of risks is covered in Section 4.1, and Section 4.3: Risk Categories Overview, which provides short definitions and outlines the most representative/common types of risk falling under each new category.

- Each newly categorized risk scenario is accompanied by a **more granular description**. Therefore, any mitigation measures WGCZ implements corresponds more precisely to the nature of the identified risk. This level of detail makes clear why specific measures are chosen and how they specifically address the identified threat. The shift towards more granular risk descriptions and the inclusion of direct references to relevant legislation, research, and technical constraints are further discussed in Section 4.2: Overall Approach, as well as throughout Section 4.3: Risk Categories Overview. These sections detail how each risk scenario now includes precise justifications for the measures it triggers.

- WGCZ has now adopted a **centralised "Mitigation Measures Register",** which is introduced and explained in Section 5.3: Mitigation Measures, and unifies each control (such as AI scanning, manual moderation, or user reporting systems) under one cohesive framework, detailing where and how it applies. By doing so, WGCZ eliminates duplications, streamlines updates, and makes it easier to measure the real-world effectiveness of each control. This also facilitates ongoing reviews and regulatory reporting, providing a single source for understanding precisely how risks are being addressed.

[1] For the purposes of this report, the term "user" refers to the concept of "recipient of the service" as defined in Article 3(b) of the DSA. Under this provision, a recipient is any natural or legal person who uses an intermediary service, whether to seek information (such as viewers) or to make information accessible (such as content uploaders). This usage is intended solely for the purposes of clarity and consistency within the context of this report.

Current Findings and Risk Profile

At the nascent ("inherent") stage of this assessment cycle — that is, before any existing controls/mitigation measures on XVideos are taken into account (see Section 5.2), most risk scenarios fell in the medium band of the five-by-five matrix, and about 10 % were classed High because of their potential for severe harm. After those scenarios were re-scored in light of the controls already deployed and their qualitative effectiveness ratings (Table 8), no scenario remained at High residual risk. A limited number persist at medium residual, reflecting challenges that cannot be eliminated solely through technical or policy measures. In this report, "inherent" and "initial" both describe the untreated risk level on the platform, whereas "residual" denotes the level that remains once current mitigations are considered; detailed definitions are provided in Section 5.4.

Illegal content distributions, including CSAM or non-consensual sexual content, represent some of the most severe inherent threats. However, combining algorithmic scanning, thorough human moderation, compliance with applicable laws and obligations (such as notice-and-takedown procedures, and mandatory reporting duties), and collaboration with NGOs and law enforcement reduces the likelihood of occurrence and their impact. XVideos employs a multi-tier detection, removal, and reporting system to mitigate risks swiftly.

Content Moderation Framework

The platform's content moderation infrastructure stands out as a core risk-control mechanism. Automated AI scanning flags potentially non-compliant content, which is then reviewed by trained human moderators, particularly in cases involving contextual or borderline material. This multi-layered approach reduces false positives (lawful content might otherwise be removed) and false negatives (harmful material overlooked). A transparent complaint-handling process allows verified uploaders and registered account holders to appeal decisions, thus upholding fairness and aligning with fundamental rights requirements.

Recommender System Transparency and Adjustability

XVideos' recommender system enhances user experience but carries an inherent risk of amplifying problematic content if not supervised carefully. To address this, the platform clarifies how location and popularity data influence recommendations while allowing users who opt-in viewing-history personalisation (profiling is off by default) to disable or reset it at any time. These measures help reduce exposure to identified risks such as harmful content loops (repeated algorithm-driven exposure to increasingly problematic content), and subtler forms of algorithmic bias or echo chambers, defined as algorithmically reinforced exposure to narrow viewpoints or repetitive content types. These risks continue to be monitored as part of the ongoing risk management activities detailed in Section 5.4.

Protection of Minors

WGCZ mitigates inadvertent underage access through page-blurring, RTA (Restricted to Adults) labels, prominent age warnings, and parental guidance resources. While these approaches deter casual or accidental exposure, stricter controls (such as parental settings or site-level age-verification solutions) may still be circumvented by using VPNs, or unsecured devices. Although simple click-based age confirmation currently remains the widespread standard across much of the EU, these additional circumvention methods reflect realistic residual risks that arise especially where stricter controls or parental limitations are actively enforced. XVideos continues to evaluate more advanced approaches to age-verification, but recognises that the widespread adoption of biometric or document-based solutions raises significant privacy questions, regulatory uncertainties, and proportionality concerns.

Advertising Integrity

WGCZ has established a clear repository and pre-publication reviews to prevent deceptive or harmful ads from weeding out misleading promotions or illicit product endorsements. All ads are visibly labelled, avoiding user confusion between sponsored material and user-generated content. User controls allow opt-out from certain targeted ads, balancing commercial viability and user privacy.

Governance and Compliance

WGCZ has designated a Compliance Officer to oversee DSA-related obligations, conducts quarterly internal DSA audits, and anticipates external audits to validate processes for content moderation and recommender systems. Partnerships, and on-going engagement to establish cooperation, with NGOs and law enforcement, such as OffLimits and StopItNow,

SWGfL, the wider InHope network, and Interpol, provide additional understanding, expertise, and technologies in CSAM and NCII detection and takedowns.

XVideos has substantially lowered the overall residual risk by implementing a layered system of content moderation, recommender transparency, Terms of Service (ToS), and ongoing collaboration with experts. However, specific issues, particularly child protection, deepfake pornography, or public health concerns, have proven more resilient to straightforward technical solutions. These residual medium-level residual risks reflect systemic challenges that no single platform can fully resolve alone. Offenders intent on child-exploitation content continue to bypass age gates or disguise CSAM, adapting with tactics such as hash obfuscation and VPN chains whenever filters tighten. Deepfake detectors likewise falter when files are heavily compressed or only a performer's face is swapped, meaning final judgments still depend on contextual evidence such as consent records, performer verification, and prompt user reports. Public health risks such as compulsive use, body-image distortion, or self-harm ideation are behavioural, beyond what any algorithm can diagnose, so mitigation hinges on evidence-based nudges and educational messaging. Because these drivers are fast-moving or psychosocial, lasting risk reduction demands an adaptive, multi-stakeholder strategy rather than a single technical control.

WGCZ will continue refining controls, partnering with specialist NGOs and law enforcement entities, and engaging with emerging research. Through this combination of vigilance, collaboration, and iterative improvements, the platform aims to sustain a safe and rights-focused environment for its users while fulfilling its DSA obligations.

2. Introduction

2.1. Purpose and scope of the report

The primary purpose of the risk assessment is to:

- identify and evaluate risks associated with all aspects of our operations that could negatively impact users, society, or our ability to comply with all relevant regulations,
- specifically address risks related to the DSA, including data transparency, reporting obligations, risk of dissemination of illegal content, negative effects for the exercise of fundamental rights, effects in relation to gender-based violence, risks for public health and minors, consequences to the person's physical and mental well-being and advertiser and commercial content integrity.

The scope of this report encompasses all processes, systems, and controls relevant to our online platform's functionality and compliance with regulations. This includes, but is not limited to, content moderation, user safety, data privacy, advertising practices, and reporting procedures. Identified risks are assessed based on their likelihood of occurrence and potential impact on various aspects, such as user trust, public safety and health, protection of minors, regulatory compliance, material and immaterial damages and our overall business objectives.

The conducted risk assessment considered the diverse regional and linguistic landscape of our user base. Despite the complexity due to such a vast landscape, we fully recognize the importance of tailoring our approach to mitigate risks specific to different regions and languages. Our approach to the challenges the linguistic landscape poses in front of us is driven by principles as recited by DSA – those of risk-based approach and proportionality. These principles are reflected not only in our risk management system but are adopted across all related activities.

The report comprises four parts to be read in conjunction with each other:

- the present risk assessment and risk management report,
- the risk management dashboard,
- the Risk Register with risk assessment overview dashboard,
- the Mitigation Measures Register with the mitigation measures overview on the dashboard.

2.2. Risk management framework

The risk assessment was conducted following the principles and guidelines established in the international standard ISO 31000: Risk Management. This standard provides a comprehensive framework for organizations to effectively identify, assess, treat, monitor, and communicate risks. During this assessment, we particularly focused on the standard's emphasis on continuous improvement and the importance of involving top management and other relevant stakeholders throughout the risk assessment process.

Building upon the principles of ISO 31000, we recognize the crucial role of leadership in fostering a strong risk management culture. Our governing body demonstrates a clear commitment to risk management by actively engaging in discussions and providing the necessary resources. This commitment translates into tangible actions, as evidenced by our top management devoting sufficient time and resources to consider measures related to risk management and actively participating in decisions regarding risk mitigation strategies.

The ISO 31000 framework aligns well with the requirements of the Digital Services Act (DSA) for Very Large Online Platforms (VLOPs). The standard's focus on systematic risk identification, assessment, and treatment ensures we proactively address potential issues related to data transparency, reporting obligations, and content integrity, as mandated by the DSA.

We chose ISO 31000 as the foundation for our risk assessment process and risk management system due to several factors:

- ISO 31000 offers a well-established, internationally recognized framework for risk management. It provides a structured approach that encompasses all stages of the risk management process, from identifying risks to evaluating, treating, monitoring, and communicating them;
- while the DSA itself doesn't explicitly mandate the use of ISO 31000, the standard's core principles align well with the objectives outlined in the DSA. These principles emphasize proactive risk identification, risk mitigation strategies, and continuous improvement – all crucial aspects of a robust risk management program under the DSA;
- the flexibility of ISO 31000 allows us to tailor the approach to the specific needs of our platform and the evolving regulatory landscape. We can integrate additional considerations specific to the DSA while leveraging the core principles of the standard.

Hence, utilizing ISO 31000 for our risk assessment and risk management process leverages several advantages:

- provides a well-established and systematic approach to managing risks across the organization,
- encourages a proactive and continuous improvement cycle for identifying, assessing, and mitigating risks,
- facilitates clear communication and collaboration among stakeholders regarding risk identification, evaluation, and treatment strategies.

2.3. Methodology used for risk assessment

Risk Assessment Steps

1. Defining the scope and objectives of the assessment
2. Employing a variety of techniques to systematically identify potential risks relevant to our operations and DSA compliance
3. Assessing the likelihood and potential impact of each identified risk using a pre-defined scoring system
4. Prioritizing risks based on their likelihood and impact scores to determine which require further mitigation efforts
5. Developing and implementing appropriate risk mitigation strategies to address identified risks
6. Continuously monitoring the effectiveness of risk controls and reviewing the risk management process at regular intervals

Data collection

A combination of data collection techniques was used to gather information for this assessment, including:

- internal workshops with relevant stakeholders from various departments (e.g., legal, compliance, IT, content moderation),
- interviews with key personnel to gain in-depth insights into specific processes and potential risks,
- review of internal documents and policies related to data management, content moderation, and regulatory compliance, including the ToS, with specific focus on our content moderation system as a whole and design of our recommender system and algorithms,
- benchmarking against industry best practices and other VLOPs.

Risk identification

To comprehensively identify potential risks, we carried out the following activities:

- brainstorming sessions with cross-functional teams to generate a broad range of risk scenarios,
- scenario planning exercises to consider potential future events and their impact on our operations,
- review of existing Risk Registers and industry threat intelligence to identify relevant risks specific to VLOPs,
- inclusion of systemic risk areas explicitly referenced in Article 34(1) of the DSA,
- consideration of how core platform systems influence the systemic risks, in line with the Article 34(2) of the DSA.

Risk assessment criteria

A scoring system was used to assess the likelihood and impact of each identified risk. The scoring criteria were defined based on pre-determined scales considering factors such as frequency of occurrence, severity of consequences, and potential for business continuity disruption.

3. Overview of the XVideos features

Xvideos.com provides adult-oriented video content and streaming services globally. Its primary function is to host, stream, and share adult video content, allowing users to interact through functionalities such as uploading, viewing, and commenting on videos. Because of the sensitive nature of the hosted content, XVideos implements stringent content moderation practices, conducts regular systemic risk assessments, and maintains transparency mechanisms to manage and mitigate the risks inherent in user-generated adult content.

The platform is intended **exclusively** for an adult user base, specifically, individuals who are legally permitted to access adult content. It is labelled accordingly while no content can be viewed unless the prospective user has confirmed that they are of the legal age (i.e., over 18 years old). XVideos serves as an intermediary within the digital adult entertainment sector by facilitating the uploading and dissemination of user-generated media. The platform also actively monitors the evolution of digital standards within the specialized adult entertainment ecosystem.

3.1. Basic Functionalities

XVideos.com, designated as a VLOP under the DSA, provides core functionalities that classify it as a hosting service and an online platform, as defined in Articles 3(g)(iii) and 3(i) of the DSA. The following functionalities are central to enabling interaction between users and facilitating the dissemination of content:

1. **Content Upload and Sharing**

XVideos allows registered users to upload, host, and publicly share adult-oriented videos. The platform stores and distributes content (in the form of videos) provided by users at their request. This user-submitted content forms the core of XVideos' operational model as a user-generated content platform.

In the main upload flow, newly uploaded content normally passes through a layered moderation flow, combining automated detection systems (such as image and metadata scanning for illegal content) with manual human review by trained moderators. These steps are intended to:

- Verify that all featured individuals are of legal age,
- Prevent non-consensual, abusive, or otherwise illegal material from being published,
- Ensure adherence to intellectual property rights and content integrity.

Submissions that pass the moderation process are made publicly accessible via streaming and integrated into the XVideos search, recommendation, and content discovery systems. Content uploaders can manage their media through user dashboards, where they can edit metadata, monitor video statistics, respond to comments, or remove content. Users also have access to mechanisms to report or appeal moderation decisions.

2. **Content Search**

The platform offers a structured content search infrastructure to support the discovery of relevant adult content based on user preferences. This system includes:

- Keyword-based search allowing users to locate specific content using descriptive terms,
- Tagging system, where content is classified by detailed tags representing categories, acts, performer names, or thematic elements,
- Category and genre browsing, enabling exploration through predefined classifications (e.g., amateur, BDSM, MILF, etc.),
- Channel-specific content, where viewers can access collections by individual uploaders or verified studios,
- Filters and sort options, including trending, most viewed, top rated, newest, or longest videos.

Search results are dynamically ranked based on relevance, keyword match, popularity, and in some cases, user-specific factors such as location or recent viewing history (if the used has actively enabled the tracking of its viewing history).

3. **User Interaction**

XVideos also enables active interaction between registered users and content creators. Every video page on XVideos includes a public comment section where registered users can leave feedback and express opinions. These comments allow viewers to share reactions, ask questions, or highlight moments from the video. Content creators can interact with their audience through replies. Also, the platform provides a simple way for registered users to express content preferences via "like" and "dislike" buttons. These votes are aggregated and displayed on the video page, contributing to overall popularity scores. Registered users can "subscribe" to their favourite content creators.

4. **Recommender Systems**

XVideos integrates a recommender system to help users discover relevant content based on contextual and behavioural signals (when a user has actively enabled viewing history). The platform explicitly discloses how this system functions in its ToS to ensure transparency in alignment with the requirements of the DSA. The platform's users retain control over certain influencing factors, such as location and content category selection.

On the homepage of XVideos, the recommender system prioritises two main criteria: the user's chosen geographical location (users in different regions can be presented with various videos on the homepage based on presumed or manually selected geographic interest) and the popularity of videos within that region (system ranks and displays content based on how often users in that country have clicked it). The number of clicks is used as a proxy for popularity, reflecting what is trending or widely viewed among users in a similar geographical area. Users can actively influence the homepage recommendations by selecting or changing their location in the online interface.

Beyond the homepage, recommendations are refined further when users navigate specific categories or content creator pages. The recommender system shows videos related to the selected sub-genre or category in these contexts and content uploaded by the chosen creator. Users impact the output directly by interacting with these filters. For example, selecting a specific tag like "POV" or browsing a particular model's content will immediately adjust the video suggestions to reflect those interests.

XVideos also offers a "related videos" feature that becomes visible during or after video playback. This part of the recommender system draws on collective viewing history across the **entire user base** (not just the individual user) and the context of the video being viewed, tags, categories, and popularity. To be clear, this feature does not rely on the profiling of any single, individual user.

When a user has actively enabled viewing history, the system can refine related video suggestions based on past interactions. This allows for a more tailored experience, surfacing content that aligns with the user's demonstrated interests. Turning this functionality on or off remains within the user's control, respecting autonomy and privacy preferences.

5. **Content flagging and reporting**

XVideos empowers users and authorities to report content that is illegal, harmful, or violates platform ToS. The reporting system includes abuse reporting forms (for CSAM, non-consensual content, terrorist content, etc.), per-video reporting buttons for user convenience, and authority Contact Point for EU regulators and law enforcement.

Reported content is immediately ghosted (made inaccessible to the users) while a human moderation team reviews the complaint. The process includes possible uploader feedback and always results in a reasoned resolution sent to the complainant (if claimant’s contact information was provided). Also, XVideos supports trusted flaggers from the European Economic Area (EEA), whose reports are prioritized, demonstrating its commitment to trusted collaboration.

3.2. Types of content hosted

XVideos hosts a wide range of adult-oriented digital content strictly governed by its ToS, community guidelines, and applicable legal standards. The platform maintains a zero-tolerance policy toward illegal, harmful, or non-consensual material and implements a multi-layered moderation and content responsibility approach.

The content on XVideos is exclusively uploaded by third-party uploaders, including amateur videos (filmed and uploaded by individuals), self-produced series or content by independent adult performers, including verified creators, images, and videos, often added manually by the uploader. All user-generated content must meet strict compliance requirements. Uploaders are obligated to confirm they are 18 years of age. Also, the platform ensures that all individuals depicted in the content are adults and have given explicit, informed, and documented consent. Violations of these terms result in immediate content removal and can trigger account termination and referral to law enforcement where applicable.

XVideos hosts content from professional adult studios, distributors, and verified commercial partners. These content providers operate under formal licensing agreements or direct publishing partnerships with WGCZ. Such content is curated, tagged, and promoted following platform rules and applicable laws and is subject to the same moderation processes as user-uploaded content.

The platform hosts display and video advertisements from entities within the adult entertainment ecosystem. These can include banner ads promoting premium adult content sites, affiliate links to camming services, dating platforms, or merchandise stores, and sponsored videos or promotions integrated into the platform's interface.

All content hosted on XVideos must comply with strict platform policies and applicable legal obligations. Prohibited content includes, but is not limited to:

- Child sexual abuse material (CSAM), including real or simulated depictions,
- Non-consensual content, including hidden camera recordings or "revenge porn",
- Content involving coercion, violence, or extreme harm,
- Terrorist, extremist, or criminal propaganda,
- Copyright-infringing material, unless uploaded by the rightful owner or licensee,
- Degrading, hateful, or discriminatory expressions that violate EU and national anti-hate speech laws.

XVideos supports responsible content creation and sharing within a strictly regulated environment, guided by a framework of user accountability, proactive moderation, and transparency as mandated by the DSA.

3.3. Role of Management

The management of XVideos plays a central role in the governance, risk mitigation, legal compliance, and continuous development of the platform. Senior leadership is actively engaged in shaping the platform's operational framework, including the approval and periodic update of internal governance policies to ensure alignment with evolving legislative requirements. This involvement extends to the implementation and enforcement of the ToS, the oversight of content control mechanisms, and the advancement of compliance initiatives under the DSA, including the designation of a dedicated compliance officer. Senior management also provides strategic oversight of systemic risk assessments, in accordance with the obligations set forth by the DSA. Key responsibilities include:

- Identifying emerging systemic and operational risks that can affect fundamental rights, public safety, or the integrity of democratic processes;
- Conducting regular internal audits and assessments of algorithmic systems, content moderation workflows, and other safeguards critical to DSA compliance;
- Reviewing and validating mitigation strategies, particularly those addressing the dissemination of illegal content, disinformation, or manipulation of users.

A core area of managerial oversight is content moderation. Management ensures that moderation teams are properly trained, resourced, and operate within clearly defined legal and ethical standards. Furthermore, senior management actively refines moderation practices by integrating feedback from content moderators, certified trusted flaggers, and relevant external stakeholders.

In terms of external engagement, WGCZ (the operator of XVideos) has established a dedicated point of contact for EU authorities and public bodies. This communication channel supports the secure, timely handling of official orders (e.g. from

regulators, or law enforcement), regulatory inquiries, and law enforcement notifications. The management team is responsible for coordinating appropriate and prompt responses to such requests. Additionally, management oversees the publication of the platform's legally mandated transparency reports, ensuring regular disclosure of moderation activity, systemic risks, and compliance measures.

Platform innovation and development also fall under management's strategic direction. In recent years, management has prioritized the integration of user safety and transparency features into the platform's technical infrastructure. This includes enhancing algorithmic accountability by making the content recommendation system more transparent and user controllable. For example, users can adjust their viewing preferences, disable personalized recommendations, or filter content by location, category, or popularity. Management is also leading initiatives to improve algorithmic explainability and to provide users with clear, accessible information about how content is ranked and surfaced.

Beyond technical and regulatory domains, management is committed to cultivating an ethical and responsible platform culture. This includes setting standards for corporate conduct, promoting digital dignity, and ensuring the platform's operations are rooted in the principles of consent, safety, and user empowerment. Internal reporting mechanisms are in place, allowing employees and contractors to report misconduct or policy violations securely and without fear of reprisal.

In summary, the active participation of XVideos' senior management is fundamental to the platform's compliance with the DSA and its overall governance. Their leadership ensures that the platform operates with accountability, prioritizes user protection, and continuously adapts to meet the evolving demands of the digital environment.

.

4. Risk Identification

4.1. Evolution from the First Risk Assessment

During the first risk assessment, WGCZ identified 38 unique risk scenarios relevant to XVideos. However, continuous monitoring of the regulatory landscape, relevant research, best practices, insights obtained through the requests for information received from the Commission, and analysis reports from leading online platforms highlighted the need to expand the scope of the assessment. The latest updates to the Risk Register significantly enhance the structure, clarity, and comprehensiveness of risk management efforts. As a result, WGCZ added 49 new risk scenarios, increasing the total from 38 to 87. This significant expansion allows for a more comprehensive consideration of evolving risks and newly identified threats related to content moderation, privacy, user safety, and DSA compliance. While in a number of cases, existing risk categories were sufficient to incorporate new threats, in others, creating new categories was deemed necessary.

One of the most notable updates is the introduction of several new risk categories, which aim to more precisely capture distinct areas of concern, in particular:

- *Gender-Based Violence* focusing on gender-based abuse, harmful stereotypes, and grooming risks;
- *Civic Discourse and Electoral Processes* consolidates risks related to disinformation, political manipulation, and election-related content integrity;
- *Negative Effects on Public Security* addresses threats from extremist groups, cyber-attacks, and criminal exploitation of the platform;
- *Protection of Public Health* focusing on issues like addiction, body image concerns, and health misinformation;
- *Fundamental Rights* which focusing on privacy and dignity violations, discriminatory practices, and freedom of expression, providing better alignment with regulatory concerns;
- *Recommender Systems and Algorithmic Design* emphasizing risks from algorithmic amplification and content personalization.
- *User Rights and Platform Policies* focusing on transparency in ToS, content moderation policies, and user protections;
- *Advertiser and Commercial Content Integrity* addresses the ethical use of advertising data, ad repository transparency, and user control over ad targeting.

Beyond these structural realignments, numerous existing risks have been reclassified to better fit within these new categories. For example, risk scenarios related to gender-based violence and revenge porn, previously categorized under *Illegal Content Distribution* (IC_6 and IC_7), have now been reassigned to the *Gender-Based Violence* (GB_1 and GB_2) category. Likewise, disinformation and misinformation risks that were previously labelled under *Manipulation of the Platform* (MA_1 and MA_3) have been shifted to *Civic Discourse and Electoral Processes* (CD_1 and CD_2). Risks involving algorithm-driven amplification of harmful content, previously part of *Systemic Risks* (SR_1, SR_2, and SR_3), now fall under *Recommender Systems and Algorithmic Design* (RS_1, RS_2, and RS_3). Similarly, risks related to unauthorized data usage and privacy violations (AF_2) have been moved from the broad *Adverse Effects on Fundamental Rights* category into a more focused *Data Privacy and Protection* category (DP_5).

Alongside the reclassification of existing risks, numerous new risk scenarios have been introduced, reflecting emerging threats and regulatory expectations. These include new risks under *Illegal Content Distribution*, such as the exploitation of slow moderation processes, inadequate content reporting features, and automated moderation mistakenly removing lawful content (IC_8 – IC_11). The *Gender-Based Violence* category now includes risks related to the recommendation system amplifying harmful content, insufficient reporting mechanisms for victims, and grooming activities targeting vulnerable users (GB_3 – GB_6). Similarly, within *Protection of Minors*, new risks highlight concerns around the misuse of minors' personal data, inadvertent exposure to harmful content, and the potential mental health impact of such exposure (PM_7 – PM_9).

Further updates introduce *Data Privacy and Protection* risks, including concerns over unauthorized tracking, weak encryption, and insufficient clarity regarding user data processing (DP_6 – DP_9). Additional risks within *Civic Discourse and Electoral Processes* emphasize the potential for platform algorithms to amplify political extremism, delays in moderating civic engagement-related content, and the monetization of politically harmful content (CD_3 – CD_6). The *Negative Effects on Public Security* category has been expanded to account for criminal and extremist groups exploiting the platform, cybersecurity vulnerabilities, and insufficient responses to cyber-attacks (PS_1 – PS_4).

Similarly, newly identified risks within *Recommender Systems and Algorithmic Design* include malicious actors exploiting algorithms to spread disinformation and harmful content, the risk of users being exposed to increasingly extreme content due to recommendation mechanisms, and inadequate adaptation of recommendation systems to regional and linguistic contexts (RS_4 – RS_7). In *Protection of Public Health*, new risks scenarios now capture concerns about body image-related harm, the dissemination of misleading health information, and the potential promotion of self-harm (PH_3 – PH_6). Additionally, the *Transparency and Reporting Obligations* category has been expanded to include risks related to failure to publish transparency reports on time and inaccurate reporting of active platform users (TR_5 – TR_6).

Beyond these substantive changes, the Risk Register also introduces new risk scenarios related to XVideos policies and advertising transparency. The new *User Rights and Platform Policies* category now outlines risks such as unclear ToS, inadequate complaint handling mechanisms, and failure to publish the ToS in all official EU languages (UR_1 – UR_6). Meanwhile, the *Advertiser and Commercial Content Integrity* category now addresses issues such as the unauthorized use of sensitive personal data for advertising purposes, incomplete ad repository records, and a lack of transparency in advertisement removals (AD_5 – AD_9).

The rationale for these changes is driven by several key factors. First, the need for improved categorization ensures that risks are grouped in a way that more accurately reflects their nature and potential impact, allowing for more precise risk mitigation strategies. Second, the recognition of emerging threats, particularly in the areas of content moderation, algorithmic influence, and privacy violations, has necessitated the addition of new risk scenarios that were previously unaccounted for. These changes align with the evolving regulatory obligations, including transparency and accountability requirements under digital governance frameworks. Lastly, the updates enhance transparency by clearly defining platform risks, improving user rights protections, and ensuring more substantial advertising and algorithmic systems oversight. The latest updates significantly enhance risk management efforts' structure, clarity, and comprehensiveness, making it easier to identify, track, and mitigate platform risks effectively.

4.2. Overall Approach

WGCZ employs a comprehensive approach to risk identification, designed to uncover a broad spectrum of potential systemic and direct risk scenarios. Systemic risks can impact the entire industry or digital ecosystem, while direct risks affect more specific aspects of operations, such as data security or content moderation.

This methodology rests on four main pillars:

- Structured workshops, involving purposefully convened sessions with multidisciplinary teams and external experts.
- Corporate knowledge, leveraging the accumulated know-how of internal teams.
- External collaboration, engaging with specialist consultants, third-party organizations, and thought leaders in cybersecurity and regulatory compliance.
- Continuous legislative and research monitoring, which includes reviewing relevant laws, regulations, and academic/industry research to ensure that newly emerging risks are identified promptly.

Structured workshops

The WGCZ risk identification methodology includes the organisation of assessment workshops, where risk scenarios in relevant areas are defined and discussed. The workshop format serves as an effective method for identifying and further evaluating risk scenarios, as it fosters open communication, allowing participants to share their thoughts and concerns based on their experience and expertise. This approach helps uncover threats and corresponding risk scenarios that might have gone unnoticed.

Workshops were conducted with representatives from key departments, including:

- Leadership
- Legal and Compliance Department
- Content Moderation Team
- IT Department

Despite the formalised process of planning and documentation, the workshops were held in a flexible open-discussion format, which encouraged active participation and brainstorming. While each workshop had unique aspects, a typical structure included discussions on new categories and specific risk scenarios. Participants collectively identified and assessed various likely risk scenarios that had not been detected in the previous iteration.

The risk identification process was based on substantive open discussions and the application of comparative analysis methods with the risk assessment practices of other major platforms. This approach enabled a comprehensive evaluation capable of adapting to evolving regulatory requirements and the operational specifics of the platform.

Although many identified risk scenarios are common to any platform, the risk identification process was specifically focused on fulfilling DSA obligations, particularly assessing systemic risks per Article 34 of the DSA. Special attention was given to the protection of minors, combating NCII, CSAM, and ensuring transparency in content moderation and reporting processes.

Corporate knowledge

WGCZ leverages substantial corporate knowledge drawn from the experience and expertise of its internal teams. The platform's content moderation team, characterised by low turnover and extensive experience, possesses deep insights into user behaviour patterns, moderation challenges, offender profiles, and recurring issues involving illegal content, such as NCII and CSAM. This experienced workforce has provided essential expertise, enabling the identification of risks primarily within the category of illegal content distribution, including the uploading of CSAM, non-consensual sexual material, exploitation of delays in moderation processes, inadequacies in moderation tools, and challenges related to regional language adaptation. In addition, the team highlighted risks related to gender-based violence, such as revenge porn, and cyberbullying, and risk scenarios concerning the protection of minors.

The internal IT team's expertise was critical in identifying and addressing technological and cybersecurity-related risks. They identified and addressed risks related to data breaches, inadequate encryption methods, and insufficient anonymization of sensitive user data. Their evaluations also contributed significantly to identifying and addressing vulnerabilities in access control mechanisms, particularly concerning minors and weaknesses in detecting malicious activities such as phishing and malware distribution. Furthermore, their technical insights informed and allowed to address risks related to potential algorithmic biases in recommender systems and inadequate user control over recommendation settings.

Internal legal and compliance experts proactively identifying compliance-related risks. Their expertise primarily informed scenarios within the fundamental rights, including privacy violations, potential discriminatory practices within platform policies, and the balancing of freedom of expression and moderation practices. Additionally, they identified risks associated with the transparency and reporting obligations required by the DSA, specifically related to timely and accurate transparency reporting and clear disclosure of content moderation and appeal processes to users. Their detailed understanding of advertising regulations also contributed to identifying and addressing scenarios within the advertiser and commercial content integrity, including misleading advertisements, improper use of sensitive user data in ad targeting, and inadequate maintenance of advertising repositories.

Legislative and research monitoring

The legislative and research monitoring pillar emphasizes continuous monitoring and proactive engagement with relevant legislation, regulatory frameworks, and academic or industry research. WGCZ systematically tracks and analyses legislative developments concerning NCII, CSAM, online safety, and privacy protection within the European Union and across other major jurisdictions, including the United Kingdom and the United States (at federal and state levels).

In addition, WGCZ conducts targeted analyses of specific legislative requirements that directly impact its operations. A recent example includes the detailed analysis of requirements for implementing age verification tools as stipulated by the Czech Republic's Radio and Television Broadcasting Council (RRTV), interpreting § 7(1)(a) and § 8(3)(f) of Act No. 132/2010 Coll. on On-Demand Audiovisual Media Services, and § 6a(1)(a) of Act No. 242/2022 Coll. on Video-Sharing Platform Services. This analysis explored the adequacy and limitations of currently used methods such as qualified disclaimers and highlighted potential privacy implications and broader consequences of adopting more sophisticated age verification mechanisms. Such detailed regulatory reviews directly inform risk scenarios identified in the *Protection of Minors* category, particularly those concerning inadequate access verification and the importance of balancing adequate age controls with users' rights to privacy.

Moreover, WGCZ's risk identification approach integrates external research, data analyses, and academic studies into its risk identification methodology, ensuring a comprehensive and evidence-based understanding of emerging online safety trends. For instance, research by Thorn (2023)[2] on youth perspectives regarding online safety provided insights into moderation-related risks such as the uploading of violent or gender-based violence material and risks involving revenge porn and cyberbullying. Likewise, findings from Lim et al. (2015)[3], Mestre-Bach et al. (2023)[4], and Wright et al. (2015)[5] informed the risk assessment of violent and sexually aggressive content, highlighting complexities around the impacts of violent pornography and its relationship to real-life violence.

The comprehensive privacy-focused studies by Vallina et al. (2019)[6] and Verizon (2023)[7] provided valuable data regarding unauthorized tracking, data breaches, and inadequate privacy compliance on some pornographic websites. This evidence directly informed identified risks in the Data Privacy and Protection category.

The analysis of research on the impact of pornography on minors and youth, such as Owens et al. (2012)[8] and UNICEF (2021)[9], reinforced the Platform's understanding of the nuanced relationship between exposure to explicit content and minors' well-being. Research exploring associations between pornography use, relationship quality, and psychological outcomes, such as studies by Perry (2017, 2020)[10], Willoughby et al. (2014)[11], and Bőthe et al. (2020)[12], help to identify risks related to public health and user well-being.

External collaboration

The platform's approach to risk identification is supported by ongoing collaboration with a range of NGOs and expert third parties who provide valuable insights into harms affecting users – particularly in relation to CSAM, NCII abuse, and related

[2] Thorn. (2023). LGBTQ+ Youth Perspectives: How LGBTQ+ youth are navigating exploration and risks of sexual exploitation online

[3] Lim, M. S., Carrotte, E. R., & Hellard, M. E. (2015). The impact of pornography on gender-based violence, sexual health and well-being: What do we know? Journal of Epidemiology and Community Health, 70(1), 3–5

[4] Mestre-Bach, G., Villena-Moya, A., & Chiclana-Actis, C. (2023). Pornography use and violence: A systematic review of the last 20 years. Trauma, Violence, & Abuse, 25(2), 1088–1112

[5] Wright, P. J., Tokunaga, R. S., & Kraus, A. (2015). A meta-analysis of pornography consumption and actual acts of sexual aggression in general population studies. Journal of Communication, 66(1), 183–205

[6] Vallina, P., Feal, Á., Gamba, J., Vallina-Rodriguez, N., & Anta, A. F. (2019). Tales from the porn. Proceedings of the Internet Measurement Conference

[7] Verizon. (June 6, 2023). Global number of data breaches with confirmed data loss from November 2021 to October 2022, by target industry and organization size [Graph]. In Statista

[8] Owens, Eric W., et al. "The impact of internet pornography on adolescents: A review of the research." Sexual Addiction & Compulsivity, vol. 19, no. 1–2, Jan. 2012, pp. 99–122

[9] UNICEF, ., 2021. Digital Age Assurance Tools and Children's Rights Online across the Globe, UNICEF: United Nations Children's Fund. United States of America

[10] Perry, Samuel L. "Pornography and relationship quality: Establishing the dominant pattern by examining pornography use and 31 measures of relationship quality in 30 national surveys." Archives of Sexual Behavior, vol. 49, no. 4, 2 Jan. 2020, pp. 1199–1213; Perry, Samuel L. "Pornography use and marital separation: Evidence from two-wave panel data." Archives of Sexual Behavior, vol. 47, no. 6, 21 Sept. 2017, pp. 1869–1880

[11] Willoughby, B. J., Carroll, J. S., Nelson, L. J., & Padilla-Walker, L. M. (2014). Associations between relational sexual behaviour, pornography use, and pornography acceptance among US college students. Culture, Health & Sexuality, 16(9), 1052–1069

[12] Bőthe, B., Tóth-Király, I., Potenza, M. N., Orosz, G., & Demetrovics, Z. (2020). High-frequency pornography use may not always be problematic. The Journal of Sexual Medicine, 17(4), 793–811.

online risks. These partnerships serve as a complementary source of expertise within the platform's broader risk assessment and mitigation framework.

EU Collaboration. WGCZ cooperates with the Czech Safer Internet Centre, bringing significant expertise in addressing issues related to NCII and CSAM. Additionally, the platform continues to work with OffLimits (Netherlands), with implementation of a CSAM hash list and preparatory work to introduce deterrence messaging and helpline services from the Stop It Now programme. These efforts aim to further tackle both the dissemination of CSAM and underlying offender behaviour through early-stage intervention.

UK Collaboration. Dialogue is ongoing with the Lucy Faithfull Foundation concerning the possible adoption of Project Intercept and the reThink Chatbot, a tool designed to redirect users searching for CSAM towards support services. Engagement is also underway with SWGfLto implement the StopNCII tool, which enables adult victims to hash and flag intimate images at risk of being shared online. This supports more effective detection and removal while maintaining user anonymity.

I**nternational Collaboration.** At the international level, WGCZ engages with InHope, the global network of internet hotlines for reporting CSAM, contributing to the development of mitigation measures. Collaboration with Interpol is also in progress, including the potential adoption of the "Worst Of" list (IWOL) – a database of domains publishing the most severe CSAM. The partnership agreement is currently under review. In parallel, engagement with Interpol's child safety data analytics team continues to explore further opportunities for coordinated action.

Engagement in Standards Development. In addition to these partnerships, the platform is also engaged in international efforts related to age assurance standards. Since February 2024, WGCZ has been collaborating with the Digital Governance Standards Institute of Canada in the development of a national age assurance standard. Through its Director of Regulation & Safety, WGCZ has participated as an active member of the Expert Drafting Team since March 2024. While initial involvement focused on shaping the structure of the work, participation has intensified in recent months, with active contributions to the drafting process and detailed discussions on key provisions. The standard is approaching completion and is expected to be opened for wider consultation in the near future. The primary focus of this involvement is to support the development of effective approaches to protecting children online, such as device-level age verification solutions, while also working to ensure that the standard does not promote or enforce, for example, site-level approaches that could negatively impact user privacy or operational feasibility. This contribution demonstrates the platform's commitment to supporting privacy-conscious and technically robust age assurance frameworks at an international level.

Multilateral Forums and Knowledge Exchange. Finally, participation in a UN co-hosted conference in New York and related industry roundtables has enhanced the WGCZ's understanding of emerging risks and international best practices, reinforcing its commitment to proactive and informed risk governance.

4.3. Risk Categories Overview

This section outlines the various categories of risks identified during the assessment. Each category aligns with potential areas of concern regarding our operations and compliance with the DSA.

By comprehensively understanding these risks, we developed effective mitigation strategies to maintain xvideos.com a responsible online platform.

1. Illegal Content Distribution

Illegal Content Distribution refers to the risks associated with the uploading, sharing, or distributing of unlawful, harmful, or unauthorized content. This category encompasses a wide range of violations, including copyright infringement, dissemination of malware, and the spread of CSAM. The presence of such content on a platform can have severe consequences, both legally and reputationally, and requires strong detection, moderation, and enforcement mechanisms.

The risk scenarios highlight the specific challenges within this category. The spread of hate speech (IC_2) demonstrates how platforms must have effective moderation tools capable of detecting and removing harmful rhetoric before it incites violence or discrimination. Exploiting slow moderation processes by malicious content creators (IC_8) illustrates how even robust content policies can fail if enforcement mechanisms are inefficient. Automated moderation systems, while valuable, also introduce risks, as seen in scenarios where lawful content is mistakenly removed (IC_10), leading to undue censorship and the suppression of legitimate speech. These examples underscore the delicate balance that must be maintained between removing illegal content and protecting users' rights.

2. Gender-Based Violence

The *Gender-Based Violence* category refers to the presence and proliferation of harmful or abusive content that specifically targets individuals based on their gender. This category encompasses a range of violations, including revenge porn, cyber harassment, and content that glorifies or normalizes gender-based violence. The risks associated with gender-based violence on digital platforms extend beyond individual harm; they contribute to systemic inequalities and can lead to severe emotional, psychological, and even physical consequences for victims.

The significance of combating gender-based violence is underscored by its status as a recognised human rights issue. Beyond legal considerations, gender-based violence poses a direct threat to user safety and well-being. Victims of cyberbullying, NCII (commonly known as revenge porn), and targeted harassment often experience long-term psychological distress, reputational harm, and, in some cases, threats to their physical safety.

Risk scenarios illustrate the key challenges associated with this category. The spread of revenge porn or cyberbullying (GB_2) underscores the need for specialized detection tools and immediate takedown procedures to protect victims from ongoing harm. The amplification of violent content by the recommendation system (GB_3) highlights the unintended consequences of algorithmic design, where harmful material can be promoted due to engagement metrics rather than ethical considerations. Additionally, this underscored the importance of reporting and removal mechanisms for abusive content (GB_4), which if and when lacking reveals a critical gap that leaves victims without proper recourse, allowing harmful content to persist and further exacerbate the damage.

3. Protection of Minors

This category addresses the risks associated with underage individuals gaining unauthorized access to adult content, the potential creation or dissemination of exploitative material involving minors, and the failure of safety measures designed to protect them from harmful exposure. In the context of an adult content platform, ensuring the protection of minors becomes even more critical. This requires robust content moderation protocols and user safety policies to ensure compliance with the DSA and international child protection laws. For example, the General Data Protection Regulation (GDPR) mandates stringent safeguards to prevent minors from having their data collected and used in unauthorized ways.

The specific risk scenarios highlight the critical challenges associated with this category. One key concern is minors' use of technical workarounds to bypass content restrictions. The scenario involving minors using VPNs and proxy servers (PM_4) illustrates the sophistication with which underage users attempt to evade parental controls and geographic barriers, necessitating more advanced detection and blocking technologies alongside more effective approaches to implementing controls across the online environment.

Another primary concern is the improper collection or profiling of minors' data (PM_7), which poses privacy risks beyond content access. On the platform, profiling is never performed by default and is only permitted to a limited extent where users have given clear consent, such as by accepting the relevant categories of cookies. Even if minors do not actively engage with the platform, data targeting to the extent allowed by the user that includes information about their online behaviours or interactions could lead to inappropriate advertising, exposure to harmful material, or breaches of data protection laws. The most extreme risk in this category is the presence or distribution of content involving minors (PM_8). This scenario represents a zero-tolerance issue, requiring that such content be immediately detected, removed, and reported to the relevant authorities.

The Protection of Minors category consolidates risks associated with underage access, data protection, and content moderation into a comprehensive concept. Each scenario within this category highlights a specific vulnerability in enforcement or technology that could place minors at risk, whether through access loopholes, data misuse, or exposure to illegal content.

4. Fundamental Rights

This category encompasses a range of risks related to privacy, discrimination, human dignity, and freedom of expression, all of which are integral to online platforms' ethical and legal responsibilities. Under Article 34(1)(b) of the DSA, VLOPs must assess and mitigate any foreseeable negative effects on fundamental rights, aligning with the Charter of Fundamental Rights of the European Union ("Charter"). These rights include human dignity, privacy, personal data protection, freedom of expression, non-discrimination, child protection, and consumer protection.

A vital risk area within this category concerns privacy violations arising from inadequate data governance and unauthorised information sharing. Under Article 7 (Respect for private and family Life) and Article 8 (Protection of personal data) of the Charter, users must be guaranteed control over their personal data, including how it is collected, stored, and shared. However, inadequate online platform practices such as the unauthorised disclosure of user data (FR_1), the monetisation of browsing preferences without consent (FR_2), and user data collection without adequate consent for targeted advertising (FR_3) directly infringe upon these rights. Such violations breach the GDPR and pose significant risks under the DSA's transparency and accountability requirements for online platforms.

Another critical dimension is discrimination in platform policies and algorithmic decision-making, which aligns with Article 21 of the Charter (Non-Discrimination). This category includes risks such as discriminatory ToS (FR_4), where content removal or enforcement practices disproportionately target marginalised groups. Unintentional or systemic algorithmic bias can lead to unfair content suppression, unequal visibility of voices, and biased moderation enforcement. The DSA requires platforms to ensure transparency in algorithmic processes and prevent discriminatory outcomes, necessitating periodic content moderation and recommendation systems audits.

The right to human dignity (Article 1 of the Charter) is also a significant concern within this category. Through their host content, online platforms must prevent exploitation, humiliation, or degradation. Risks such as the dissemination of degrading or exploitative content (FR_5) and the creation and spread of deepfake content featuring individuals in explicit or harmful situations without consent (FR_6) are particularly alarming. Deepfake pornography can be recognised as a growing digital threat. Platforms that fail to detect and promptly remove such content can violate the dignity protections enshrined in the Charter and reinforced by the DSA.

Equally important is protecting freedom of expression and information (Article 11 of the Charter). The DSA emphasises the importance of freedom and pluralism in the media, requiring platforms to ensure fair, non-arbitrary content moderation practices. However, overcorrections in content moderation can lead to excessive restrictions on speech, as seen in overly restrictive content moderation, leading to the unjustified removal of lawful adult content (FR_7). The arbitrary removal of user-generated content without transparent appeal mechanisms (FR_8) is another key risk, undermining the right to fair and balanced content moderation policies.

While the *Fundamental Rights* category directly addresses core human rights concerns, many fundamental rights violations overlap with other risk categories. Addressing these concerns cannot be done in isolation; instead, it requires a coordinated approach that integrates multiple aspects. For instance:

- The rights of the child (Article 24 of the Charter) are closely linked to the *Protection of Minors* category mentioned above.
- Consumer protection (Article 38 of the Charter) is strongly related to risks within the *Transparency and Reporting Obligations category*. Platforms must ensure that users receive accurate, clear, and timely information about their rights, particularly regarding advertising practices, content moderation, and data usage. Risks such as misleading or deceptive advertisements (AD_1–AD_9) threaten consumer rights by exposing users to fraudulent, exploitative, or non-compliant commercial content. Additionally, a lack of transparency in moderation policies (UR_1–UR_6) can prevent users from understanding their rights regarding content removal, complaint handling processes, and moderation restrictions.
- The right to non-discrimination intersects with algorithmic bias and fairness risks found in the Recommender Systems and Algorithmic Design category. Modern platforms use automated decision-making systems to moderate content, recommend media, and enforce platform policies. However, algorithmic models can perpetuate and amplify biases, leading to unintentional discrimination, disproportionate exposure to harmful material, or the suppression of viewpoints. For example, biased content recommendation systems (RS_1–RS_7) can create echo chambers that reinforce harmful narratives or silence marginalized voices by disproportionately filtering their content.
- Freedom of expression and information is affected by multiple overlapping risk categories, particularly *Illegal Content Distribution*, *Gender-Based Violence*, and *Transparency and Reporting Obligations*. Platforms face the dual challenge of removing illegal or harmful content while protecting lawful expression. Risks such as overly restrictive content moderation (FR_7) and arbitrary content takedowns without due process (FR_8) threaten freedom of expression by silencing creators, journalists, and activists. At the same time, insufficient moderation of harmful content (IC_2, GB_2, GB_4) can allow hate speech, cyberbullying, and gender-based violence to thrive, directly harming affected users and undermining their rights to dignity and equal treatment.

5. Civic discourse and electoral processes

The risks categorised under *Civic Discourse and Electoral Processes* share a common theme: the unintentional or deliberate spread of politically misleading or manipulative content in spaces not designed for such discussions. This category refers to the potential risks associated with disseminating politically charged or deceptive content on a platform, even when the platform is not primarily intended for political discourse. It is a trite observation that adult content platforms are not in principle associated with civic discourse or electoral influence risks. With that said WGCZ cannot entirely exclude the possibility that XVideos could be misused for political manipulation or misinformation. It is not inconceivable that such a risk could occur through creator-generated content, paid advertisements, or algorithmic content recommendations. Political groups, activists, or even state-sponsored disinformation actors can attempt to hijack discussion spaces or insert misleading content into video descriptions, comments, or ad placements. Additionally, the increasing integration of AI-generated content and deepfakes adds a new dimension of risk, as manipulated media could be used to spread misinformation about public figures or electoral processes. These concerns may become more pertinent during major election cycles, social movements, or politically sensitive global events.

Several risk scenarios illustrate how these issues can theoretically manifest, even on an adult content platform like XVideos. Coordinated fake news campaigns and political manipulation (CD_1 and CD_2) highlight how organised efforts (bot-driven misinformation campaigns or coordinated deceptive content uploads) can exploit the platform to spread misleading narratives. Algorithmic amplification of polarising content (CD_3) underscores the risks posed by engagement-driven recommendation systems, which can prioritise controversial or emotionally charged material without considering its potential to misinform or incite division. Generating revenue from harmful political content (CD_5) raises ethical and reputational concerns, as the platform could be perceived as monetising misleading narratives, divisive content, or politically sensitive material without adequate oversight.

6. Negative effects on public security

This category refers to the risks associated with the exploitation of the platform by criminal, extremist, or otherwise dangerous actors who can use it as a vehicle for illegal activities. In general, an adult content platform would not be considered a high-risk environment for public security threats such as terrorist recruitment, human trafficking, or extremist propaganda. With that said, WGCZ cannot wholly discount that any online platform with global reach, user-generated content, can be vulnerable to misuse. Criminal networks and extremist groups continuously seek alternative digital spaces to operate under the radar, leveraging platform features for illicit purposes.

Risk scenarios illustrate how public security threats can theoretically manifest in an adult content environment. The use of the platform by extremist groups or criminal organizations (PS_1 and PS_2) highlights that dangerous actors can attempt to operate in digital spaces that do not typically attract law enforcement attention. Whether by spreading extremist propaganda, , or using the platform to target vulnerable individuals, these threats pose significant risks to public safety if they were realised. Additionally, insufficient response to cyber-attacks (PS_4) underscores the growing cybersecurity dimension of public security concerns. If XVideos fails to adequately protect itself against data breaches, hacking incidents, or coordinated cybercrime operations, it can become a gateway for identity theft, financial fraud, or mass data exploitation.

7. Data Privacy and Protection

This risk category encompasses the main aspects of how user data is collected, processed, stored, and shared, which is crucial. Robust data protection measures are essential in an adult content platform where users expect high confidentiality. Given the delicate nature of user data, which often involves private viewing histories (where enabled by the user), and personal data/information, the consequences of privacy breaches can be particularly severe. Given the stigma, risks, and personal nature of adult content consumption, users are susceptible to the security of their personal data. Leaked viewing histories can lead to personal embarrassment, blackmail, or financial fraud, making privacy concerns uniquely high stakes for an adult platform. Also, protecting personal data is a cornerstone of responsible tech governance. Users entrust platforms with their information, assuming that it will be handled securely and only used as explicitly consented.

A data breach (DP_1) is the most immediate and well-known threat, where personal information is exposed due to hacking, insider threats, or insufficient cybersecurity defences. Given the delicate nature of adult content consumption, such a breach could have devastating consequences for affected users, from reputational damage to identity theft. Weak encryption protocols (DP_8) highlight the technical vulnerabilities that cybercriminals can exploit to intercept sensitive data during transactions, logins, or internal data transfers. Failure to properly anonymize data (DP_9) exposes a unique risk where even "de-identified" user information can be reconstructed, allowing individuals to be traced back to their digital

activity. This is particularly concerning in cases where governments, hackers, or malicious third parties seek to uncover user identities for surveillance, extortion, or social harm.

Although *Data Privacy and Protection* is its category, it is closely intertwined with other risk areas that involve data security failures or privacy infringements. In the *Protection of Minors* category, unauthorized data collection related to minors (PM_7) presents serious legal and ethical concerns, particularly if the ToS are violated and children's data is inadvertently processed. Similarly, in the *Fundamental Rights* category, privacy violations (FR_1, FR_2, FR_3) overlap with data protection concerns, as they reflect how user data (whether browsing history, personal details, or payment transactions) can be improperly collected, shared, or monetized without consent. Additionally, the *Advertiser and Commercial Content Integrity* category raises concerns about how advertisers leverage user data. If personal data is used to target individuals with ads based on content preferences of delicate nature (AD_5), the platform risks crossing ethical and legal boundaries, leading to regulatory scrutiny and user backlash.

The scenarios grouped under this category all share a common objective: ensuring that delicate user information remains secure, confidential, and protected from unauthorized access or exploitation. They address technical vulnerabilities (weak encryption, poor anonymization), procedural risks (inadequate user consent or data-sharing policies), and external threats (hacking, unauthorized third-party access).

8. Protection of Public Health

The *Protection of the Public Health* category addresses the potential impact of adult content consumption on users' physical and mental well-being. While adult platforms are designed for entertainment, research is uncertain as to the association of consumption with negative outcomes in users' psychological health, body image perceptions, and overall behavioural patterns. Additionally, online platforms can also be, theoretically, misused to spread misinformation regarding sexual health, relationships, and body standards, which can lead to harmful or misleading beliefs.

Risk scenarios illustrate the potential harm within this category. Addiction concerns (PH_1) highlight the risks associated with compulsive content consumption, which can interfere with daily life, relationships, and psychological well-being, although researchers have declined to classify excessive or compulsive pornography use as an addictive behaviour. The spread of misinformation (PH_3) demonstrates the risks of false or misleading health claims about sexual health, contraception, or body enhancement procedures. The reinforcement of harmful body image ideals (PH_4) demonstrates risk of unrealistic representations of intimacy and physical appearance, although researchers have not found that pornography consumption causes such ideals.

9. Transparency and Reporting Obligations

Transparency and Reporting Obligations category covers XVideos's responsibilities to clearly and accurately disclose how it manages content, moderate user interactions, utilises algorithms, and complies with regulatory requests. This category addresses scenarios involving how openly XVideos shares its internal moderation activities, algorithmic decision-making processes, and adherence to legal transparency mandates established by DSA.

Under the DSA, specific transparency obligations require platforms to regularly disclose detailed reports outlining their content moderation practices, including data on the volume and nature of content removals, automated moderation tools, complaint-handling processes, and engagement with trusted flaggers. Platforms must also provide clear information on their algorithmic recommendation systems, explaining how they influence content visibility and how users can modify or influence recommendations.

Risk scenarios illustrate how transparency risks can manifest. For instance, inadequate disclosure of algorithmic processes (TR_2) highlights the DSA's explicit requirement (under Articles 27 and 38) to provide transparency regarding the logic and parameters behind algorithmic content recommendations and moderation decisions. Users and regulators must understand why specific content is prioritised, demoted, or removed. Lack of clarity can make users feel unfairly treated or manipulated, prompting regulatory inquiries and loss of trust. Similarly, scenarios involving late or incorrect transparency reports (TR_5 and TR_6) directly correspond to the DSA obligations outlined in Articles 15, 24, and 42. These reports' errors, omissions, or delays signal potential non-compliance with the DSA and could invite regulatory action.

10. User Rights and Platform Policies

This risk category addresses whether users receive transparent, timely, and accurate information about their rights, responsibilities, and the rules governing their interactions on the platform. It encompasses how rules are communicated, how effectively users can contest decisions or appeal content removal, and whether significant policy changes are proactively and transparently disclosed to the user base. The importance of this risk category is closely tied to the principle of consumer protection (Article 38 of the Charter). Users have a fundamental right to understand what actions or content could lead to restrictions, suspensions, or other restrictive measures on their accounts.

Risk scenarios highlight the potential threats associated with inadequate user communication and unclear policies, directly intersecting with obligations outlined under the DSA. For instance, a lack of clarity in content moderation policies (UR_1) poses significant risks of user confusion and accidental policy violations. Under Article 14 of the DSA, platforms must clearly communicate their content moderation practices in their ToS. Similarly, failing to announce significant updates to the ToS (UR_4) proactively contravenes Article 14(2), which mandates that platforms must inform users of any substantial changes to their ToS. The absence of a concise, easily accessible, and machine-readable summary of the platform's ToS (UR_6) directly conflicts with Article 14(5), which requires platforms to present their policies transparently and in user-friendly language. This includes summarising key policies and rules in a clear and easily understandable format. Failure to comply with these transparency provisions undermined XVideos' credibility with users and regulatory bodies.

Importantly, *User Rights and Platform Policies* intersect significantly with other risk categories focused on customer protection, most notably *Data Privacy and Protection*, *Transparency and Reporting Obligations*, and *Advertiser and Commercial Content Integrity*. For example, inadequate communication about user data handling and privacy rights directly affects how clearly users understand and trust the platform's data practices. Transparent reporting obligations complement user rights by ensuring users have visibility into how moderation decisions are made and communicated, while accurate and fair advertising practices rely heavily on transparent and understandable user consent and disclosure policies. All risk scenarios under these categories focus on transparency, fairness, and clear communication between the platform and its users.

11. Recommender systems and algorithmic design

The *Recommender Systems and Algorithmic Design* category refers to the risks associated with how online platforms use automated algorithms to select, prioritise, and present content to users. Algorithmic systems could unintentionally amplify harmful or even illegal content. Issues include inadvertently promoting extreme or violent materials, trapping users in echo chambers or feedback loops, and limiting their exposure to diverse content alternatives. Additionally, a lack of user-friendly controls means users can have little influence over what content the algorithms repeatedly suggest, intensifying potential harm.

This risk category matters because poorly regulated recommender systems can severely impact user experience. Algorithms designed primarily to drive engagement can favour controversial content, inadvertently creating feedback loops that expose users repeatedly to distressing material. Also, on an adult content platform, there is a need to be especially vigilant to avoid the amplification of problematic content through algorithmic suggestions, such as content depicting non-consensual acts, revenge porn, or exploitative imagery.

Given the potential negative impact of inadequate recommender systems, the DSA specifically addresses these algorithmic risks (particularly in Articles 27 and 38). Key scenarios highlight these risks associated with recommender systems. For instance, boosting polarising or harmful content (RS_1) demonstrates a core risk inherent to engagement-driven algorithms, where negativity often increases user interaction, inadvertently promoting divisive or harmful material. The scenario involving harmful content loops (RS_5) illustrates a significant problem: when users who interact briefly with non-consensual or violent material can inadvertently trigger algorithms to suggest increasingly harmful content, trapping users in escalating cycles of harmful exposure. Furthermore, the absence of user-friendly mechanisms to adjust recommendations (RS_7) underscores a significant issue of autonomy, leaving users powerless and unable to redirect or influence the content they receive.

These algorithmic risks intersect notably with several other risk categories, underscoring their systemic nature. For example, *Illegal Content Distribution* is closely linked, as recommender systems can unintentionally spread illegal material. Likewise, *Gender-Based Violence* intersects strongly, as algorithms could amplify revenge porn, cyberbullying, or harassment, inadvertently exacerbating gender-based abuse. Moreover, in *Civic Discourse and Electoral Processes*, recommender systems can inadvertently promote misinformation or polarising political content. Also, the *Protection of*

Public Health category intersects here, as algorithm-driven content loops could propagate harmful narratives about body image or mental health.

12. Advertiser and Commercial Content Integrity

This risk category encompasses scenarios related to misleading or undisclosed advertisements, promotion of illegal products, improper use of sensitive personal data for ad targeting, and inadequate transparency in advertising practices. It also includes ensuring the accuracy and completeness of advertising repositories and proper documentation when ads are removed.

Articles 26, 39, and 41 of the DSA explicitly mandate increased transparency, documentation, and reporting obligations for advertising. These requirements closely align with several risk scenarios. For instance, misleading advertisements that are not clearly distinguished from organic content (AD_1) pose potential violations of advertising transparency regulations and create confusion among users. Similarly, the improper use of personal data for ad targeting (AD_5) presents both regulatory and ethical risks, particularly when explicit user consent is not obtained or properly documented. Additionally, incomplete or inaccurate advertising repositories (AD_8) and lack of proper documentation for ad removals (AD_9) undermine compliance and accountability, hindering the platform's ability to demonstrate adherence to transparency requirements.

Furthermore, multiple other risk categories illustrate the systemic nature of advertising integrity issues. *Data Privacy and Protection* is directly linked, as improper data handling for ad targeting overlaps significantly with broader privacy and consent concerns. Unauthorized tracking, profiling, or targeted ads using sensitive user data could simultaneously violate both advertising integrity and legal regulations (e.g. DSA). *Fundamental Rights* intersect with this category, as unethical advertising practices can infringe upon users' privacy and dignity, particularly when personal or intimate data is used without clear consent. *Transparency and Reporting Obligations* are also connected, as the DSA explicitly requires comprehensive advertising documentation and transparency reports, making advertising integrity a critical aspect of regulatory compliance.

5. Risk Assessment

5.1. Influence of Core Systems on Systemic Risks

In alignment with Article 34(2) of the DSA, this risk assessment provides a structured evaluation of how core operational systems, features, and internal policies of the platform can directly or indirectly influence the systemic risks outlined in Article 34(1).

This assessment goes beyond identifying risk exposure by explicitly analysing the design and functioning of the platform's most influential systems: recommender systems, content moderation, ToS and enforcement mechanisms, advertisement presentation systems, and data-related practices. Each of these operational domains has the potential to either mitigate or amplify systemic risks, depending on how they are implemented, governed, and monitored.

To ensure an evidence-based approach, this assessment cross-references the specific risk scenarios, which serve as the foundation for understanding the pathways through which systemic risks can emerge, escalate, or be controlled. Furthermore, the analysis also considers the impact of intentional manipulation of the service and regional differences, recognizing that systemic risks are not uniformly distributed across the EU. Also, by integrating these dimensions, XVideos provides a comprehensive overview of its systemic risk landscape, serving as the foundation for developing risk mitigation strategies under Article 35 of the DSA.

Recommender Systems and Algorithmic Systems

Recommender systems and other algorithmic tools are essential in how users experience and engage with content on XVideos. These systems determine what is prioritized and suggested across user feeds, search results, and content carousels. Recommender systems are designed to enhance user experience and maximize relevance. If these systems are not carefully managed, they can significantly amplify systemic risks. The platform's assessment of its algorithmic architecture reveals multiple risk pathways (see Table 1).

Table 1: Influence of recommender and algorithmic systems on scenario-based systemic risk assessment

Risk ID	Risk Scenario	Core System Influence
RS_1	Algorithmic amplification of polarizing or harmful content	Algorithms trained to optimise for engagement metrics, such as clicks, watch time, or shares, can inadvertently prioritise illegal or incompatible content. This can be especially problematic when malicious actors exploit platform dynamics by producing intentionally harmful content. The risk of deepening user exposure to illegal or extreme material is further amplified when such content evades initial detection and remains within the system. The effectiveness of content moderation mechanisms, therefore, plays a crucial role.
RS_3	Platform's reliance on problematic content for user engagement	
RS_4	Malicious actors or automated bots generate content that spreads disinformation, scams, or exploitative material, which then gets promoted on the platform	
RS_5	Users interacting with non-consensual or violent content are subsequently exposed to increasingly harmful material due to the recommender system's algorithmic behavior	
GB_3	The platform's recommendation system amplifies content that endorses or normalizes gender-based violence	The recommender system can potentially become a vector for societal harm by disproportionately promoting content that normalises gender-based violence or perpetuates unrealistic and harmful body image standards. These types of content can initially evade advanced moderation due to their aesthetic appeal and staged nature (scenes of violence can be fictitious, i.e. the violent scene does not involve real violence, but rather "role-playing") but still potentially contribute to long-term systemic risks by shaping user perceptions and reinforcing harmful norms.
PH_4	Publishing content that reinforces unattainable or harmful body image ideals	

RS_2	Lack of diversity in content recommendations creating echo chambers	A lack of sensitivity to regional, cultural, and linguistic nuances in algorithmic decision-making can lead to disproportionate exposure of specific user populations to harmful or illegal content. Content flagged or deprioritized in major geographic regions could remain active and highly visible in other linguistic areas due to algorithmic detection or moderation coverage gaps.
RS_6	The recommender system does not adequately consider regional or linguistic variations, leading to the promotion of harmful or illegal content within specific cultural contexts	
IC_11	Inadequate adaptation of content moderation to regional languages and cultural contexts leaves users in certain areas more exposed to harmful or inappropriate content	
RS_7	The platform lacks a user-friendly or easily accessible mechanism for users to modify their content recommendations	The absence of accessible user controls to adjust recommendation preferences limits user autonomy and opportunities for self-directed risk reduction. Without transparent or practical tools to opt out of algorithmic personalisation or to reset recommendation settings, users can be continuously exposed to content they deem harmful or inappropriate.

Content Moderation Systems

Effective content moderation underpins the platform's capacity to prevent the proliferation of illegal and incompatible. This process involves a combination of automated detection tools (such as AI-driven classifiers) and human oversight (content moderation team). The following table highlights how different shortcomings in these systems contribute to systemic risks on the XVideos.

Table 2: Influence of content moderation systems on scenario-based systemic risk assessment

Risk ID	Risk Scenario	Core System Influence
IC_7	Inadequate or not robust enough content moderation tools and systems	Automated detection tools often operate on machine learning models that are designed to identify and flag prohibited content. However, due to the inherent limitations of AI, these systems can produce false negatives (missed illegal or incompatible content). Content that should be flagged as violating ToS or community standards can remain undetected. This gap can be exploited by malicious actors who rapidly upload illegal material (for example, copyrighted content, hate speech, CSAM, NCII) before the system identifies and removes it. On the other hand, these systems can produce false positives (legitimate content erroneously taken down). Users can see their lawful posts incorrectly removed or restricted, damaging trust in the platform's moderation processes and, in some cases, infringing upon freedom of expression rights (see also FR_7, FR_8). Over time, these errors can deter lawful creators or alienate user communities.
IC_8	Malicious content creators exploit slow moderation processes to upload and distribute illegal content before it is detected	
IC_10	Automated moderation systems mistakenly classify lawful content as illegal, leading to unnecessary takedowns	
GB_4	Lack of sufficient reporting and removal mechanisms for victims of abusive content	Insufficient reporting or complaint mechanisms impact the platform's responsiveness to victims of abusive or otherwise illegal content. When users lack transparent, accessible processes to complaint takedowns or report harmful material, there is a heightened risk of prolonged harm, such as the continued circulation of gender-based violence content or potentially the unchallenged presence or even monetisation of abused individuals on XVideos.
UR_1	Users do not receive clear and comprehensive information regarding the platform's content moderation policies, procedures, and tools	
FR_7	Excessively restrictive content moderation leads to the unjustified removal of lawful adult content, restricting creators' right to freedom of expression	Overly aggressive or opaque moderation can also negatively affect freedom of expression, discouraging users from sharing legitimate content or engaging in discussions on provided content. Transparency in moderation decisions and a fair and timely appeals process are essential to preserving users' fundamental rights. Without these safeguards, users can perceive moderation as arbitrary or unjust, eroding trust and reducing the sharing of accepted adult content.
FR_8	Arbitrary removal of user-generated content without a transparent appeal process curtails users' right to free speech	

IC_11	Inadequate adaptation of content moderation to regional languages and cultural contexts leaves users in certain areas more exposed to harmful or inappropriate content	Insufficient resourcing for minority languages and regional dialects can create asymmetrical protection across user communities. Content moderation tools and moderator training often focus on major languages, leaving less prevalent dialects or regional expressions at risk of either inadequate enforcement (harmful content not being removed) or excessive removal (misinterpretation of context). This mismatch contributes to systemic inequities. Users in certain regions can be more exposed to illegal or incompatible content and culturally aware support.

Terms of Services (ToS)

ToS establish the foundational rules and expectations governing user interactions on XVideos. They define what is permissible, how complaints are handled, and the potential consequences for violations. Beyond setting ground rules, the ToS embody the platform's commitment to safeguarding fundamental rights and ensuring legal clarity for all users. If improperly formulated or inconsistently enforced, the ToS can give rise to systemic risks by enabling discriminatory practices, eroding transparency, or undermining trust in the platform's governance. Several risk scenarios provided in Table 3 underscore how ToS development and enforcement can exacerbate or alleviate systemic risks.

Table 3: Influence of the ToS on scenario-based systemic risk assessment

Risk ID	Risk Scenario	Core System Influence
FR_4	The ToS disproportionately affect marginalized groups, leading to violations of the right to non-discrimination	The very existence of the ToS carries risks connected primarily with the enforcement clarity of policies and their possible discriminatory nature. If enforcement mechanisms are vague or applied disproportionately, marginalized groups can find their content or accounts unfairly restricted. This can occur due to implicit bias in moderation decisions or insufficient training for staff and automated systems handling reports. Moreover, ambiguously worded sections in the ToS can also lead to inconsistent application of community standards, creating confusion and perceived injustices. Linguistic and cultural accessibility also can lead to unclarity and discrimination. Failure to provide the ToS in local languages (UR_5) or to adapt them to regional legal requirements can create situations where users are unaware of what behaviour is permitted. This lack of clarity often results in inadvertent violations and subsequent sanctions.
UR_2	The ToS lack adequate details about the platform's internal complaint-handling system	
UR_3	Users are not provided with clear and easily accessible information about the reasons that can lead to restrictions on their access to the platform	
UR_5	The platform does not publish the ToS in all the official languages of the EU Member States where it operates	
UR_4	Significant modifications to the ToS are implemented without properly or promptly informing users	Transparency and clear communication are crucial for user empowerment and fairness. Users who cannot easily understand the rationale behind ToS rules are less likely to follow these rules. Inadequate communication around ToS changes (e.g. unclear revisions, insufficient notice periods, and limited user-facing documentation) can degrade user trust. Moreover, the lack of a concise, machine-readable version of the ToS means users and oversight authorities may find it harder to interpret or verify compliance, which increases the risk of misunderstandings or disputes. Ultimately, it can lead to legal uncertainty through vague language or convoluted enforcement workflows and undermine the perceived user's trust.
UR_6	A concise, easily accessible, and machine-readable summary of the ToS is not made available	

Advertisement Systems

Advertising is the only source of revenue on XVideos and a critical avenue for user engagement. However, the mechanics of selecting, targeting, and delivering ads can introduce systemic risks (see Table 4) if not managed with robust oversight and transparency. Using sensitive personal data and potentially misleading or harmful advertisements call for stringent standards to protect user rights.

Table 4: Influence of advertisement systems on scenario-based systemic risk assessment

Risk ID	Risk Scenario	Core System Influence
AD_1	Misleading advertisements not clearly distinguished from regular content	The presence of unmonitored advertisements on the platform can pose risks of user manipulation and exploitation. This is particularly concerning when ads are indistinguishable from user-generated content, potentially misleading users about the commercial or promotional nature of the material—especially in cases involving undisclosed sponsorships or endorsements. Such practices may constitute violations of consumer protection laws. Additionally, advertising systems that rely on sensitive personal data—such as sexual orientation, health status, or religious beliefs—raise significant privacy and ethical concerns.
AD_2	Advertisements promoting illegal or harmful products	
AD_3	Breach of user trust through undisclosed sponsorships	
AD_4	Failure to disclose information about advertisements required by applicable laws	
AD_5	The platform's advertising system uses sensitive personal data (e.g., sexual orientation)	
PM_7	Utilising minors' personal data for profiling, which may result in the presentation of inappropriate or harmful advertisements	Minors' data might be collected despite them being forbidden from accessing adult content and ads, which might cause them to receive ads unsuitable for their developmental stage. This breaches regulatory frameworks designed to safeguard children's privacy.
AD_7	The platform does not maintain a repository of advertisements	Under Article 39 of the DSA, VLOPs are required to maintain a public or regulatory-facing repository of advertisements that includes essential details such as sponsor identity, targeting parameters, and the duration of each ad's display. However, several issues can undermine the effectiveness of this repository and, by extension, the XVideos' compliance with the DSA. If the repository does not contain required information about advertisements, or the platform does not record and explain when an ad is removed or disabled, including the ToS violations or legal grounds for doing so. Also, regulators, civil society organisations, and other stakeholders cannot effectively audit the platform's advertising ecosystem without a reliable ad repository. This may facilitate the spread of misinformation (CD_1, CD_2) or other illegal/harmful content disguised as legitimate advertisements.
AD_8	The ad repository contains incomplete or inaccurate data, including misidentified advertisers, missing targeting parameters, or incorrect details about the ad's reach	
AD_9	When ads are removed or disabled, the platform does not document the reasons for removal within the ad repository.	
CD_1	Coordinated fake news campaigns influencing public opinion	
CD_2	Political manipulation through targeted disinformation	

Data-Related Practices

The collection, processing, and sharing of user data form the basis of key platform functionalities, including recommender systems and targeted advertising. However, these same practices can give rise to significant privacy and security vulnerabilities. The scenarios in Table 5 reflect the principal risks of such vulnerabilities, which can undermine user trust, cause non-compliance with EU data protection regulations and infringe on users' fundamental right to privacy and dignity.

Table 5: Influence of data-related practices on scenario-based systemic risk assessment

Risk ID	Risk Scenario	Core System Influence
DP_1	Data breach exposing user information	Adult content platforms typically collect data beyond basic personal information, including viewing habits. Such records can reveal intimate user preferences, making any data breach potentially more damaging and stigmatising than breaches in less sensitive industries. Insufficient encryption, unpatched software or technical
DP_2	Unauthorized tracking and data collection practices	

DP_8	Weak or inadequate encryption methods for protecting user data	misconfigurations, or unauthorised internal access to systems handling sensitive user records compromises users' safety, reputations, and the platform's reliability.
DP_3	Inadequate security measures leading to account hijacking	Given the particularly sensitive nature of adult content, consent must be explicit, well-informed, and easily retractable. Users expect granular control over how their personal information, especially sexual or intimate content preferences, is collected and retained for marketing or analytics purposes. Potential pitfalls, for example, include ambiguous consent forms that do not clarify how personally identifiable information is used and unclear or buried opt-in/opt-out mechanisms.
DP_5	User data being used without consent in a way that invades privacy	
DP_6	Insufficient clarity regarding the processing and usage of user data	The absence of transparent processes enabling users to exercise their data protection rights (the rights of access, rectification, and erasure) undermines transparency. It can leave users feeling powerless over their personal information. Similarly, vague or overly complex data processing statements reduce comprehensibility, and fuel opacity perceptions.
DP_7	Absence of mechanisms enabling users to exercise their data protection rights	
DP_9	Failure to properly anonymize user data, increasing the risk of user identification	In the context of adult content, failure to properly anonymise or pseudonymise user data significantly increases the risk of personal exposure and harassment, including blackmail or discrimination. Even partial data points can reveal a user's identity, preferences, and behaviours. For adult platforms, such exposures may be significantly more harmful than in other industries. Areas of concern include, but are not limited to, storing full user identifiers alongside browsing or viewing logs that indicate explicit preferences, or inadequate use of hashing or tokenisation for unique data identifiers.
FR_1	Unauthorized disclosure of user data, including identities, viewing activity, and payment details, compromising users' right to privacy	Exploitative data practices on adult content platforms—such as selling sensitive viewing habits or linking personal identifiers to explicit content consumption—violate users' rights to privacy and dignity. The non-consensual sharing or commercialization of such delicate data can cause serious harm to an individual's personal life and mental well-being. Critical concerns in this area include, first and foremost, the profiling of users for advertising or monetization purposes without explicit and informed consent. Equally problematic is the sharing or sale of datasets that reveal explicit viewing behavior or expose potentially sensitive personal attributes.
FR_2	Users' browsing and viewing preferences, including sensitive content choices, are monetized by third parties without consent, breaching their right to privacy	
FR_3	User data is gathered without adequate consent and shared with advertisers, violating the right to privacy	

Intentional Manipulation and Exploitation of the Service

Although adult content platforms do not typically serve as forums for political debate, WGCZ cannot altogether exclude the possibility of intentional misuse and exploitation by malicious actors aiming to spread harmful material, commit fraud, or otherwise bypass moderation controls. WGCZ has assessed the potential impact of inauthentic behaviour, such as automated exploitation and fake accounts, on its ability to prevent and respond to systemic risks (see Table 6).

Table 6: Influence of intentional misuse and exploitation on scenario-based systemic risk assessment

Risk ID	Risk Scenario	Core System Influence
RS_4	Malicious actors or automated bots generate content that spreads disinformation, scams, or exploitative material, which then gets promoted on the platform	Attackers or unauthorised content creators may exploit any time lag in moderation or vulnerabilities in content review systems to disseminate illegal material (e.g., non-consensual sexual acts, CSAM, or content facilitating human trafficking). Within an adult content context, perpetrators may upload such content under the guise of legal adult material, hoping that high volumes of daily uploads or reliance on automated moderation might delay detection. In some cases, criminal organisations could potentially leverage these loopholes to facilitate or advertise human trafficking. Malicious content creators may
CD_1	Coordinated fake news campaigns influencing public opinion	

CD_2	Political manipulation through targeted disinformation	capitalise on slow moderation to circulate prohibited content before it is flagged and removed. Additionally, extremist or terror groups could communicate on adult platforms to avoid detection, posing a broader public security concern.
CD_5	Generating revenue from content that threatens the integrity of democratic processes	
IC_8	Malicious content creators exploit slow moderation processes to upload and distribute illegal content before it is detected	Attackers or unauthorized content creators may exploit any time lag in moderation or vulnerabilities in content review systems to disseminate illegal material (e.g., non-consensual sexual acts, CSAM, or content facilitating human trafficking). Within an adult content context, perpetrators may upload such content under the guise of legal adult material, hoping that high volumes of daily uploads or reliance on automated moderation might delay detection.
PS_1	Terrorist or extremist groups leveraging the platform to recruit members under the guise of adult content	
PS_2	Criminal organizations utilizing the platform for human trafficking or exploitative activities	
DP_4	Users creating fake accounts, impersonating other individuals or legal entities	Fake profiles and impersonation can be used to bypass identity verification, orchestrate scams, or coerce users into revealing personal information, leading to extortion or blackmail. This risk is magnified in an adult content environment, where heightened privacy concerns and stigmatization may discourage victims from reporting such incidents, thus emboldening malicious actors. Moreover, invasive data practices can compound these threats by exposing sensitive user details, fuelling further exploitation or harassment.
DP_5	User data being used without consent in a way that invades privacy	

Regional and Linguistic Specificities

Regional and linguistic differences can significantly affect the platform's ability to moderate content effectively, enforce its ToS, and ensure compliance with key transparency obligations, particularly in the multilingual environment of the European Union. WGCZ assessed local dialects and user language preferences and identified several specific risks and threats (see Table 7).

Table 7: Influence of regional and linguistic specificities on scenario-based systemic risk assessment

Risk ID	Risk Scenario	Core System Influence
IC_11	Inadequate adaptation of content moderation to regional languages and cultural contexts leaves users in certain areas more exposed to harmful or inappropriate content	Due to varying dialects and cultural nuances, content moderation and recommendation systems may fail to identify illegal or incompatible materials appropriately. Automated detection tools and machine-learning models often rely on training data sets that are not comprehensive enough for all EU languages and regional variants. As a result, users in underrepresented language communities may be disproportionately exposed to harmful content. The recommender system may also inadvertently amplify such content in these regions due to limited language-specific settings.
RS_6	The recommender system does not adequately consider regional or linguistic variations, leading to the promotion of harmful or illegal content within specific cultural contexts	
UR_5	The platform does not publish the ToS in all the official languages of the EU Member States where it operates	The absence of localised ToS versions deprives non-English-speaking or minority-language users of clear guidance on platform rules. Users who cannot access the ToS in their native language are at a heightened risk of unknowingly violating the platform's policies or remaining unaware of their rights under these policies. This lack of linguistic inclusivity also undermines trust and could lead to allegations of discriminatory practices.

5.2. Inherent Risk Assessment

In ISO 31000, risk management standards, **inherent risk** is defined as the level of risk an organisation faces **before any controls or mitigating measures** are applied. It reflects the baseline exposure to potential harm, assuming no protective mechanisms are in place.

Likelihood assessment criteria: The likelihood of a risk materialising has been assessed according to the following five categories (each category corresponds to an escalating level of probability):

- **Very unlikely (Score 1):** Very low probability of occurring within the next year (e.g., no historical incidents, strong mitigating controls in place elsewhere)
- **Unlikely (Score 2):** Low probability of occurring within the next year (e.g., rare historical occurrences, some potential vulnerabilities)
- **Possible (Score 3):** Moderate probability of occurring within the next year (e.g., occasional historical occurrences, some vulnerabilities identified)
- **Likely (Score 4):** High probability of occurring within the next year (e.g., frequent historical occurrences, significant vulnerabilities identified)
- **Very likely (Score 5):** Almost certain to occur within the next year (e.g., ongoing incidents, critical vulnerabilities identified)

Impact Assessment criteria: The impact of risk covers the level of harm that can be inflicted on users and WGCZ itself. Impact scores range in the following categories:

- **Insignificant (Score 1):** Minimal to no potential harm to users, society, or the organization. Regulatory intervention is unlikely, and any reputational impact would be isolated.
- **Minor (Score 2):** Potential for some harm to users (e.g., exposure to inappropriate content) or society (e.g., spread of misinformation). The organization could face minor reputational damage or regulatory warnings.
- **Moderate (Score 3):** Increased potential for harm to users (e.g., safety risks, privacy breaches) or society (e.g., manipulation, erosion of trust). The organization could experience moderate reputational damage or regulatory investigations.
- **Major (Score 4):** Significant potential for harm to users (e.g., widespread exposure to harmful content) or society (e.g., manipulation of elections). The organization could face substantial reputational damage, potential suspension of operations, or significant fines.
- **Critical (Score 5):** Severe potential for harm to users (e.g., exploitation, psychological harm) or society (e.g., major disruption of democratic processes). The organization could face severe reputational damage, license revocation, or complete platform shutdown.

It's important to emphasize that even in instances where the current risk assessment cannot predict significant inherent impact to WGCZ e.g. regulatory fines, lawsuits etc., the potential harm to users and society (objects of protection) still exists. Our control environment is designed with this dual purpose in mind. By strengthening our internal controls, we not only safeguard the organization but also protect users and society from the potential consequences of these risks. This approach ensures a comprehensive risk management strategy that prioritizes user safety and societal well-being, while acknowledging the interconnectedness between organizational resilience and the protection of the objects outlined in the DSA.

To arrive at an inherent risk score, each identified risk is rated on both its likelihood and impact dimensions. These two ratings are then multiplied to produce a composite score: $\boldsymbol{Inherent\ Risk\ Score = Likelihood\ Score \times Impact\ Score}$

To visualise the inherent risk level evaluation, WGCZ employs a 5×5 risk matrix that cross-references likelihood (y-axis) with impact (x-axis). Each axis is scored on a scale of 1 to 5 (as described above), and the intersection of the two axes indicates the inherent risk score:

Figure 1: Inherent Risk Scoring Matrix

Risk Value Matrix (x-impact; y-probability)		Insignificant	Minor	Moderate	Major	Critical
		1	2	3	4	5
Very Unlikely	1	1	2	3	4	5
Unlikely	2	2	4	6	8	10
Possible	3	3	6	9	12	15
Likely	4	4	8	12	16	20
Very Likely	5	5	10	15	20	25

Source: WGCZ Risk Register, RA DASHBOARD

The resulting inherent risk score falls into one of three categories:

Low Risk (**Score 1–5, Green**)**:** Risks that fall into the low category represent scenarios where the potential for harm is minimal, and the likelihood of occurrence is relatively remote. These risks are characterized as isolated or rare incidents with negligible impact. Such scenarios have limited platform exposure or narrow scope.

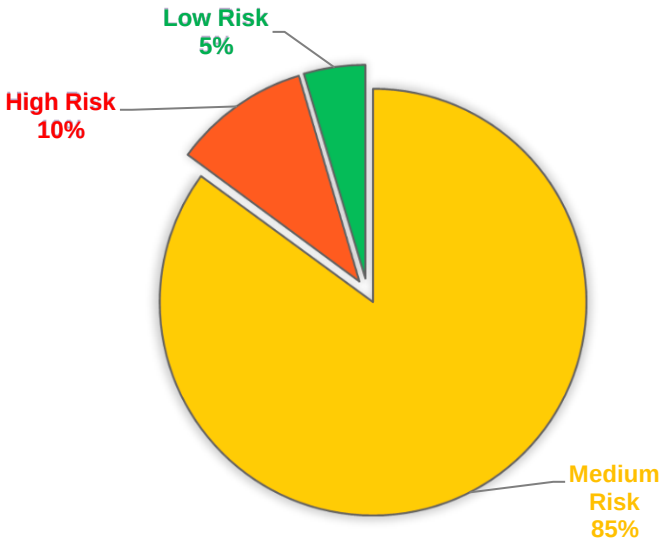
Medium Risk (**Score 6–15, Yellow**): Medium risks represent a moderate level of concern. These scenarios have a meaningful probability of occurrence or a more substantial potential impact, though not at a high level. These risk scenarios include known vulnerabilities or precedents that potentially affect the platform and its users, and they also may impact possible build-up to more severe issues if left unaddressed.

High Risk (**Score 16–25**, **Red**): High risks are the most severe regarding both likelihood and impact and may indicate serious systemic vulnerabilities, threatening platform integrity and safety. These scenarios typically include high-probability events with potentially significant effects on users or the platform (e.g., significant legal, financial, or reputational consequences).

WGCZ evaluated 87 risk scenarios across 14 defined risk categories to determine the inherent risk score. The average inherent risk score across all categories is 11, placing the overall risk exposure at the medium level. This score represents a moderate but notable baseline vulnerability of the platform.

The assessment results plotted in Figure 2 show that most risks (85%) fall within the medium risk range. While not immediately critical, medium risks may escalate if left unmanaged, requiring ongoing monitoring and appropriate control measures. 10% of the assessed risk scenarios were classified as high risk. The highest scoring risks are associated with areas such as the upload and distribution of CSAM, the use of the platform to facilitate or normalize gender-based violence, the exploitation of minors by-passing access controls, and the dissemination of non-consensual or humiliating content (e.g., NCII) that infringes on users' dignity. These risks are severe due to their potential for legal, reputational, and user safety consequences, and they highlight the platform's exposure to systemic vulnerabilities in highly sensitive areas. Only 5% of the risk scenarios were categorized as low inherent risk. These scenarios are generally associated with isolated use cases or areas where systemic relevance to the platform is minimal, such as rare cases of regional political manipulation.

Figure 2: Results of inherent risk assessment



Source: WGCZ Risk Register, RA DASHBOARD

When examining the average inherent risk score across categories (see Figure 3), it becomes evident that certain areas present greater risk concentrations than others. However, when assessing systemic risks, it is essential to evaluate the severity of individual risks through inherent risk scores and consider the number of risks identified within each category. The frequency of risks provides important insight into the breadth of the risk's exposure. In the aggregate, a category with a high number of medium-scoring risks may pose a level of exposure comparable to or even greater than that of a category with fewer but higher-scoring risks. In other words, volume amplifies significance. Multiple interconnected risk scenarios may indicate systemic weaknesses that cannot be fully captured by score alone. Therefore, considering both inherent risk score and risk concentration offers a complete understanding of vulnerabilities and where WGCZ should prioritise its resources.

Figure 3: Comparison of inherent risk by category

Risk Category	Number of Risks	Inherent Risk Score
Fundamental Rights - Dignity Violations	2	16
Protection of Minors	9	13
Gender-Based Violence	6	13
Illegal Content Distribution	11	12
Transparency and Reporting Obligations	6	12
Negative effects on public security	4	12
Data Privacy and Protection	9	10
User Rights and Platform Policies	6	10
Fundamental Rights - Privacy Violations	3	10
Advertiser and Commercial Content Integrity	9	10
Protection of Public Health	6	9
Recommender systems and algorithmic design	7	9
Fundamental Rights - Freedom of expression	2	9
Fundamental Rights - Discriminatory Practices	1	6
Civic discourse and electoral processes	6	5
Total	**87**	**11**

Source: WGCZ Risk Register, RA DASHBOARD

When considering both the inherent risk score and the number of risks in each category, it was found that *Illegal Content Distribution* is the category with the highest total of 11 distinct risk scenarios and has an average inherent risk score of 12.

The volume of risks in this area reflects the platform's wide variety of threats regarding content legality and enforcement efficiency, ranging from CSAM and hate speech to moderation delays and regional moderation gaps.

Protection of Minors and *Gender-Based Violence* closely, with 9 and 6 risk scenarios, respectively, both averaging an inherent risk score of 13. These categories feature some of the platform's highest-scoring individual risks, such as minors managing to access adult content despite the prohibition and by by-passing age assurance processes (score 25) and the distribution of non-consensual (e.g. NCII) or violent material. While they do not contain the highest number of risk scenarios, their consistently elevated scores across scenarios indicate a concentration of high-severity issues.

Fundamental Rights – Dignity Violations category, which includes only 2 risk scenarios, has the highest average category score of 16, driven by the severity of cases involving deepfake exploitation and degrading content. Although the number of risks is low, the critical nature of each scenario elevates this category to the top of the inherent risk score comparison.

A group of categories, including *Transparency and Reporting Obligations, Negative Effects on Public Security, User Rights and Platform Policies, Privacy Violations*, and *Advertiser and Commercial Content Integrity*, each contain 4 to 9 risk scenarios and average scores around 10 to 12. These categories represent important areas of moderate to elevated concern and contribute significantly to the overall risk score.

Categories such as *Protection of Public Health, Recommender Systems and Algorithmic Design, and Freedom of Expression* fall into the lower-medium risk band with average scores of 9. Each contains between 2 to 7 risks, generally tied to amplifying harmful content, misinformation, and algorithmic blind spots. These are considered significant due to architecture, and automation in influencing user experience, including the potential for both amplifying harmful content and under-reporting.

At the lower end of the spectrum, *Civic Discourse and Electoral Processes* include 6 risks but hold the lowest average score of 5. Most of these risks are assessed as very unlikely to occur or currently limited in scope, often due to platform content specificity. *Fundamental Rights – Discriminatory Practices category*, while limited to a single identified scenario, has an inherent risk score of 6. Although this represents a relatively low average, discrimination's ethical and legal sensitivity requires ongoing monitoring despite the lower baseline exposure.

Inherent risk assessment results reflect a risk landscape where risk level (as measured by score) and risk volume (as measured by a number of risk scenarios in each category) shape the platform’s priorities. Categories with high scores but fewer scenarios, such as dignity violations, require focused intervention on specific severe risks, while high-volume, medium-risk categories, such as illegal content or advertising practices, demand scalable governance structure (e.g., policies, decision-making procedures, risk controls, compliance processes) that can keep working effectively even as the scale of operations increases (more users, more content, more markets, or more complexity). This approach ensures mitigating high-score risks and widespread risks in the platform’s ongoing operation.

5.3. Mitigation Measures

5.3.1. Mitigation Measures Register

The platform's risk management system and related risk assessment report in the past iteration relied on a dedicated Risk Register and detailed risk cards. Each risk card included all essential information about each risk scenario, such as particular risk assessment and mitigation measures/controls, including justification and effectiveness. While this method offered a convenient, high-level picture of how risks might evolve (based on probability and impact scoring), it had several limitations.

The visual format of each risk card inherently emphasized the risks themselves rather than the mitigation measures to be applied. Although descriptions of the measures were included, they were typically presented as bullet points or summaries. As a result, it was often difficult to track how measures were implemented or why certain ones were considered more critical than others. In other words, the cards were more risk-focused than measure-focused.

Over time, the Risk Register expanded to cover 87 different risk scenarios, making it increasingly burdensome to keep every individual card updated, significantly when a core mitigation measure (e.g., a new AI scanning tool) changed. Minor updates had to be reflected across multiple cards, which increased the likelihood of errors or inconsistencies.

Additionally, because each risk card listed measures somewhat independently, there was little consolidation or cross-referencing of shared controls. The same measure might appear under different names (e.g., “AI-based scanning,” “AI scanning tool,” or “automated hate speech detection”), causing confusion and making it more challenging to evaluate the overall effectiveness of a given control across multiple risk scenarios.

WGCZ has adopted a **new risk management system component known as the "Mitigation Measures Register" in response to the challenges inherent in the previous Risk Register**. This registry is specifically designed to provide a cohesive, measure-centric view of how the platform addresses various risk scenarios. Key elements of the Mitigation Measures Register include:

- A systematic listing of mitigation measures, with each entry detailing the measure’s name, type, purpose, and assigned ownership.
- Categorization of each measure as “preventive” (aimed at avoiding unwanted events), “detective” (designed to identify issues that have bypassed preventive controls), “corrective” (intended to resolve or mitigate issues that have already occurred), or “reactive” (implemented in response to unforeseen or exceptional incidents to mitigate impact and ensure continuity).
- An assessment of each measure’s effectiveness, rated as “high” (strong ability to mitigate the targeted risk; technically robust, actively enforced, and/or clearly reducing harmful outcomes), “moderate” (functioning and contributing to risk reduction, but conditionally effective and part of a broader approach), or “low” (limited impact due to technical or contextual constraints; narrow scope or reliance on inconsistent user action).
- Explicit mapping of each measure to specific risk scenarios (e.g., illegal content, data breaches), clearly demonstrating how it addresses the identified risks.

Having a single "source" for all measures ensures that updates (for example, refining an age-verification approach or deploying a new content moderation measure) are recorded clearly and in a standardised format. This eliminates duplication, mismatched references, and inconsistencies that arose when the same measure was described differently across multiple risk cards.

Also, rather than fixating on individual risks' probability and impact scores, this measure-centric approach keeps all teams focused on how effectively each measure performs. For example, if the moderation team has been expanded but response times to flagged content show no improvement, that discrepancy becomes immediately visible. It prompts a deeper evaluation of the measure's real-world impact and encourages continuous refinement.

Regulatory bodies, auditors, and other stakeholders can also more easily assess WGCZ's approach to risk management by reviewing a consolidated list of all measures. This is especially beneficial for responding to formal information requests, transparency reporting obligations, or external audits, as it removes the need to piece together data from multiple disparate documents.

Control effectiveness assessment

The assessment of control effectiveness was conducted qualitatively, considering a range of factors that were likely to influence practical effectiveness.

To ensure consistency and comparability across all assessed measures, a predefined rating scale was applied. This scale categorizes each measure’s effectiveness as high, moderate, or low, based on qualitative criteria reflecting its practical risk mitigation capacity. The definitions of these rating levels are provided in Table 8.

Table 8: Control effectiveness scale

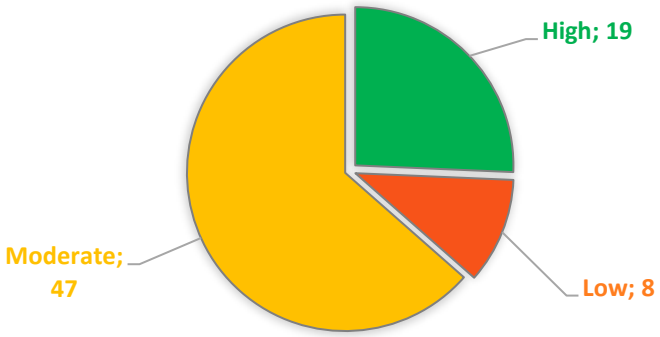
Effectiveness Level	Description
High	The control demonstrates a strong ability to mitigate the targeted risk. It is either technically robust, actively enforced, or clearly reduces harmful outcomes.
Moderate	The control is functioning and contributes to risk reduction, but its effectiveness is conditional. Often part of a broader multi-layered approach, but not fully sufficient on its own.
Low	The control has limited impact on the addressed risk or is constrained by technical or contextual limitations. It may detect only a narrow range of cases or rely on user action without consistent enforcement.

Source: WGCZ Mitigation Measure Register, Dashboard

Only those criteria that were relevant to the nature of the measure and supported by available data were taken into account, including but not limited to technical reliability (such as documented error rates or known detection limitations), or user engagement (in cases where the measure relied on user action, awareness, or active configuration). Although each measure received an overall effectiveness rating, it was acknowledged that its actual impact on residual risk could vary depending on the nature of the specific risk scenario. This relationship between control measures and their varying influence on residual risk is further addressed in the following section.

A total of 74 mitigation measures were assessed for effectiveness. The results are visualized in Figure 4 and reflect the overall distribution of effectiveness ratings across the platform's risk control landscape. The majority of measures (48 out of 74) were rated as having moderate effectiveness. These controls typically contribute to risk reduction but show limitations that affect their overall reliability or consistency. In many cases, their impact is conditional – shaped by factors such as user engagement, resource intensity, scope of detection, or the need for continuous updates and manual oversight. These measures include, for example AI scanning, user reports, Internal VPN and Firewalls and access restrictions, user-controlled content, or dedicated reviews. A smaller portion (19 out of 74 measures) were assessed as having high effectiveness. These controls demonstrate strong risk mitigation capacity, either due to their technical robustness, targeted scope, or consistent enforcement. These measures include, for example clear visual marking of advertisements and informational label on ads, manual moderation process and content ghosted measure, or IS security measures. Only 8 measures received a low effectiveness rating. These controls tend to face significant technical or contextual limitations, such as low accuracy rates, dependence on user action without systemic reinforcement, or limited scope of application. While some may still serve a complementary role, their standalone impact on reducing platform risk remains limited. These measures include, for example MD5 hash, notice mechanism for trusted flaggers, terms of service, warnings about adult content, or age confirmation. While their standalone impact is limited, especially without additional enforcement layers, they can still play a meaningful role in specific contexts, for instance, deterring casual underage access or reinforcing user awareness of platform boundaries.

Figure 4: Results of control effectiveness assessment



Source: WGCZ Mitigation Measure Register, Dashboard

5.3.2. Mitigation Measures in Place

a) Content Moderation

The platform's content moderation framework is designed to detect and remove illegal content, prevent the spread of harmful or non-consensual material, and balance fundamental rights (such as freedom of expression, non-discrimination, or child rights) with users' safety and well-being. Below are short descriptions of the primary measures/controls applied.

Automated Tools (MD5 Hash, AI Scanning, Thorn Safer, Google SafetyNet, Hive Classification)

These automated detection systems quickly flag known illegal content (including CSAM, hate speech, or violent material) by matching uploads against hash databases such as MD5 or scanning new content via artificial intelligence. They are essential for rapid, large-scale detection of high-risk materials and serve as an initial filter that can significantly reduce the prevalence of illicit content before it reaches public view. Thorn Safer compares uploads against existing CSAM databases, and Google SafetyNet and Hive Classification help identify violent, hateful, or otherwise prohibited media. Specifically, these automated tools address the illegal distribution of incompatible content such as copyrighted material, hate speech, malware or phishing links, non-consensual sexual acts, gender-based violence, and animal abuse. They also help mitigate the risk of inadequate moderation tools (for example, due to regional language gaps) or other situations where large-scale scanning is critical to user safety. Applying and ongoing improvement of this detection methodologies demonstrates compliance with Article 35(1)(d) of the DSA, which specifically references the need to adapt and test algorithmic detection.

Manual Moderation Process & Multi-Layered Review

The moderation team performs in-depth reviews of any content flagged by automated systems or user reports, also verifying participant age and consent and checking for content that may be illegal or non-consensual. This additional human oversight is crucial because algorithms can misinterpret context. Moderators detect nuances (e.g. cultural references or borderline cases) where automated filters might err, significantly lowering the incidence of false positives and negatives. By deploying such a multi-layered approach (automated plus human review), the platform aligns with Article 35(1)(c) of the DSA, which highlights the importance of adapting effective content moderation processes. Manual moderation is indispensable for resolving a wide range of risks, including, for example, illegal content distribution of various types (IC_1 to IC_11) and gender-based violence (GB_1 to GB_6). Also, it mitigates scenarios were specific contexts, such as user age, local culture, or intent, must be accurately assessed.

User Reports (Notice and Action Mechanism) & Notice Mechanism for Trusted Flaggers

Under this mechanism, any user (including trusted flaggers) can submit a report if they encounter content that appears illegal or incompatible. The relevant teams receive alerts to review the reported content, drawing on human judgment. User reports identify potential violations that might elude algorithmic tools, especially in nuanced or context-specific content cases. The platform complies with DSA Articles 16 and 17, which mandate user-friendly notice and action procedures. This user report mechanism is crucial for a broad spectrum of risks, including illegal content (IC_1 through IC_11), gender-based violence (GB_1 through GB_6), and situations involving misinformation or harmful content that automated systems might not fully capture.

Also, certain NGOs, law enforcement bodies, and specialized organizations enjoy "trusted flagger" status, allowing their notices to receive priority handling by moderators. This process ensures that especially severe or sensitive cases undergo swift review and action. By collaborating with trusted flaggers, the platform fulfils Article 22 of the DSA and the broader emphasis on expert coordination in Article 35(1)(g). This system mitigates high-risk scenarios including, but not limited to, CSAM (IC_4), non-consensual sexual acts (IC_5), various forms of gender-based violence (GB_1 to GB_6), and other forms of illegal content where timely intervention is crucial.

Content Ghosting

The content ghosting measure temporarily hides flagged content from public view without permanently deleting it, allowing moderators time to conduct a proper review. During this review window, only the content uploader and authorised internal teams can access the material. This protects the broader user community from potential exposure to illegal or harmful content. This approach strikes a balance between prompt intervention and procedural fairness, aligning with the DSA's

requirements to manage risk while upholding user rights. Content ghosting is especially valuable in cases where rapid isolation of suspect material can prevent harm.

Complaint-Handling System

The complaint-handling system provides a formal channel for users to appeal content or account restrictions. Affected users can file a complaint that triggers a structured review process, enabling relevant teams to assess whether the original decision was justified. This process upholds fundamental rights by ensuring moderation outcomes remain transparent, proportionate, and reversible when errors are identified. It aligns with Article 17 of the DSA, which mandates internal complaint-handling mechanisms and helps mitigate several important risks, such as automated tools mistakenly classifying lawful content as illegal or false positives resulting from algorithmic scanning or mistaken assessments. For example, users who believe their content was unfairly flagged as hate speech (IC_2) or non-consensual sexual material (GB_2) can seek swift remediation if the content is found to be lawful. Moreover, the complaint mechanism strengthens the overall moderation ecosystem by incorporating continuous user feedback, reducing the impact of tool limitations (IC_7). It also lowers the risk of unjustified censorship or denial of lawful expression, safeguarding freedom of speech following fundamental rights obligations (FR_7 and FR_8).

Moderation Team Training

Moderation team training is a comprehensive measure that enhances WGCZ's overall systemic risk management ecosystem. By providing ongoing education on emerging issues and developing cultural or linguistic sensitivities, training empowers human moderators to make accurate, and context-informed decisions. This reinforces the importance of internal processes for content moderation. Its broad scope addresses a wide array of risk scenarios. For instance, in risk scenarios related to illegal content distribution (IC_1 to IC_11), adequately trained moderators are best equipped to identify and respond to issues like unauthorized copyrighted material, hate speech, malware or phishing links, CSAM, non-consensual acts, or animal abuse, ensuring that the platform's monitoring policies are enforced with due care for user rights. In gender-based violence risks (GB_1 to GB_6), training prepares moderators to recognize and respond promptly to different manifestations of gender-based violence, including explicit physical harm to "revenge porn" or subtler forms of harassment or stereotyping. Concerning freedom of expression risks (FR_7 and FR_8), a well-informed team is less likely to mistakenly remove lawful content or overlook contextual factors, thereby reducing instances of undue censorship and helping strike the correct balance between removing illegal material and protecting user rights. It has systemic relevance because moderation team training is foundational to how the platform manages content. In other words, it weaves through the entire risk management framework. This ongoing training measure significantly strengthens the platform's capacity to handle diverse vulnerabilities.

b) Terms of Service (ToS)

The ToS and its enforcement mechanisms establish legal clarity, set expectations for user conduct, support regulatory compliance, and serve as an enforcement framework against illegal and incompatible content. These measures are essential to fulfilling several obligations under the DSA, particularly those relating to transparency, accessibility, user protection, and uniform rule enforcement.

Comprehensive ToS

The ToS form a comprehensive document outlining contractual obligations, rules, and user responsibilities. They provide clear and detailed guidelines on acceptable user submissions, intellectual property rights, content moderation policies, user reporting mechanisms, and protections against illegal and incompatible content types. Specifically, the following critical clauses within the ToS are directly referenced in the platform's risk mitigation measures:

- Intellectual Property Rights (Point 6) – This provision prohibits users from uploading or sharing copyrighted material or any content protected by intellectual property rights without proper authorization. This directly addresses risks related to copyright infringement (IC_1).
- User Submissions (Point 7) – The ToS establish clear guidelines governing user-generated content, explicitly prohibiting submissions of hate speech (IC_2), content depicting violence or gender-based harm (GB_1, GB_2),

and material related to illegal or exploitative activities. This clause is a legal basis for promptly moderating and removing problematic submissions and helps mitigate risks associated with malicious or abusive content uploads.

- Content Moderation (Point 8) – This section defines the processes related to XVideos's content moderation practices, including the "notice and action mechanism," enabling users to report illegal content. It also outlines procedures for content removal decisions, thereby addressing risks related to the timeliness and effectiveness of moderation activities (IC_9, UR_3). This measure ensures clear accountability and transparency regarding moderation decisions.

- Child and Non-Consensual Content (Point 3) – The ToS explicitly and vigorously prohibit the uploading, sharing, or promotion of CSAM, NCII, or content depicting non-consensual sexual acts, underscoring the zero tolerance of the platform for any such content. This crucial provision directly addresses severe, high-impact risks, such as the distribution of CSAM (IC_4) and non-consensual sexual content (IC_5), reinforcing the platform's zero-tolerance approach to exploitation and abuse.

- Terrorism and Physical Harm Violence (Point 4) – Under this clause, the platform explicitly forbids content that promotes terrorism, extremist ideologies, violent acts, or physical harm toward others. This provision addresses significant public security risks, such as the potential exploitation of the platform by extremist or terrorist groups (PS_1, PS_3).

- Access and Age Restrictions (Point 2) – Clear age restrictions are established in the ToS, explicitly prohibit access to the platform by anyone under the age of 18. In addition to this prohibition, users must affirmatively attest, at the point of access, that they meet this age requirement. Together, this provision mitigates explicitly risks involving minors' exposure to adult material and inappropriate content (PM_1, PM_2, PM_5, PM_7). The ToS also outline measures such as content labelling, age confirmation, and parental control instructions to protect minors further.

- Creation of an Account (Point 5) – This section establishes strict rules around account creation, explicitly prohibiting using fake identities or impersonating other individuals or entities. This clause mitigates risks associated with identity theft, fraud, and manipulation by malicious actors creating false accounts (DP_4, RS_4).

- Internal Grievance System and Dispute Resolution (Point 13) – The ToS clearly describe internal complaint-handling processes, dispute resolution mechanisms, arbitration procedures, and guidelines for users wishing to appeal moderation decisions. This comprehensive approach addresses procedural fairness and accountability risks, providing precise mechanisms for challenging content and account restrictions (UR_2, UR_3, FR_8).

The comprehensiveness of the ToS provides legal clarity and transparency and serves as the legal backbone for moderation actions. Moreover, by defining explicit processes for reporting, appeals, and dispute management, the ToS forms a framework that reinforces compliance with Article 35(1)(b) of the DSA and safeguards fundamental user rights.

Public Availability & EU-Language Versions

To ensure accessibility and clarity across all jurisdictions in which the XVideos operates, the ToS are publicly available in all official languages of the EU. This approach also fulfils the requirement outlined in Article 14(6) of the DSA. Such availability is crucial for enabling users to fully understand the rules governing their interactions with the platform, particularly given the complex nature of issues like intellectual property rights, content moderation policies, prohibited material, and mechanisms for dispute resolution.

Providing multilingual versions, WGCZ mitigates specific risk scenarios, notably those related to linguistic misunderstanding or unintentional non-compliance arising from language barriers. For instance, ensuring users clearly understand moderation standards in their native language directly reduces the risk scenario IC_11, where inadequate adaptation of content moderation to regional languages and cultural contexts could otherwise lead users in certain areas to be disproportionately exposed to illegal or incompatible content.

Furthermore, clear multilingual ToS enhances user rights and fairness (UR_5), explicitly addressing the risk that users could fail to adhere to rules they do not fully comprehend due to language limitations. In addition, consistent availability in all EU languages supports equitable enforcement and helps mitigate potential discriminatory practices or perceptions thereof (FR_4), as it ensures all users, irrespective of linguistic or cultural background, have equal opportunity to grasp the platform's policies.

Summaries of ToS

WGCZ provides consist understandable summaries of its ToS. These plain-language summaries clearly outline critical policies such as prohibited content, moderation procedures, user rights, reporting mechanisms, and avenues for appeals. This measure aligns directly with Article 14(5) of the DSA, which mandates that platforms must offer concise, easily accessible, and machine-readable summaries of their terms.

Providing simplified summaries significantly mitigates risks associated with user misunderstanding of moderation policies or platform rules, specifically addressing risk scenarios such as insufficient clarity of content moderation policies (UR_1), lack of easily accessible ToS summaries (UR_6), and overly complex reporting procedures preventing effective user flagging of harmful content (IC_9).

c) Data Privacy and Security

The primary objectives of data privacy and security measures implemented by the platform are to protect user data against unauthorized access and empower users to exercise meaningful control over their personal data. Given the delicate nature of data collected on platforms offering adult content, robust security and privacy measures are essential to mitigate systemic risks such as unauthorized disclosures of user identities, and preferences. They also help prevent risks related to unauthorized tracking and data collection practices, inadequate security leading to account hijacking, and improper data processing transparency.

Privacy Policy & Cookie Policy

XVideos has established clear and comprehensive Privacy and Cookie Policies. These policies describe the types of personal data collected (including sensitive content preferences and usage data), methods of data processing, purposes for which the data is collected, and data retention periods. Notably, the Cookie Policy provides users with clear information on cookie usage and offers practical tools enabling users to consent to, manage, or completely opt out of cookies.

This measure directly mitigates the risks identified in scenarios related to unauthorized or unclear tracking of users' browsing and viewing preferences (FR_2, DP_2), unauthorized data sharing with third parties (FR_3), insufficient transparency regarding data usage (DP_6), and non-consensual usage of user data for profiling or advertising purposes (PM_7). Moreover, these policies align closely with the DSA's transparency requirements (Article 35(1)(i)), ensuring users are well-informed about how their data is processed and empowering them to manage their consent and privacy preferences actively.

Information Security Measures (Encryption, Firewalls, and Two-Factor Authentication)

The platform employs robust information security (IS) measures, including end-to-end encryption for user chats, firewall protections, access restrictions, and mandatory two-factor authentication (2FA) for sensitive account operations. These technical controls minimise the risk of unauthorised data access, breaches, and account hijacking (DP_1, DP_3).

Given the platform’s nature, handling personal user information requires stringent protective measures to safeguard users against severe privacy violations (FR_1). Implementing strong encryption, strict firewall rules and multi-factor authentication directly minimises the potential damage of security breaches.

Regular Code Reviews & Hacker One Program

To proactively detect and address potential security vulnerabilities, WGCZ conducts regular code reviews involving systematic inspections by developers and security experts. Also, the platform participates in the HackerOne vulnerability testing program to independently test and identify security flaws within the platform infrastructure.

These ongoing proactive assessments significantly reduce the risks of vulnerabilities leading to unauthorized data disclosure, breaches, or account compromise (DP_1, DP_3, DP_8). This approach aligns with best practices for robust security and is indirectly required by the DSA, specifically under Article 35(1)(f), regarding reinforcing internal processes and resources to detect systemic risks.

Versions Lifecycle Policy of Storage

WGCZ employs a data-lifecycle policy for storage of personal data. Retention periods are purpose- and category-specific and are set out transparently in the platform's publicly available Privacy Policy. Once the applicable period for a given category expires, the data is automatically deleted or irreversibly anonymised, substantially reducing the volume of stored sensitive information. This measure is crucial in minimising risks associated with unnecessary retention of personal data, unauthorised disclosure, or misuse (DP_1, DP_9). By systematically removing or anonymising older data, the platform aligns itself with GDPR's data minimisation principle, effectively reducing the potential impact of data breaches and unauthorised access.

User-Controlled Tracking & Data Collection

WGCZ provides straightforward, user-friendly interfaces that empower users to actively manage their data and privacy preferences. Users can easily activate or deactivate data-driven features such as tracking cookies and personalized ads, where allowed by the user. This measure mitigates risks arising from unauthorized or undesired data collection and profiling (FR_2, FR_3, DP_2, DP_5, DP_6). Empowering users with granular control over tracking and data collection practices directly addresses the requirement for transparency and user empowerment in Article 35(1)(i) of the DSA.

d) Protection of Minors

The vital objective of the platform's user protection strategy is to *effectively* prevent underage users from accessing adult or otherwise harmful content, reduce potential harm to vulnerable users, and proactively uphold rigorous child protection standards in line with the DSA. This set of measures directly addresses critical risk scenarios identified in the platform's risk assessment, such as inadequate verification procedures allowing minors unauthorized access, technological circumvention of safeguards, inadvertent exposure due to mislabelling, and risks associated with minors developing unhealthy behaviours after exposure. Per the DSA Article 35(1)(j), WGCZ has taken targeted measures to protect the child's rights, including age assurance and parental control tools, and tools aimed at helping minors signal abuse or obtain support. The platform is also actively involved, has invested and continues to invest to advance effective solutions for preventing minors from accessing adult content.

Age Confirmation & Explicit Content Warnings

A fundamental step in protecting minors on the platform involves mandatory age confirmation for all users accessing adult content. Before viewing such content, each user is explicitly asked to confirm that they meet the required age threshold. This confirmation is accompanied by a clear, prominent, and explicitly worded warning about adult content, informing users upfront about the nature of the material they are about to access. This is an essential but necessary measure to deter underage viewers and signal to adults that the platform hosts sensitive content.

Self-declaration mechanisms remain the only tools that balance proportionality, accessibility, and privacy. As highlighted in the European Digital Rights (EDRi) position paper (2024)[13], stricter verification mechanisms may infringe on fundamental rights, mainly when they rely on biometric data or identity documents. The Open Technology Institute[14] has concluded that no existing tool fully meets the combined criteria of effectiveness, inclusiveness, and data protection. Therefore, qualified disclaimers remain a foundational control, particularly when integrated into a broader risk mitigation framework.

Although self-declared age assurance represents a balanced control, XVideos acknowledges it as potentially circumventable by determined and/or unsupervised minors, which is reflected in the relevant risk scenarios. However, this measure addresses risks associated with insufficient or easily bypassed age verification (PM_1, PM_3).

Page Blurring & RTA Label

Addressing risks such as minors inadvertently accessing adult content due to content labelling inaccuracies (PM_5) or through devices left logged in by adults (PM_2), the platform implements page blurring on its landing pages. Initially,

[13] EDRi. (2023, October). Online age verification and children's rights: Position paper. European Digital Rights

[14] Forland, S., Meysenburg, N., & Solis, E. (2024). Age Verification: The Complicated Effort to Protect Youth Online. Open Technology Institute

obscuring explicit visual content is a barrier to immediate unintended exposure. It creates an additional step, requiring deliberate user action before content is visible.

Simultaneously, the platform applies adequate content labelling through the universally recognised RTA (Restricted to Adults) label. Labelling content as "adult-only" significantly aids parents and guardians in configuring effective browser-level or device-level filters. These filters leverage such standardised labelling to block access pre-emptively.

Parental Guidance

International studies, including those referenced in the OECD Risk Typology[15], emphasize the critical role that parents and educators play in teaching children about online safety and helping them navigate digital risks. For example, the OECD (2020)[16] highlights the importance of parental and educational support in fostering digital resilience. Similarly, Burns & Gottschalk (2019)[17] stress that understanding and mitigating risks requires a nuanced, comprehensive approach—one that involves parents and educators, rather than relying solely on technical solutions. Recognizing this, XVideos provides comprehensive parental guidance resources. These materials offer instructions for parents and guardians on implementing robust parental controls, using browser security features, and setting up third-party filtering tools.

This measure addresses multiple risks related to minors bypassing platform-level controls through technological means such as VPNs (PM_4), exploiting weak password practices (PM_3), or manipulating customer support procedures (PM_6). The platform enhances minors' protection beyond its immediate technological perimeter by empowering parents and guardians with knowledge and practical tools.

Ongoing Assessment of Age Assurance Measures

Acknowledging there is a potential challenge posed by minors circumventing existing access control measures (particularly highlighted by risks PM_1, PM_3, and PM_4), XVideos actively undertakes ongoing assessments of age verification tools. This measure involves the platform continuously evaluating the effectiveness, reliability, privacy implications, and proportionality of emerging age verification technologies, including methods such as facial age estimation, Bank iD, and the EU Digital Identity Wallet. The goal is to ensure robust and *effective* protection of minors from exposure to adult content while simultaneously upholding fundamental user rights, data protection standards, and the principle of proportionality mandated under EU regulations.

Based on the studies, implementing adequate age verification has consistently proven challenging from technological and human rights perspectives. Currently, available methods often encounter significant limitations, including technical immaturity, risks to personal privacy, risks of data breaches and resulting crime — including extortion of compromised users, susceptibility to circumvention, discriminatory effects, and potential infringement upon users' anonymity. Therefore, continuous evaluation of these technologies is important for identifying reasonable solutions.

Research from key international institutions underscores the complexity and shortcomings of existing age verification tools. The French Commission on Information Technology and Liberties (CNIL)[18], and Australia's eSafety Commission[19] all concluded in recent assessments that, as of 2024, no existing age assurance technology fully meets requirements for accuracy, inclusiveness, privacy protection, and proportionality. Each institution emphasized the necessity of supplementary measures, such as user education, clear labelling, content moderation, and proactive parental involvement, to enhance online safety.

[15] OECD. (2021). *Children in the digital environment: Revised typology of risks* (OECD Digital Economy Papers No. 302). OECD Publishing. https://doi.org/10.1787/9b8f222e-en

[16] OECD (2020), Draft Recommendation of the Council on Children in the Digital Environment

[17] Burns, T. and F. Gottschalk (eds.) (2019), Educating 21st Century Children: Emotional Well-being in the Digital Age, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/b7f33425-en

[18] Commission Nationale de l'Informatique et des Libertés. *(2022)*. *Online Age Verification: Balancing Effectiveness, Privacy, and User Rights*. Paris, France

[19] Australian eSafety Commissioner. *(2023)*. *Roadmap for Age Verification and Assurance Technologies: Issues, Challenges, and Recommendations*. Sydney, Australia.

The European Commission's comprehensive mapping study (2024)[20] highlighted critical limitations and ethical considerations associated with the widespread use of age verification technologies. In particular, the study stressed that age verification mechanisms should not become barriers to legitimate access, specifically cautioning against the digital exclusion of users who lack official identification, those from disadvantaged socio-economic groups, refugees, migrants, or prioritizing online anonymity. Thus, digital services must strike a careful balance between child protection goals and preserving users' fundamental rights.

The EDRi position paper (2023)[21] does not recommend the widespread adoption of biometric or document-based age verification measures, citing substantial privacy risks, heightened digital exclusion, and potential for discriminatory outcomes. Instead, EDRi advocates for the cautious, rights-oriented use of age declaration systems combined with education and digital literacy interventions, urging policymakers and platforms to critically assess the proportionality and necessity of intrusive methods.

Based on these insights, XVideos proactively gathers and analyses data from pilot studies, expert consultations, emerging research, and legal assessments. Continuous evaluation of age assurance technologies is therefore important for identifying reasonable solutions.

e) Advertising Integrity

WGCZ recognises that advertising practices present specific risks and that such risks require targeted controls. The primary objectives of the advertising integrity measures are to clearly distinguish advertisements from user-generated content, prevent harmful or illegal advertisements (such as illicit product promotions or deceptive offers), and ensure robust compliance with obligations stipulated in Article 35(1)(e) of the DSA to adapting advertising systems and adopting targeted measures. These objectives directly address identified risk scenarios in the *Advertiser and Commercial Content Integrity* category (AD_1 to AD_9). To manage these risks effectively, XVideos has implemented several dedicated measures.

AD Repository & Pre-Publication Review

WGCZ maintains and makes publicly available a comprehensive AD Repository in compliance with Article 39 of the DSA, a structured database containing all approved advertisements' details and key metadata such as advertiser identity, targeting parameters, and historical performance. Before ads are permitted to appear publicly, they undergo a pre-publication review performed by trained personnel. This review systematically verifies each advertisement's compliance with legal requirements and platform policies, including explicit checks for misleading claims, deceptive product promotions, undisclosed sponsorships, and the overall appropriateness of advertised content.

This measure directly mitigates several significant advertising-related risks identified in the risk register. Specifically, it ensures advertisements are clearly distinguished from regular content and accurately represented, thereby addressing the risk of misleading advertisements (AD_1). Proactively preventing the promotion of illegal or harmful products also directly mitigates the risk scenario involving prohibited advertisements (AD_2). The measure further reduces the risk of breaches of user trust due to undisclosed sponsorships by verifying transparency concerning sponsorship details and advertiser disclosures (AD_3). Additionally, it confirms that advertisements comply fully with applicable legal disclosure requirements, addressing the scenario where required ad disclosures might otherwise be omitted (AD_4).

Moreover, the measure explicitly addresses the scenario involving the absence of an advertisement repository (AD_7) by establishing and maintaining the structured AD Repository. Through systematic data management and periodic verification, XVideos ensures the repository remains accurate, complete, and current, addressing risks related to incomplete or inaccurate AD Repository data (AD_8). Finally, by documenting why certain advertisements are rejected or removed, the measure partially mitigates the risk associated with insufficient documentation of advertisement removals within the repository (AD_9).

[20] European Commission, Directorate-General for Communications Networks, Content and Technology, Raiz Shaffique, M., van der Hof, S., Mapping age assurance typologies and requirements – Research report – Full report, Publications Office of the European Union, 2024

[21] EDRi. (2023, October). *Online age verification and children's rights: Position paper*. European Digital Rights

By thoroughly vetting advertisements before their public release, XVideos significantly reduces the likelihood of users encountering fraudulent, deceptive, illegal, or otherwise harmful advertising content. Furthermore, the ongoing management of the AD Repository enhances transparency and accountability.

Clear Visual Marking & Informational Labels on Ads

XVideos has implemented clear visual distinctions that unmistakably identify advertisements to maintain transparency and prevent user confusion or deception (AD_1, AD_3). Every advertisement displayed on the platform is labelled prominently with either the explicit designation "Ad" or an identifiable informational ("i") icon. Clicking this icon provides detailed sponsor information, targeting rationale, and specifics about the ad's purpose and origin. This measure ensures that users can immediately differentiate sponsored content from organic, user-generated posts, substantially reducing risks associated with undisclosed sponsorships and deceptive promotions and fully addresses the requirements of Article 26 of the DSA.

User Control Over Ad Targeting

Personalised advertising involves privacy considerations and may expose users to content they find intrusive or inappropriate (AD_5, AD_6). XVideos empowers users by providing control over advertising preferences. Through transparent cookie banners and easily accessible user settings, service recipients can manage or restrict how their data is used for ad-targeting purposes, which substantially reduces privacy risks and compliance with transparency obligations according to Articles 35(1)(e) and 35(1)(i) of the DSA.

f) Governance & Compliance

The governance and compliance framework is essential for ensuring alignment with the **DSA** and other applicable regulatory obligations. As a VLOP, XVideos is required to maintain transparent, documented, and continuously monitored internal processes and governance-related measures designed to ensure that risk mitigation is implemented, overseen, resourced, and systematically reviewed for effectiveness.

Compliance Officer Nomination

WGCZ has formally designated a Compliance Officer, whose primary responsibility is to oversee the implementation of the platform's obligations under the DSA. This role is supported by allocating dedicated resources (financial and personnel) to ensure that compliance initiatives are effectively carried out across all functional areas. This step has centralized accountability for regulatory compliance, enabling a coordinated approach to risk mitigation and legal adherence. DSA-related responsibilities are not treated in isolation by individual teams but are approached systemically and consistently. This measure directly supports the platform's obligations under Article 35(1)(f) of the DSA, which requires the reinforcement of internal processes and supervision, particularly concerning the detection and mitigation of systemic risks. The appointment of a Compliance Officer, combined with adequate resourcing, is instrumental in addressing risks associated with transparency and reporting obligation (as seen in scenarios TR_1 through TR_6).

Internal DSA Audit

The platform has implemented a formal process for conducting internal audits focused on DSA compliance alongside regular updates to internal procedures. These audits assess the effectiveness of content moderation protocols, recommender system design, advertising review workflows, and mechanisms for handling user complaints. Where deficiencies are identified, procedural changes are introduced to address gaps and bring practices in line with current legal expectations or emerging risks. Also, these audits are essential for internal accountability, as they check that all teams are involved in implementing risk mitigation measures. This approach again aligns with Article 35(1)(f) of the DSA, reinforcing internal supervision and continuous evaluation. It also supports Articles 39 and 42 concerning transparency and reporting. These audits are especially relevant to risk scenarios such as failure to report on moderation actions (TR_1), incorrect reporting of active users (TR_6), recommender systems amplifying harmful content (RS_5), or ad repository containing incomplete or inaccurate data (AD_8).

Reporting to authorities

WGCZ has also implemented a structured and well-documented process for reporting to authorities, including appointing a point of contact and establishing workflows for responding to notifications from competent EU authorities. This ensures the platform can respond to such requests and fulfil its obligations under Articles 9 and 10 of the DSA. The reporting mechanism also strengthens compliance with Article 35(1)(f), which requires prompt cooperation with enforcement and regulatory bodies as part of risk mitigation. This measure is central in addressing several high-risk scenarios, particularly those involving criminal or harmful activity where law enforcement intervention is necessary. These include CSAM (IC_4), and non-consensual sexual content (IC_5) distribution. Also, this measure mitigates the risk of non-compliance with regulatory requests for platform data (TR_3) itself.

g) Research & Cooperation

The WGCZ risk management framework's critical pillars are continuous research monitoring and strategic cooperation. These efforts are designed to anticipate emerging threats, particularly those associated with illegal content, and the social consequences of adult content. Also, this approach ensures that mitigation measures are grounded in evidence and expert knowledge.

This category includes the following components:

- Research-based review of risk factors (including content harms and user vulnerabilities);
- Strategic cooperation with law enforcement;
- Ongoing partnerships with NGOs and other trusted third parties.

These activities are directly aligned with the DSA's expectations under Article 35(1)(f), which emphasize the need to reinforce the internal processes, resources, testing, documentation, or supervision of any of their activities, particularly regarding detecting systemic risk.

Research Monitoring on Violent Content and CSAM

As part of its commitment to evidence-based risk governance, the platform has instituted a structured system for continuously reviewing academic research, expert studies, and NGO publications related to CSAM, NCII, and gender-based violence. These reviews inform the platform's risk assessment, as well as design and implementation of specific mitigation measures.

(i) Review of Research on CSAM

Recent studies confirm that CSAM is overwhelmingly distributed via peer-to-peer (P2P) platforms, encrypted messaging apps, and darknet marketplaces, rather than through mainstream, regulated adult content services. Key findings include:

- The UNODC's Study on the Effects of New Information Technologies on the Abuse and Exploitation of Children (2015)[22] and the ECPAT (2018) report[23] emphasizes that CSAM is primarily exchanged in non-commercial, unmoderated environments such as torrent systems, hidden services, or private communication platforms.
- Huikuri (2023)[24] highlights the importance of anonymity as a key enabler for CSAM users, suggesting that platforms with robust moderation, identity controls, and audit trails are less attractive for CSAM dissemination.

These insights validate the platform's long-standing position that although CSAM remains a very high inherent risk (due to the extreme harm involved), its design and controls reduce its attractiveness as a vector for such content. Among other measures, these factors have contributed to lowering the inherent risk from "High" to residual risk "Medium" for risk that users upload and share CSAM (IC_4).

[22] United Nations Office on Drugs and Crime. Study on the Effects of New Information Technologies on the Abuse and Exploitation of Children. Vienna, 2015

[23] ECPAT International. (2018). Trends in Online Child Sexual abuse material. Bangkok: ECPAT International (April 2018)

[24] Huikuri, S. (2023). Users of Online Child Sexual Abuse material. Journal of Police and Criminal Psychology, 38(4), 904–913.

Importantly, these same research findings also inform the risk associated with sharing prohibited material, such as violent or exploitative content involving minors (PM_8). While this risk scenario focuses more broadly on non-sexual but still harmful material involving minors (e.g., depictions of physical abuse, coercion, bullying, or humiliation), the distribution patterns identified in CSAM studies are similarly applicable. Studies indicate that exploitative content involving minors (whether sexual or non-sexual) is overwhelmingly shared on platforms with limited oversight and accountability. Therefore, evidence from CSAM-focused research is directly relevant to PM_8, reinforcing the conclusion that mainstream adult platforms with structured content review processes and low levels of anonymity are not commonly used for the dissemination of such content.

(ii) Research Monitoring on Violent Content and Gender-Based Violence

This measure form is relevant to scenarios involving the dissemination or normalisation of violent and non-consensual content. The targeted risk scenarios include the uploading and sharing of material depicting non-consensual sexual acts and gender-based violence. To ensure that the platform's content moderation policies remain responsive to current societal concerns, WGCZ has undertaken a series of ongoing reviews of academic literature and policy reports. While public debate often assumes a straightforward causal relationship between exposure to violent or aggressive pornographic material and real-life behaviours, the research consistently shows a more complex picture, in particular:

- Lim et al. (2015)[25] examined the impact of pornography on gender-based violence, sexual health, and well-being. The study identified that while violent content is prevalent in some pornography, there is no consistent evidence that it directly causes real-world aggression. It highlighted methodological limitations in existing literature and noted the importance of contextual and personal factors in determining any impact.

- Wright et al. (2015)[26] conducted a meta-analysis on pornography consumption and sexual aggression. They found statistically significant but modest correlations between consumption of pornography and attitudes supportive of sexual violence, though not all consumers displayed these tendencies. They emphasized that other factors, such as personality traits and social norms, often moderated these outcomes.

- Mestre-Bach et al. (2023)[27] reviewed 20 years of research on pornography and violence. Their findings suggested that while some studies observed associations between violent pornography and intimate partner sexual assault, others found no direct link, mainly when accounting for individual predispositions (e.g., antisocial traits, sensation seeking, or belief in rape myths).

- The World Health Organization (2024)[28], in its global fact sheet on violence against women, identified key contributing factors to sexual and intimate partner violence (including gender inequality, substance abuse, and early exposure to violence) but did not recognize pornography as a primary risk factor on its own.

Cumulative evidence and the strength of the platform's existing controls (see point b): Content Moderation in this section) have had a positive impact on the residual risk levels associated with the following scenarios: non-consensual sexual acts (IC_5), material containing gender-based violence (IC_6), and revenge porn or harassment (IC_7). Studies related to gender-based violence, harmful stereotypes, and the algorithmic amplification of violence also contributed to lowering the risk levels for scenarios GB_1 through GB_5. Nonetheless, this measure remains subject to continuous review, particularly considering emerging technologies (e.g., generative AI or synthetic abuse content) and ongoing social developments.

Review of Research on Adult Content and Public Health

This measure supports WGCZ's commitment to fostering user well-being, mitigating risks to vulnerable populations and promoting the responsible use of adult content. WGCZ reviewed academic research concerning adult content's psychological, behavioural, and social effects. Key studies and institutional reports informing this review include:

[25] Lim, M. S., Carrotte, E. R., & Hellard, M. E. (2015). The impact of pornography on gender-based violence, sexual health and well-being: What do we know? Journal of Epidemiology and Community Health, 70(1), 3–5

[26] Wright, P. J., Tokunaga, R. S., & Kraus, A. (2015). A meta-analysis of pornography consumption and actual acts of sexual aggression in general population studies. Journal of Communication, 66(1), 183–205

[27] Mestre-Bach, G., Villena-Moya, A., & Chiclana-Actis, C. (2023). Pornography use and violence: A systematic review of the last 20 years. Trauma, Violence, & Abuse, 25(2), 1088–1112

[28] World Health Organization: WHO. (2024, March 25). Violence against women

- Owens et al. (2012)[29] conducted a meta-analysis of pornography's impact on adolescents. They found inconclusive evidence regarding harmful behavioural outcomes, highlighting substantial methodological inconsistencies across studies. Some research suggests the potential for increased risk-taking, while others see no correlation.
- Sinković et al. (2013)[30] found no direct association between the frequency of pornography use and risky sexual behaviour, mainly when accounting for early exposure and personality traits such as sexual sensation seeking.
- Bőthe et al. (2020)[31] demonstrated that high-frequency pornography use alone does not predict problematic behaviour. Psychological distress, such as depression or anxiety, is more likely to occur when users experience moral incongruence with their consumption, not from the content per se.
- Willoughby et al. (2014)[32] examined the relationship between pornography use and social behaviour in young adults. They found that frequent users were not necessarily more socially isolated or dysfunctional and that personal values and relational context play a key role in how content is experienced.
- Perry (2017, 2020)[33] explored how pornography affects romantic relationships. He found that mutual and consensual use often has neutral or positive effects on relationship satisfaction. Issues tend to arise when one partner consumes pornography privately or excessively without the other's knowledge or comfort, leading to relationship strain.
- Stefanska et al. (2022)[34] investigated the prevalence of atypical or potentially harmful sexual fantasies. They concluded that many pornography users do not exhibit dangerous or deviant tendencies and that interest in illegal or violent content is statistically rare.

These findings suggest that the public health risks associated with adult content are nuanced and context dependent. While there is legitimate concern about early exposure among minors or excessive, compulsive use in adults, current evidence does not support blanket assumptions of harm across all user groups. Instead, risk is shaped by individual characteristics, social context, and how content is consumed.

The limited or inconsistent causal evidence linking pornography consumption to negative mental health outcomes, sexual aggression, or social dysfunction has influenced the assessment of risks related to the protection of minors. These include age verification weaknesses and inadvertent exposure, which are maintained at a Medium residual risk level, reflecting the limitations of existing technologies and the need for continued parental and systemic involvement (PM_1 through PM_9). This also affects the evaluation of risks related to mental health, addictive behaviours, and sexual wellness education, which are rated as medium or low, depending on the specific scenario and the user segment affected (PH_1 through PH_6).

Cooperation with Law Enforcement Agencies

The platform maintains formalised, operational channels for cooperation with law enforcement authorities at the national and international levels. This cooperation is grounded in established procedures and technical capabilities, enabling swift and lawful responses to illegal or high-risk content incidents. Key aspects of this cooperation include:

[29] Owens, Eric W., et al. “The impact of internet pornography on adolescents: A review of the research.” Sexual Addiction & Compulsivity, vol. 19, no. 1–2, Jan. 2012, pp. 99–122

[30] Sinković, M., Štulhofer, A., & Božić, J. (2013). Revisiting the association between pornography use and risky sexual behaviors: The role of early exposure to pornography and sexual sensation seeking. Journal of Sex Research, 50(7), 633–641

[31] Bőthe, B., Tóth-Király, I., Potenza, M. N., Orosz, G., & Demetrovics, Z. (2020). High-frequency pornography use may not always be problematic. The Journal of Sexual Medicine, 17(4), 793–811.

[32] Willoughby, B. J., Carroll, J. S., Nelson, L. J., & Padilla-Walker, L. M. (2014). Associations between relational sexual behaviour, pornography use, and pornography acceptance among US college students. Culture, Health & Sexuality, 16(9), 1052–1069

[33] Perry, Samuel L. “Pornography and relationship quality: Establishing the dominant pattern by examining pornography use and 31 measures of relationship quality in 30 national surveys.” Archives of Sexual Behavior, vol. 49, no. 4, 2 Jan. 2020, pp. 1199–1213

[34] Stefanska, E. B., Longpré, N., & Rogerson, H. (2022). Relationship between atypical sexual fantasies, behavior, and pornography consumption. CrimRxiv

- Designated points of contact: WGCZ has designated contact points to handle removal orders, data preservation requests, and investigative inquiries submitted by competent authorities.
- Manual reporting protocols: Moderation teams are trained to escalate cases involving CSAM, human trafficking, and terrorist content in the event such risks materialise. Verified cases, including IP addresses, metadata, upload logs, and related account activity, are manually reported to ensure that law enforcement agencies receive actionable intelligence.
- Data preservation and legal compliance: The platform follows strict procedures to comply with obligations under EU and national data retention laws. This includes preserving digital evidence when legally mandated and cooperating in criminal proceedings involving vulnerable victims.

This cooperation framework is particularly relevant to high-severity risk scenarios, such as users uploading and sharing CSAM (IC_4), as well as public security threats, including the distribution of hate-driven or violent content that threatens public safety (PS_3). In each of these cases, the availability and demonstrated use of rapid communication channels with law enforcement has directly contributed to a downward adjustment of the residual risk level. The platform's ability to escalate and report serious threats in real time significantly mitigates potential systemic harms and aligns with regulatory expectations.

In addition, as described in Section 4.2: Overall Approach, the platform is engaged in ongoing dialogue with Interpol, including preparations for the adoption of the "Worst Of" List (IWOL) – a centralized database of the most severe CSAM-hosting domains. Further engagement with Interpol's child safety data analytics team is underway, ensuring the platform remains aligned with international enforcement efforts.

Cooperation with NGOs and Trusted Third Parties

As introduced in Section 4.2: Overall Approach, WGCZ collaborates with various NGOs and expert third parties to support the identification of emerging risks. This section expands on that foundation by describing how these partnerships function as operational mitigation measures that directly reduce residual risk in specific scenarios.

Beyond regulatory enforcement, the platform cooperates actively with NGOs, civil society actors, and specialized third-party groups. These collaborations provide essential expertise, particularly in areas such as NCII, youth exploitation, and gender-based violence. By integrating these external perspectives into its governance structure, the platform can respond more effectively to complex and high-sensitivity harms.

Among the key partnerships is OffLimits (Netherlands), which has supported WGCZ in implementing a CSAM hash-matching system and preparing to introduce deterrence messaging and helpline services in collaboration with the Stop It Now programme. This partnership also contributes to shaping internal policies on victim-sensitive moderation practices. The Czech Safer Internet Centre further advises the CSAM and NCII detection platform, focusing on early-stage intervention models that align with national regulatory expectations. In the United Kingdom, the platform is working with SWGfL to integrate the StopNCII tool, which enables victims to hash and pre-emptively flag intimate images at risk of being shared without requiring the image to be uploaded. This tool enhances detection accuracy while preserving user anonymity.

Further engagement with the Lucy Faithfull Foundation in the UK is ongoing to explore the possible adoption of Project Intercept and the ReThink Chatbot. These tools are designed to redirect users searching for CSAM toward appropriate support services, targeting demand-side behaviour and reinforcing deterrence messaging. At the international level, the platform collaborates with InHope, the global network of internet hotlines dedicated to combatting CSAM. This cooperation supports harmonized reporting procedures, shared classification standards, and the broader development of global mitigation frameworks.

These partnerships play an instrumental role in addressing several key risk scenarios. Specifically, they enhance the platform's ability to mitigate risks related to the uploading or sharing of non-consensual sexual acts (IC_5), the distribution of content intended to harass or harm individuals (GB_2), inadequate reporting and response mechanisms for victims of abuse (GB_4), and the sharing of prohibited exploitative content involving minors (PM_8).

As also noted in Section 4.2, these efforts are complemented by the platform's participation in the development of international standards on age verification (through its role in a Canadian expert working group), as well as its ongoing engagement in multilateral forums and industry roundtables, including a United Nations co-hosted conference in New York.

5.4. Residual Risk Assessment

Residual risk refers to the level of risk that remains after the implementation of controls. While inherent risk provides insight into baseline exposure assuming no mitigation, residual risk evaluates the effectiveness of WGCZ's operational and compliance environment in reducing that exposure. By comparing inherent and residual risk levels, the platform measures risk reduction, identifies persisting vulnerabilities, and prioritizes areas for continuous monitoring.

Unlike inherent risk, residual risk is evaluated qualitatively rather than quantitatively. Applying a qualitative approach to residual risk stems from the inherent interdependence of the controls and risk scenarios. Although control measures have certain effectiveness ratings in the Mitigation Measure Register, their de facto mitigating effectiveness is contingent upon the specific characteristics of the risk scenario and the broader operational context in which they are deployed.

For example, the "notice and action" reporting system, allowing users to flag illegal or harmful content, is highly effective for clear-cut violations, such as uploading violent or explicit non-consensual content. These are easily identifiable and generate prompt reports, triggering rapid removal. On the other hand, it could be less effective for algorithmic amplification of misinformation or gender-based stereotyping, where users may not perceive content as harmful or may themselves support or engage with it. So, the same reporting system reduces residual risk significantly in one case but offers limited mitigation in another.

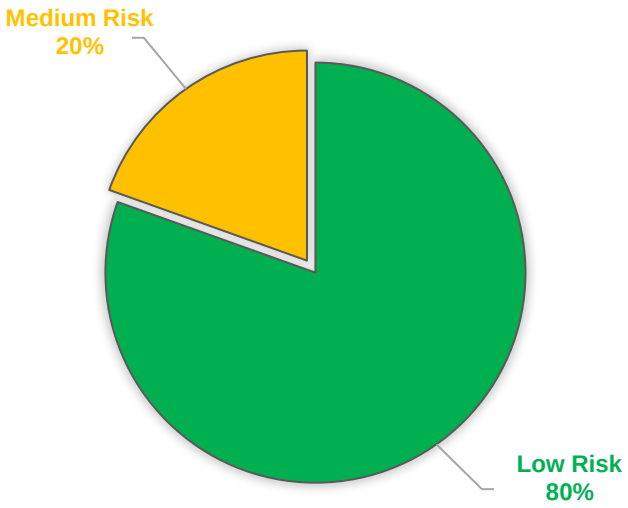
In addition to scenario-specific differences, the effectiveness of a measure is also shaped by its interaction with other measures, it means the same control may give varying results depending on the configuration of the overall mitigation strategy. For example, a high-level content recommendation algorithm designed to reduce exposure to harmful material may demonstrate limited impact if not aligned with user-controlled settings. When such an algorithm is augmented with user-facing tools, its capacity to mitigate risks increases measurably. This highlights the importance of not assessing controls in isolation but as part of a system where combinatory effects can significantly influence residual risk.

In the current risk assessment, residual risks were classified into two levels:

Low Risk (Green): The measures are sufficient, and the remaining exposure is minimal. These risks are well-contained under current operations but should be periodically reviewed to ensure ongoing effectiveness.

Medium Risk (Yellow): Relevant controls have been implemented; however, a meaningful level of residual exposure persists due to limitations (e.g. control coverage, scalability, responsiveness, or specificity to the risk scenario). These risks may not present immediate threats, but they are more systemic or have the potential to escalate if left unmanaged. Also, some medium residual risks stem from scenarios that were assessed as high inherent risks. In such cases, applying multiple mitigation measures has significantly reduced the overall risk, but not to a low level. These scenarios reflect partial success in risk mitigation but also highlight the need for ongoing monitoring.

Figure 5: Results of residual risk assessment



Source: WGCZ Risk Register, RA DASHBOARD

The residual risk assessment revealed a significant overall reduction in risk exposure following mitigation. Of the 87 assessed scenarios, 70 (80%) were rated as low residual risk, indicating that existing controls are effective.

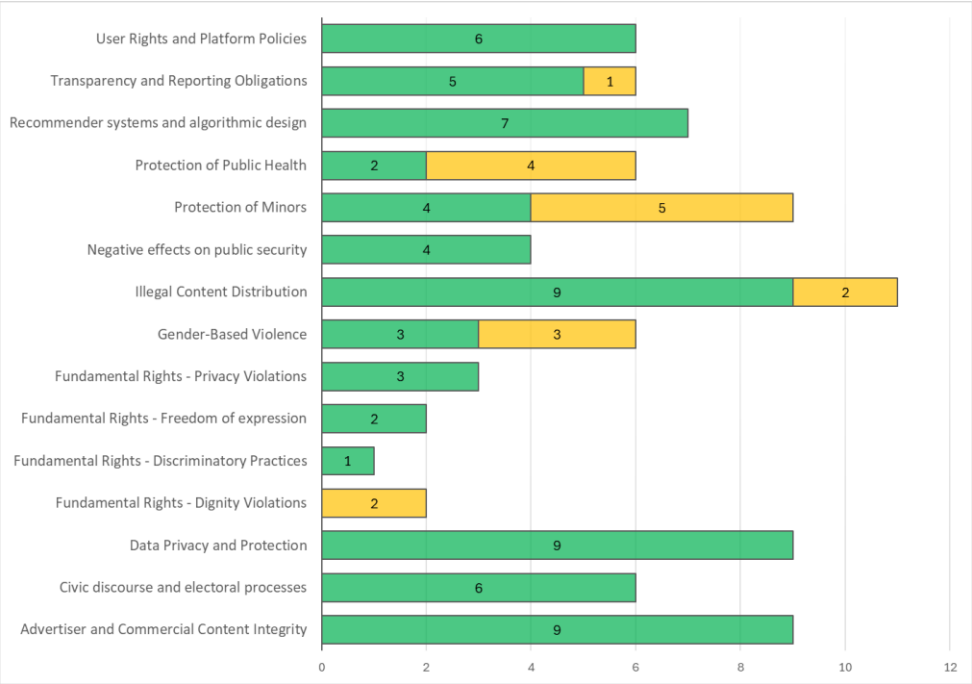
Instead, 17 scenarios (20%) were evaluated as medium residual risk, which is distinguished into two types:

Persistent medium risks (Inherent Risk = Medium → Residual Risk = Medium): These risks remain at a medium level even after applying all currently available mitigation measures. This is **not due to a failure of controls** but rather a **reflection of the intrinsic characteristics of the platform's services**. For example, several risks within the *Protection of Public Health* category, such as the potential for users to develop unhealthy behavioural patterns, addiction, or distorted perceptions of body image, remain inherently tied to the nature of adult content platforms. Even with robust content warnings, health disclaimers, and educational partnerships, the residual exposure cannot be eliminated. Specific societal or psychological impacts may occur regardless of platform intent or moderation practices. In such cases, the strategic focus is not on complete elimination, which would be unfeasible, but on active monitoring, research-based interventions, and user awareness. This reasoning applies similarly to gender-based violence scenarios where harmful stereotypes are normalised and scenarios involving exposure of minors, where certain systemic and behavioural factors can only be reduced but not entirely eradicated by technical measures. Therefore, these medium residual risks are best described as service-adjacent risks, which means risks are managed responsibly but cannot be fully separated from the type of content or services offered.

De-escalated high risks (Inherent Risk = High → Residual Risk = Medium): The second group of medium residual risks consists of scenarios initially classified as high inherent risk but have been reduced to medium through specific mitigation measures applied. The distribution of CSAM or non-consensual sexual content is among the most severe threats to user safety and XVideos integrity. These were assigned high inherent scores, but through a combination of different measures, the likelihood of these scenarios has been significantly reduced. Despite de-escalation, a medium residual risk remains, as the adversarial nature of the threat means there is always a possibility of new abuse vectors emerging. Thus, **while the risk is no longer at the highest level, it still requires active oversight.**

Notably, no scenarios were evaluated as high residual risk, signifying that even the most severe inherent risks, such as those related to child safety, privacy violations, and abusive content, have been significantly mitigated through WGCZ's layered control framework. It also indicates that XVideos is well-positioned from a regulatory compliance standpoint and has embedded risk mitigation into its operational model in an adaptable way.

Figure 6: Results of residual risk assessment by category



Source: WGCZ Risk Register, RA DASHBOARD

WGCZ also analysed residual risks across 14 predefined categories to get insight into the concentration of risk exposure and the effectiveness of implemented controls from the other point of view. The analysis reveals that most risk categories have been successfully reduced to exclusively low residual risk (8 from 14 categories). Specific categories still contain a notable proportion of medium-risk scenarios because of the complexity of the threat landscape and the limits of mitigation strategies given the nature of the platform's services, in particular:

- **Transparency and Reporting Obligations:** This category includes 6 risk scenarios, of which 5 have been mitigated to a low residual level. One scenario related to reporting the average monthly number of active service recipients (TR_6) remains at medium residual risk. This scenario was initially rated as medium inherent risk and retains this classification post-mitigation due to ongoing uncertainty about how the risk must be addressed in practice. Although several mitigation measures have been implemented, including appointing a Compliance Officer, setting up an internal DSA audit process, and procedural updates, residual risk persists. This is mainly attributable to the absence of a harmonised methodology or official guidance from the European Commission on calculating and reporting this metric. Without such clarity, XVideos is left to interpret requirements individually, which increases the likelihood of reporting misinterpretations or future regulatory corrections. This risk will remain at a medium level until formal, standardised guidance is issued and internal systems can be aligned accordingly.

- **Protection of Public Health:** This category includes 6 risk scenarios, of which 4 remain at medium residual risk (PH_1, PH_2, PH_4, PH_6) due to the platform's inherent content type and the complex nature of the risks. These scenarios concern issues such as the contribution to addictive behavior patterns, the lack of adequate sexual education, the promotion of unattainable or harmful body image ideals, and the potential exposure to content that may encourage unhealthy or self-harming behaviors. Each of these was initially assessed as a medium inherent risk, and the residual risk remains unchanged following mitigation, not due to lack of effort but because of the indirect and societal nature of these risks. While WGCZ has implemented important safeguards, such as displaying warnings on sensitive content, conducting reviews of public health research, cooperating with NGOs, and restricting specific types of depictions, XVideos faces structural limitations in its ability to control how users experience or internalise the material. The psychological and behavioural impact of adult content can vary widely based on the user's age, mental health, and social context, and this variability makes risk elimination infeasible. Moreover, some of these harms occur gradually and are influenced by broader cultural and media consumption trends. Xvideos cannot replace comprehensive sexual education, nor can it fully counteract unrealistic body image standards that are reinforced across multiple online environments. Despite targeted mitigations, the potential for

indirect but meaningful harm remains. The medium residual classification for these scenarios reflects a realistic and responsible assessment of what is within the platform's control. WGCZ's strategy in this area must remain focused on ongoing risk monitoring, and responsive adaptation to new research. Until more explicit public health frameworks are developed about adult content, these risks will require balancing harm reduction and users' freedom of access.

- **Protection of Minors:** This category includes 9 risk scenarios, of which 5 remain at medium residual risk (PM_1, PM_2, PM_3, PM_4, PM_9), primarily due to the inherent complexity of preventing minors from accessing adult content. These risks involve issues such as by-passing age assurance mechanisms, circumvention through VPNs or shared devices, and the potential mental and physical harm caused by exposure to explicit material. All five were initially rated as high or medium inherent risks, reflecting both high likelihood and impact despite implementing robust controls, including age confirmation systems, RTA labelling, warnings, page blurring, 2FA, and ongoing assessment of verification tools. Residual exposure remains due to behavioural circumvention, limited control over off-platform factors, and technological limitations. These scenarios require continued refinement of access controls and monitoring of age-verification tools trends.

- **Illegal Content Distribution:** This category contains a relatively high number of risk scenarios (11 total), most of which have been mitigated to a low residual level (9 risk scenarios). However, two scenarios remain at medium residual risk: the upload and sharing of CSAM (IC_4) and the distribution of non-consensual sexual acts (IC_5). These were initially rated as high inherent risks due to the high value of potential impact. Despite extensive mitigation measures (e.g. combination of human and automated review, notice and action mechanism, review of research on CSAM impact, cooperation with law enforcement agencies and NGOs, ToS specific sections), these threats persist at a medium level because of the deliberate and evolving strategies employed by malicious actors, and additional measures should be applied.

- **Gender-Based Violence:** Of the 6 risk scenarios in this category, 3 remain at medium residual risk (GB_1, GB_2, GB_5). The medium-risk scenarios relate to the upload of violent content, cyberbullying, and the spread of harmful gender-based stereotypes. All three were originally rated as high or moderate inherent risks due to the societal harm, reputational exposure, and user safety concerns they represent. Despite a wide range of mitigation measures, including automated tools for violence detection (Google SafetyNet API, Hive classification), manual moderation, multi-layered content review, user reporting, and active cooperation with law enforcement, these risks persist due to the difficulty of identifying subtle or context-dependent content, as well as challenges in detecting intent and nuance in user interactions.

Category with Only Medium Residual Risk

Fundamental Rights – Dignity Violations category contains 2 risk scenarios, both remaining at medium residual risk. The scenarios concern the dissemination of humiliating or degrading content and the creation and distribution of deepfake pornography, both of which were initially classified as high inherent risks due to the severe psychological and legal major impact they can cause. These risks threaten individuals' right to dignity and privacy, particularly when manipulated or non-consensual material is published to shame or harm.

Extensive mitigation measures have been applied, including deploying advanced AI scanning tools to detect manipulated content, using notice-and-action systems supported by trained human moderators, integrating trusted flagger networks and community reporting mechanisms, and active partnerships with NGOs and law enforcement agencies. In addition, WGCZ continues to invest in research on generative AI technologies to understand emerging threats better and has adopted strict ToS provisions that explicitly prohibit such content (ToS - violence related prohibitions and point 7 - User submissions).

Despite these efforts, the highly sensitive nature of these risks and the technological challenges associated with detecting deepfakes, and contextually abusive content mean that a complete reduction to low residual risk is not currently feasible. These scenarios remain deeply intertwined with broader societal and technological dynamics, which limits the platform's ability to resolve them in isolation. Therefore, the medium residual risk classification reflects a realistic approach, acknowledging the limits of current technology and the gravity of the potential impact.

6. Risk Management Strategy

6.1. Risk Management Strategy

In accordance with the defined risk assessment methodology, WGCZ decided to apply the following risk management strategies based on the residual risk scores:

- For risk scenarios with a residual score of **Low**, the applied strategy is **Accept**; this signifies that, after considering the mapped controls and mitigation measures, the remaining level of risk exposure is considered acceptable. It's important to clarify that "Accept" does not equate to inaction. Instead, it reflects a well-informed decision based on a robust control environment. This environment is designed to continuously monitor and evaluate the effectiveness of these controls. Should any weaknesses or changes in the risk landscape be identified, the control environment would trigger a prompt response. This could involve implementing additional controls, enhancing existing controls, or even reassessing the risk strategy if necessary. In essence, by accepting low residual risk, we acknowledge that the current controls are sufficient, but we remain vigilant and prepared to adapt should circumstances change. This ongoing monitoring and management process ensures the continued effectiveness of our controls and the continued acceptability of our residual risk exposure.

- For risks with a residual score of **Medium**, the applied strategy is **Reduce** (Mitigate); This strategy reflects the recognition that current mitigation efforts have lowered the inherent risk but that further actions are still necessary to reduce the risk to manage it responsibly, because complete elimination is not feasible. The Reduce strategy involves the design, development, implementation, and continuous management of additional or enhanced measures.

6.2. Status of Action Plan Implementation

This section provides an overview of the implementation status of action items identified in the previous Risk Assessment Report (2024) and reflects the state of progress as of April 2025. These action items represent targeted responses to identified systemic and content-related risks and are integral to the platform's ongoing risk mitigation framework.

Collaboration with NGOs

Engagement with NGOs and expert third parties is ongoing to support the identification and mitigation of risks associated with CSAM, NCII, and other harms affecting users and the platform. These collaborations serve as a complementary input to the internal risk assessment and mitigation processes.

Cooperation continues with OffLimits (Netherlands) on implementing a CSAM hash list and preparatory work to introduce deterrence messaging and helpline services from the Stop It Now programme. These efforts aim to address the dissemination of CSAM and offender behaviour and prevention strategies. The relationship with InHope, the global network of internet hotlines for reporting CSAM, remains active and contributes to the development of relevant mitigation measures cooperation with the Czech Safer Internet Centre remains in place, bringing significant expertise in addressing issues related to NCII and CSAM.

Recently, participation in a UN co-hosted conference in New York and related industry roundtables has contributed to a broader understanding of emerging risks and best practices.

As of April 2025, dialogue with the Lucy Faithfull Foundation (United Kingdom) is ongoing regarding the possible adoption of Project Intercept and the reThink Chatbot, which directs users searching for CSAM to support services. Discussions are also underway with SWGfL (United Kingdom) to implement the StopNCII tool, which allows adult victims to hash and flag intimate images at risk of being shared online. This supports more efficient detection and takedown processes while protecting user anonymity. Furthermore, engagement with Interpol is ongoing, including reviewing a proposed partnership to adopt the "Worst Of" list (IWOL), a database of domains publishing the most severe CSAM. The partnership agreement is currently under review. Engagement with Interpol's child safety data analytics team is also ongoing to explore further areas of cooperation.

Collaboration with NGOs focusing on age verification

The platform's collaboration with the Digital Governance Standards Institute of Canada on developing a national age assurance standard continues as an integral part of its ongoing efforts to support international standard-setting in this area. While initial contributions focused on shaping the structural foundations of the standard, the platform's involvement has further intensified, with active participation in drafting activities through WGCZ's Director, Regulation & Safety.

As the standard nears finalisation and prepares for public consultation, the platform's contribution focuses on ensuring that any adopted frameworks strike an appropriate balance between adequate age assurance and user privacy protection, technical feasibility, and proportionality.

Enhanced Disclaimers

No updates have been made to the disclaimers beyond the internal assessment conducted earlier this year. While the disclaimer remains concise, it clearly states that the site is not intended for minors. The Legal team concluded that the current wording sufficiently fulfils its purpose, and no further modifications are planned. This action item is, therefore considered complete unless future reassessment indicates the need for change.

Age Verification Tools Development Monitoring

In August 2024, efforts were initiated to enhance the monitoring of age verification tools. A comprehensive analysis, "Age Verification Tools: Analysis and Reference Review," was conducted to critically examine the requirements for implementing age verification tools, particularly considering the Radio and Television Broadcasting Council's (RRTV) interpretation of relevant legislative provisions, as already mentioned in section Legislative and research monitoring of the Section 4.2: Overall Approach. The research highlighted key concerns regarding the adequacy of commonly used age verification methods, such as qualified disclaimers, which RRTV deems insufficient for effectively verifying users' ages. The analysis also explored the complexities of age verification, addressing technological limitations, potential risks to user privacy, and the broader regulatory implications. It provided a well-rounded overview of current research and identified the need for more robust and privacy-conscious verification mechanisms.

As of April 2025, no further significant steps have been taken regarding implementing new age verification monitoring measures. However, this action plan remains relevant for future efforts as the challenges related to age verification persist and regulatory expectations continue to evolve.

Awareness Sub-Site Development

There has been no progress on the development of the dedicated awareness sub-site. Despite the continued relevance of the action item, particularly in relation to raising awareness about the inconclusively studies association of adult content with negative outcomes in mental health, compulsive use, and sexual wellness, no concrete steps have been taken toward its implementation.

Comprehensive Review of Moderation Processes and Procedures

A comprehensive review of moderation processes and procedures is currently underway in line with the principle of continuous improvement and in accordance with best practices. Following the revised start date of December 2024, the review is progressing according to an updated timeline, with completion scheduled for April 30, 2025. This extension allows for a more in-depth review process, including complete documentation and alignment across all involved teams.

The information collection phase has already been completed. Descriptions of the moderation process were gathered from both the content moderation team (operational perspective) and the tech team (technical perspective) and have been compared. Current efforts are focused on harmonising definitions and process steps.

Also, the work is now centred on mapping the full moderation lifecycle, with a clear distinction between human-led and automated components. This is being validated directly by content moderators to reflect operational realities accurately. Additionally, an updated content moderation process chart is being finalised. It will encompass all stages of the content lifecycle—including new content uploads, new and existing content reviews, and escalation mechanisms in cases involving illegal or non-compliant material.

The process chart will accompany a detailed working paper, providing comprehensive documentation of each step. These efforts reflect the platform's commitment to strengthening governance structures through ongoing monitoring and systematic review.

DSA Compliance Management System Implementation

In April 2024, the Methodology for DSA Compliance document was developed to establish a structured and systematic framework for ensuring ongoing alignment with the DSA. This document outlines the processes and methodologies used to identify, prioritise, and meet DSA compliance obligations.

In line with the principle of continuous improvement and as part of the regular compliance cycle, further enhancements to the Compliance Management System have been underway as of April 2025. Current activities focus on updating key compliance documents, including the Compliance Policy, Compliance Statute, and Risk Assessment Guidelines. These updates reflect evolving regulatory expectations, incorporate internal developments, and strengthen procedural clarity, ensuring the system remains dynamic and responsive.

External Audit of CMS

In keeping with the platform's commitment to continuous improvement and compliance with Article 37 of the DSA, an external audit of the Compliance Management System is currently ongoing, conducted by CERTICOM s.r.o. in line with the revised deadline of April 23, 2025. A draft audit report has already been obtained, and the final Audit Report is scheduled for publication on July 20, 2025.

Internal Compliance Audit

The DSA requirements analysis has already been completed, and the internal compliance audit is progressing as planned. By comparing the platform's current processes against applicable DSA provisions, this analysis has helped inform and enhance ongoing internal processes.

The internal audit of the moderation process, initiated in December 2024, is also moving forward on schedule. Process documentation review and update have been finalized, and efforts to standardize workflows are underway. Detailed mapping of the moderation lifecycle is ongoing, with work focused on finalizing an updated moderation process chart and a comprehensive working paper outlining roles, responsibilities, and interdependencies.

Concurrently, the internal audit of the recommender system is in the analysis phase. This includes information-gathering sessions and a questionnaire on ranking criteria and system behaviour. A structured chart and accompanying documentation are currently being prepared.

Development of Standardised Reporting Procedures

Progress is underway to align reporting procedures with the DSA Implementing Regulation to meet the reporting requirements by the established deadline of July 1, 2025. Cooperation with the Tech Team has commenced, focusing on optimizing data collection methods to support structured and compliant reporting. Present efforts are directed toward mapping content categories specified in the regulation to ensure that the data extraction system is fully aligned with reporting obligations. The resulting outputs will also contribute to the DSA Transparency Database.

6.3. Action plan for each medium – high risk

Following the identification and assessment of risks, the following table outlines the action plan for mitigating those risks. The action plan details the specific controls and measures currently in place, as well as additional mitigation strategies planned for implementation. The focus is on addressing risks with the residual medium risk as identified in the Section 5.4: Residual Risk Assessment.

Table 9: Medium residual risk scenarios and risk management roadmap

Risk ID	Risk Scenario	Measures in place	Additional measures to implement
IC_4	Users uploading and sharing CSAM (Child Sexual Abuse Material)	- Automated tools: AI scanning, Thorn Safer, Google SafetyNet API, Hive classification (all), Vercury fingerprint database - Manual moderation process (control of participant age and consent, detecting illegal content, and copywriting material) - Multi-layered content review (combination of human and automated review) - Moderation team training - User reports (notice and action mechanism) - Terms of Service - Point 3 - Child and non-consensual content, Point 7 - User submissions, Point 8 - Content moderation - Review of research on CSAM impact - Cooperation with law enforcement agencies - Cooperation with NGOs - Reporting to Authorities	- Maintain and expand collaborative studies on CSAM offender behaviours with Lucy Faithfull Foundation and law enforcement to stay updated on emerging trends; - Incorporate Interpol's "Worst Of" list (IWOL) into the existing fingerprint repository;
IC_5	Users uploading and sharing material depicting non-consensual sexual acts	- Automated tools: Thorn Safer, Google SafetyNet API, Hive classification (all), Vercury fingerprint database - Manual moderation process (control of participant age and consent, detecting illegal content, and copywriting material) - Multi-layered content review (combination of human and automated review) - Moderation team training - User reports (notice and action mechanism) - Terms of Service - Point 3 - Child and non-consensual content, Point 7 - User submissions, Point 8 - Content moderation - Cooperation with NGOs - Cooperation with law enforcement agencies - Reporting to Authorities	- Continue partnerships with NGOs to deepen understanding of non-consensual content patterns and psychological harm; - Collaborate with other platforms for real-time removal of known NCII;
GB_1	Users uploading and sharing material containing non-simulated violence, including gender-based violence	- Automated tools: Google SafetyNet API, Hive classification (Violence), Thorn Safer - Manual moderation process (control of participant age and consent, detecting illegal content, and copywriting material) - Multi-layered content review (combination of human and automated review) - Moderation team training - User reports (notice and action mechanism) - Cooperation with law enforcement agencies - Terms of Service - violence related prohibitions + Point 7 - User submissions, Point 8 - Content moderation	- Continue analysing global studies on patterns of violent adult content to refine detection thresholds and moderator guidelines - Formalize MoUs with victim support organizations for ongoing feedback on detection efficacy

GB_2	Users uploading and sharing material with the aim of harming certain persons, such as "revenge porn" or "cyberbullying"	- Automated tools: Google SafetyNet API, Hive classification (Violence), Thorn Safer - Manual moderation process (control of participant age and consent, detecting illegal content, and copywriting material) - Multi-layered content review (combination of human and automated review) - Moderation team training - User reports (notice and action mechanism) - Cooperation with law enforcement agencies - Terms of Service - violence related prohibitions + Point 7 - User submissions, Point 8 - Content moderation	- Keep assessing the psychological and societal impacts of revenge porn, in partnership with crisis helplines - Partnering with GBV-focused NGO
GB_5	Spreading harmful stereotypes that reinforce gender inequality	- Research monitoring on violent content impacts (gender-based violence, non-consensual sexual act, physical aggression etc.) - Automated tools: Google SafetyNet API, Hive classification (Violence), Thorn Safer - Manual moderation process (control of participant age and consent, detecting illegal content, and copywriting material) - Multi-layered content review (combination of human and automated review) - Moderation team training - User reports (notice and action mechanism) - Cooperation with law enforcement agencies - Terms of Service - violence related prohibitions + Point 7 - User submissions, Point 8 - Content moderation	- Ongoing academic research partnership to systematically evaluate the effect of adult content on gender stereotypes - Engage with gender-equality NGOs to refine policy and moderation practices
PM_1	Inadequate access verification systems allowing minors access to adult content	- Age confirmation - Ongoing assessment of age verification measures - Warnings about adult content - Adequate content labelling (RTA label) - Page blurring - Terms of Service - Point 2 - Access - Parental guidance and control instructions	- Maintain regular monitoring of age-assurance tools and legislative developments, contributing to standard-setting dialogues (e.g., Canadian Age Assurance Standard); - Strengthen involvement in working groups and public policy forums dedicated to child online safety and data protection;
PM_2	Minors access adult content on devices left logged in by adults, due to lack of session timeouts or secondary content access verification	- Parental guidance and control instructions - Adequate content labelling (RTA label) - Terms of Service - Point 2 - Access - Warnings about adult content - Age confirmation - Page blurring	- Continue user education on securing personal devices, log-out steps, and password policies;
PM_3	Weak password policies and software vulnerabilities allow minors to bypass restrictions and access adult content	- 2FA authentication - Warnings about adult content - Age confirmation - Page blurring - Adequate content labelling (RTA label) - Parental guidance and control instructions - Password Policy - Alerts for low password complexity - Terms of Service - Point 2 - Access - Ongoing assessment of age verification measures	- Implement forced password resets for outdated or repeated credentials; - Continue routine vulnerability assessments focusing on credential brute-forcing;

PM_4	Minors use advanced tools like VPNs or proxy services to circumvent geo-restrictions or age verification measures, gaining unauthorized access	- Ongoing assessment of age verification measures - Parental guidance and control instructions - Warnings about adult content - Age confirmation - Page blurring - Adequate content labelling (RTA label)	- Continue scoping advanced anti-circumvention solutions, balancing user privacy with minor protection needs;
PM_9	Exposure to adult content could have adverse effects on minors' physical and mental well-being, potentially leading to addiction or the development of unhealthy behaviors	- Terms of Service - Point 2 - Access - Warnings about adult content - Page blurring - Age confirmation - Parental guidance and control instructions - Adequate content labelling (RTA label) - Review of research on adult content impact - Review of research on the impact of adult content on the public health - Ongoing assessment of age verification measures	- Continue or expand collaboration with child-welfare NGOs to address early exposure prevention methods;
FR_5	The dissemination of content featuring exploitation, humiliation, or degrading material infringes upon the right to human dignity	- Automated tools: AI scanning, Thorn Safer, Google SafetyNet API, Hive classification (all) - Manual moderation process (control of participant age and consent, detecting illegal content, and copywriting material) - Multi-layered content review (combination of human and automated review) - User reports (notice and action mechanism) - Notice mechanism for trusted flaggers - Moderation team training - Terms of Service - violence related prohibitions + Point 7 - User submissions, Point 8 - Content moderation - Review of research on adult content impact - Cooperation with law enforcement agencies	- Continue research with victim advocacy groups to keep exploitation definitions updated;
FR_6	The creation and sharing of deepfake or manipulated content portraying individuals in explicit situations without consent harm their dignity	- Automated tools: AI scanning, Thorn Safer, Google SafetyNet API, Hive classification (all) - Manual moderation process (control of participant age and consent, detecting illegal content, and copywriting material) - Multi-layered content review (combination of human and automated review) - User reports (notice and action mechanism) - Notice mechanism for trusted flaggers - Moderation team training - Terms of Service - violence related prohibitions + Point 7 - User submissions, Point 8 - Content moderation - Review of research on adult content impact - Ongoing review of research and assessment of possibilities of generative AI models - Cooperation with law enforcement agencies - Cooperation with NGOs	- Continue external collaboration with digital identity experts to refine face-matching algorithms; - Establish a partnership with a specialized AI ethics consultancy to ensure compliance with emerging legislation; - Tracking newly discovered AI manipulative trends;

PH_1	Platform contributing to public health concerns such as addiction on adult materials or mental health disorders	- Review of research on the impact of adult content on the public health - Warnings about adult content	- Continue networking with experts and NGOs focusing on the health and social impacts of adult content (such as OffLimits);
PH_2	Platform not giving enough education about healthy attitudes towards sex and sexual wellness	- Review of research on the impact of adult content on the public health - Review of research on adult content impact - Warnings about adult content	- Continue networking with experts and NGOs focusing on the health and social impacts of adult content (such as OffLimits);
PH_4	Publishing content that reinforces unattainable or harmful body image ideals	- Review of research on the impact of adult content on the public health - Review of research on adult content impact - Cooperation with NGOs	- Continue networking with experts and NGOs focusing on the health and social impacts of adult content (such as OffLimits)
PH_6	Hosting content that could promote self-harm or endorse unhealthy behavioural patterns	- Automated tools: AI scanning, Thorn Safer, Google SafetyNet API, Hive classification (all) - Manual moderation process (control of participant age and consent, detecting illegal content, and copywriting material) - Multi-layered content review (combination of human and automated review) - User reports (notice and action mechanism) - Moderation team training - Terms of Service - Point 4 - Terrorism and Physical Harm Violence	- Monitor mental health studies that define self-harm or suicidal ideation in digital content; - Systematically evaluate older content for any endorsements of unhealthy or extreme behaviors;
TR_6	The platform provides incorrect data on the average monthly number of active service recipients	- Compliance Officer nomination - Internal DSA Audit - Process/Procedural Changes and Updates	- Monitor new EC or regulatory directives to ensure the chosen metrics align with official expectations; - Join industry cooperation to co-develop standardised reporting; - Document methodology used internally for transparency;

The action plan outlined in this section provides a comprehensive roadmap for mitigating the identified risks with a residual score of "Medium." By implementing the additional mitigation measures alongside existing controls, WGCZ aims to significantly reduce the likelihood and impact of these risks.

The timeline for implementing the suggested additional measures is set for 2025 and is expected to have an ongoing nature. Overall oversight is assigned to the Compliance Officer, who will be responsible for coordinating with relevant teams and ensuring steady progress in implementation. Management will remain actively engaged in the effective oversight of the systemic risks identified above, to assess progress and address any potential roadblocks. Ongoing monitoring and evaluation of the effectiveness of these mitigation measures will be crucial to ensuring WGCZ's continued compliance with regulations and its commitment to a safe and responsible online platform.

6.4. Risk monitoring and reporting procedures

Effective risk management requires a continuous process of monitoring and reporting. This chapter outlines the procedures in place to ensure timely identification and escalation of emerging risks, as well as regular communication of risk management activities.

The risk monitoring process will be conducted through a combination of ongoing activities and periodic reviews. Ongoing activities include:

- staying informed of industry trends and developments that may introduce new or exacerbate existing risks;
- regularly reviewing internal data and reports to identify potential risk indicators,
- monitoring regulatory changes and updates that might impact the risk landscape for WGCZ,
- encouraging a culture of risk awareness within the organization, where employees are empowered to report any concerns or potential risks they encounter.

Periodic reviews will be conducted as mentioned above – current periodicity is set to quarterly management reviews given the urgency and regulatory framework development. These reviews will involve a comprehensive assessment of the risk register, considering any changes in the risk landscape, the effectiveness of existing and newly implemented measures and controls, and the need for potential updates to the risk management strategy

7. Conclusion

WGCZ's updated Risk Assessment Report demonstrates the platform's commitment to responsible governance and continuous improvement in managing the complexities of an adult content environment. By expanding from 38 to 87 distinct scenarios and refining them into dedicated risk categories, WGCZ has effectively addressed the need to capture a broader spectrum of systemic risks, including gender-based violence, nuanced data privacy concerns, misinformation, and advertiser integrity. The more granular mapping of these risks and shifting to a measure-centric framework have allowed WGCZ to target each risk with evidence-based mitigation strategies.

Across all categories, the platform's mitigation control/measures (e.g., multi-layered moderation approach, transparent recommender system, robust ToS) have collectively reduced the likelihood and impact of initially high and medium risks. No risk remains high after applying existing controls, underscoring the measures' effectiveness.

Nonetheless, the analysis reveals persistent medium-level risks in sensitive areas, such as the protection of minors, non-consensual or exploitative content, deepfakes, and public health concerns. While technical measures, policy updates, and user awareness initiatives have diminished these risks substantially, they cannot be fully eradicated without sustained, multi-stakeholder engagement. Broader societal and technological factors require solutions that evolve parallel to emerging threats and regulatory developments.

Going forward, WGCZ's risk management approach will focus on three main themes:

- **Enhanced collaboration:** XVideos will intensify partnerships with NGOs, law enforcement, and specialised task forces, particularly in CSAM detection, disinformation control, and gender-based violence prevention.

- **Ongoing research:** By systematically monitoring legislative changes, AI advancements (including generative technologies), and academic research on the societal impacts of adult content, WGCZ aims to keep its controls up to date. Evaluating emerging age-assurance methods and exploring improved user guidance are prime examples of where evolving insights can further strengthen safety.

- **Audit and transparency:** Formalising a Compliance Officer role, regular internal audits, and work toward external evaluations underscore WGCZ's commitment to transparent governance. The centralised Mitigation Measures Register, user-facing policies, and reporting procedures will continue to help ensure DSA obligations remain front and foster user trust in the platform's systems.

In summary, this second Risk Assessment demonstrates clear progress in identifying and mitigating systemic and direct risks inherent in the platform's operations. Although certain residual risks remain, WGCZ's approach is agile, collaborative, and rooted in high user protection standards, legal compliance, and respect for fundamental rights. By maintaining and refining this proactive stance, XVideos is well-positioned to navigate future challenges and uphold a safe, accountable, and legally compliant service for its global user community.