XVIDEOS

RISK ASSESSMENT REPORT

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

Created on	**April 2026**
Conducted by	**WebGroup Czech Republic, a.s.**

Content

1. Executive Summary

The 2026 Systemic Risk Assessment for XVideos.com platform (also referred to as "platform" or "XVideos"), conducted by WebGroup Czech Republic, a.s., reg. no. IČO 29145465, with registered seat at Krakovská 1366/25, Praha 1, 110 00, Czech Republic (also referred to as "WGCZ"), the provider of the platform, provides a comprehensive evaluation of systemic risks associated with the platform's operation as a Very Large Online Platform ("VLOP") under the Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) ("Digital Services Act" or "DSA").

Key Changes from Previous Assessment Cycle

Compared to the previous assessment cycle (version 2.0, updated on 23 April 2025), WGCZ has significantly enhanced the identification, assessment, and documentation of systemic risks. Previously, a single risk register covered 87 scenarios across various categories. The current assessment introduces a more advanced framework that distinguishes systemic risks, risk drivers, and mitigation measures. This structure offers greater detail and operational clarity, making it easier to identify potential harms and the specific service features, governance weaknesses, or technical systems that may contribute to them.

Key changes WGCZ has implemented include:

- The most important methodological development is the adoption of a two-layer analytical model. The first layer addresses systemic risks within the meaning of Article 34(1) DSA, namely the adverse outcomes that may arise on the service, such as the dissemination of illegal content, adverse effects on fundamental rights, and negative impacts on civic discourse, public health, and users'[1] physical and mental well-being. The second layer addresses risk drivers (factors influencing the systemic risks) within the meaning of Article 34(2) DSA, namely the service features, governance arrangements, and operational factors that may cause those harms to arise or intensify.

- A second major change is the formalization of the residual-risk methodology. Unlike the previous risk assessment cycles, which relied more heavily on descriptive assessment, this assessment cycle introduces a structured two-step model for measuring residual risk. First, the model evaluates the extent to which existing mitigation measures reduce the underlying inherent risk of a given systemic-risk scenario. This effect is captured through the ***Risk Reduction Factor ("RRF")***, which reflects the median of operational effect of the controls linked to that risk. Second, the model captures the extent to which exposure may remain even where those controls operate as intended. This remaining influence is reflected in the ***Driver Impact Score ("DIS")***, which measures the ongoing effect of the relevant Article 34(2) DSA risk drivers. Taking together, these elements produce a more consistent and transparent residual-risk score that combines the effect of controls with the remaining influence of the platform's operational features.

- The governance of the assessment has also been strengthened. This assessment cycle is supported by a clearer cross-functional review process involving the relevant operational, legal, moderation, technical, advertising, and support functions. Validation workshops and review sessions were used to test individual scenarios, drivers, and mitigation measures, and the reasoning behind key judgments was documented in a more auditable manner. This significantly improves the clarity and defensibility of the assessment.

- It is important to interpret the remaining high residual-risk findings as a risk-management output rather than as a finding of non-compliance with the DSA. For a VLOP hosting adult user-generated content, the continued identification of a limited number of high residual-risk scenarios is a foreseeable and methodologically appropriate outcome where the potential harm is severe, adversarial, behaviour-driven, or structurally difficult to eliminate through platform-level controls alone. A high residual-risk rating means that the scenario remains a priority for further reduction, monitoring, and governance follow-up. It does not mean that the platform has failed to implement

[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.

mitigation measures or that a breach of the DSA has occurred. Accordingly, the numerical residual-risk values should be read together with the methodology, the layered control environment, and the risk-management strategy assigned to each scenario.

- Finally, the traceability model has been expanded. Each systemic-risk entry now links the risk description to the relevant drivers, the connected mitigation measures, the responsible functions, the quantified values, and, where relevant, the applicable risk strategy, so that the assessment forms part of a continuous governance chain rather than a standalone annual exercise.

Expanded Risk Coverage, Risk Drivers, and Control Environment:

- The revised methodology has resulted in a broader and more granular risk universe. In the current assessment cycle, WGCZ assessed 92 systemic-risk scenarios across 13 risk categories, reflecting a more platform-specific understanding of how harm may arise across uploads, comments, advertising surfaces, discovery features, account governance, and private messaging. This expansion does not reflect an artificial increase in risk, but a more mature methodology that distinguishes more clearly between harmful outcomes, the factors that drive them, and the measures intended to address them.

- In parallel, WGCZ developed a distinct Risk Driver Register identifying the principal service-level, governance, and operational factors that influence the likelihood or severity of systemic harm. The main driver categories identified in this cycle include content moderation, age assurance and access controls, advertising systems, data-related practices, manipulation and inauthentic use, regional and linguistic disparities, recommender and algorithmic systems, Terms of Service ("ToS") governance, and authority contact-point arrangements. This strengthens the assessment by explaining not only what harm may arise, but also through which concrete product features, design choices, governance weaknesses, or enforcement limitations may materialize or worsen.

- The 2026 assessment also confirms that WGCZ has established a broad and layered mitigation framework. Rather than relying on a single safeguard, the Platform applies a combination of preventive, detective, corrective, and reactive measures across the main areas of exposure. These include automated detection tools, content moderation controls, ToS and policy controls, recommender-system and user-choice measures, data privacy and security controls, minor-protection measures, advertising-integrity controls, support channels, and uploader-verification mechanisms. To improve traceability and governance, the current cycle also builds on the Mitigation Measures Register, which records existing controls, ownership, risk links, and assessed effectiveness.

Current Findings and Risk Profile

At the inherent stage of the assessment, before existing mitigation measures are considered, the overall risk profile remains elevated. WGCZ assessed the average inherent risk score across all categories at approximately 12, placing the platform's baseline exposure in the medium range. Compared with the previous assessment cycle, this reflects both the more granular methodology and stricter scoring of risks linked to minors, privacy, illegal content, and sexual exploitation. Across the 92 assessed scenarios, approximately 46% fall within the medium-risk band, 35% are assessed as high, and 20% as low.

The highest inherent risk concentrations arise in areas where the severity of harm is greatest and where the platform processes or hosts particularly sensitive content. The highest average inherent risk scores were recorded for Data Privacy and Protection Risks and Privacy & Family Life, followed by Protection of Minors, Gender-Based Violence, and Dissemination of Illegal Content. At the same time, Dissemination of Illegal Content remains the largest category by volume, with 26 distinct scenarios covering, among others, non-consensual intimate content, CSAM, grooming, exploitative sexual services, illegal links, malware, and coercive or abusive content.

Once the current mitigation framework is considered, the overall risk profile improves materially. Of the 92 assessed scenarios, 52 were classified as low residual risk, 33 as medium residual risk, and 7 as high residual risk. This demonstrates that the platform's existing controls reduce exposure meaningfully across most categories. At the same time, certain risks remain resistant to full mitigation, particularly where harm is severe, adaptive, behaviour-driven, or structurally difficult to prevent through ex ante controls alone.

Residual high-risk exposure is concentrated in two categories only: Protection of Minors and Dissemination of Illegal Content. Within the Protection of Minors category, the most serious residual risks concern minors accessing pornographic

content, being systematically exposed to explicit content through content-delivery systems, encountering sexually explicit advertising, experiencing broader pornography-related harm to well-being, and being exposed to user contact once access has already been gained. Within Dissemination of Illegal Content, the remaining high residual-risk scenarios are narrower and relate primarily to CSAM-link sharing and grooming or coercive contact involving minors through private messages.

Overall, the findings show a platform whose baseline exposure is inherently elevated given the adult-content environment, but whose control environment materially reduces risk across a substantial part of the assessed landscape. The remaining high residual risks are limited in number and concentrated in minors-related exposure and a narrow subset of illegal-content scenarios involving CSAM or grooming. **This supports the conclusion that the platform's control framework is functioning across most areas, while a limited number of particularly sensitive scenarios continue to require prioritised attention.**

Content Moderation Framework

The platform's content moderation framework remains one of the most important elements of its overall risk-control architecture. WGCZ applies a layered system combining automated detection tools, structured human review, reporting channels for users and third parties, escalation procedures, complaint-handling mechanisms, and repeat-offender enforcement rules.

Automated tools are used to identify known harmful material, suspicious patterns, and other risk indicators at scale. These tools are complemented by trained human moderators who assess flagged content in context, including legality, age, consent, copyright, and compliance with the Terms of Service. The moderation framework also includes internal classification procedures for manifestly illegal content, multilingual review capacity, dedicated notice-and-action tools, priority treatment of Trusted Flagger notices, and an appeals workflow for affected users. Taken together, these measures help reduce both under-enforcement and over-enforcement risk.

Recommender System Transparency and Adjustability

The platform's recommender and content-discovery systems are relevant both from a user-experience perspective and from a systemic-risk perspective. Recommendation logic affects which content is surfaced, prioritised, and repeatedly shown, and may therefore contribute to cumulative exposure or amplification of harmful patterns if not appropriately governed.

WGCZ addresses this area through internal review and user-facing controls. The platform explains the main factors influencing recommendations, including location, popularity, category context, and, where a user has actively enabled that function, viewing-history personalisation. Profiling is not enabled by default. Users may also influence what they see by selecting categories, tags, performers, or geographic settings, and by enabling or disabling viewing-history-based personalisation. These measures support transparency and user autonomy, although recommender-related mitigation remains partly indirect and therefore subject to continued monitoring.

Protection of Minors

Protection of minors remains one of the most sensitive and challenging areas of the assessment. The platform currently applies a set of layered safeguards intended to reduce accidental or casual underage exposure in a proportionate manner. These include adult-content warnings shown on entry, default blurring of content before access is confirmed, age self-confirmation, mandatory acceptance of the ToS prior to viewing, automated RTA metadata tagging, and guidance on parental controls and filtering tools.

These measures play an important role in reducing immediate exposure and introducing friction before explicit material becomes accessible. However, they are not equivalent to robust age verification and cannot fully prevent circumvention by determined users or through unsecured devices, shared environments, or technical workarounds. For that reason, minors-related risks remain among the most difficult to reduce to low residual levels and continue to require prioritised attention.

The area of age verification tools is highly dynamic and evolving rapidly, both from a technological and regulatory perspective. While measures such as user age verification at the website level (which WGCZ has already tested and implemented in certain jurisdictions) may effectively restrict minors' access to a particular service, they do not address

minors' exposure to online pornography more broadly and therefore do not fully resolve the underlying issue. In practice, minors can readily access many alternative websites that do not implement comparable safeguards, and under the current legal framework it remains beyond the effective enforcement capacity of regulators to address such services comprehensively.

Accordingly, website-level solutions have inherent limitations in terms of overall effectiveness. A more comprehensive approach would require age verification or parental control mechanisms at the level of the user's device or operating system, combined with filtering across online services.

Advertising Integrity

Advertising systems represent an important intersection between monetisation and user protection. In the adult-content environment, advertising risk is heightened by the possibility that deceptive, exploitative, or otherwise harmful commercial content may appear alongside user-facing content or target potentially vulnerable users.

WGCZ has implemented a structured advertising-control framework that includes pre-publication review, post-publication monitoring and enforcement, advertiser certification and eligibility controls, ad-transparency notices, an advertising repository, and impression-related advertising data to support traceability and review. These measures reduce the likelihood that harmful or non-compliant advertisements are published, remain active, or are delivered without sufficient transparency. While some residual risks remain, the assessment concludes that the platform has established a meaningful and operationally embedded advertising-integrity framework.

2. Introduction

2.1. Purpose and scope of the report

The primary purpose of the risk assessment is to:

- identify and evaluate risks across our operations that may negatively affect users, society, or our regulatory compliance,
- specifically address DSA-related risks, such as data transparency, reporting obligations, dissemination of illegal content, impacts on fundamental rights, gender-based violence, public health and minors, individual well-being, and the integrity of advertising and commercial content.

This report covers all processes, systems, and controls relevant to our platform's functionality and regulatory compliance. Areas include content moderation, user safety, data privacy, advertising practices, and reporting procedures. Risks are assessed by their likelihood and potential impact on user trust, public safety and health, protection of minors, regulatory compliance, damages, and our business objectives.

The risk assessment accounts for the diverse regional and linguistic landscape of our user base. We recognize the need to tailor our approach to address risks specific to different regions and languages. Our strategy is guided by the DSA principles of a risk-based approach and proportionality, which are integrated into our risk management system and related activities.

This report is intended to be read alongside the following related documents:

- this Risk Assessment Report, which serves as the framework document describing the methodology, the overall assessment concept, and the risk-mitigation approach;
- the Risk Register, which includes the following elements: the Systemic Risk Register, the Risk Drivers Register, the Systemic Risks Identification Catalogue and the Risk Register Dashboard summarising assessment outcomes;
- the Mitigation Measures Register, which records current controls together with ownership and effectiveness information; and
- the Action Plan, which captures completed and planned follow-up actions resulting from the assessment and subsequent monitoring.

All documents are maintained to ensure ready access for internal stakeholders, external auditors, and regulators. The assessment is reviewed and updated annually to reflect new risks, revised controls, operational changes, and regulatory developments.

2.2. Risk management framework

The risk assessment was conducted following the principles and guidelines established in the international standard ISO 31000:2018 - Risk Management - Guidelines ("ISO 31000"). 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 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. Role of the Compliance Function and Senior Management

Within WGCZ, the Compliance Function serves as the second-line assurance function and operates independently from Senior Management. For this report, it combines two roles: the statutory Compliance Officer under Article 41 DSA and the internal Risk Assessor within the ISO 31000 framework.

Its role is to provide ongoing assurance that systemic risks are identified, documented, and addressed proportionately, and that the Platform continues to meet obligations under the DSA.

Key responsibilities of the Compliance Function in this area include:

- designing and maintaining the ISO 31000-based methodology, including scoring criteria and assessment logic;
- chairing risk workshops and updating the relevant registers;
- preparing the annual systemic-risk assessment together with any ad hoc updates required under Article 34 DSA;
- assigning, monitoring, and verifying mitigation measures recorded in the Mitigation Measures Register; and
- reporting directly to Senior Management on systemic risks and potential non-compliance, including the ability to escalate concerns or issue warnings where urgent action is required.

Senior Management is responsible for strategic oversight and final accountability. It sets the platform's risk appetite, approves risk-management documentation, and ensures the overall approach aligns with the organization's regulatory and business obligations.

The principal responsibilities of Senior Management in this context include:

- determining risk appetite and approving the key risk-management documentation;
- confirming the selected strategy for risk response and ensuring that required measures are implemented;

- reviewing the annual systemic-risk assessment and its supporting documentation;
- ensuring that risk-management efforts remain aligned with strategic objectives and regulatory expectations; and
- allocating the resources necessary for mitigation, monitoring, and follow-up work.

Together, the Compliance Function and Senior Management form the core governance architecture for systemic-risk oversight. One provides independent assessment and escalation. The other supplies strategic direction, approval, and resources. The framework depends on their interaction.

3. Overview of the XVideos features

XVideos 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, private messaging 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, 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, POV, 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 user 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 may leave feedback and express opinions. These comments allow viewers to share reactions, ask questions, or highlight moments from the video, while content creators may interact with their audience through replies. The platform also allows registered users to express content preferences through “like” and “dislike” buttons. These votes are aggregated and displayed on the video page, contributing to overall popularity scores. Registered users may also “subscribe” to their favourite content creators. In addition, XVideos provides a private messaging functionality that enables direct one-to-one communication between registered users, including content creators, thereby extending interaction beyond public-facing engagement features.

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, NCII, 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 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:

- **Illegal content** that violates laws applicable within the EU and other jurisdictions:
    - CSAM and NCII,
    - Real sexual violence,
    - Terrorism-related content,
    - Extortion, blackmail, harassment,
    - Zoophilia and bestiality (real and simulated),
    - Scatophilia,
    - Real incest content,
    - Stolen private content,
    - Content involving underage individuals in non-sexual contexts, if misused; as this is distinct from CSAM and suggests material that would not, for example, constitute a legally indecent image or, an entirely innocent image, but one where the context is problematic, this should be under incompatible rather than illegal.
- **Incompatible content** that does not meet legal thresholds for illegality but violates ToS:
    - False tagging or misleading descriptions,
    - Unauthorised promotions or advertisements,
    - Private information disclosure,
    - Content with very low technical quality,
    - Spam and other platform abusive behavior,
    - Content suggesting illegal activities through language or context, potentially including credible portrayals of illegal behaviour.

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. Approach used for risk identification

The risk-identification phase was further developed in this cycle to make the assessment better grounded in evidence, easier to trace, and more clearly linked to the DSA structure. The work combined external research, internal expertise, and cross-functional validation instead of relying on a single input source.

Two-layer analytical structure

A central methodological change was adopting a two-layer analytical structure. This model separates the adverse outcomes in Article 34(1) DSA from the operational and design features in Article 34(2) DSA that contribute to them.

- The first layer concerns **Systemic Risks ("SR")**. These are the categories of harm the DSA requires VLOPs to assess, including illegal content dissemination, adverse effects on fundamental rights, protection of minors, public health, and similar societal or individual harms. In the WGCZ documentation, these risks are recorded in the Systemic Risk Register.
- The second layer concerns **Risk Drivers ("DR")**. These do not describe the harm itself but the operational, technical, and governance factors that may increase the likelihood or severity of harm. They include recommender-system design, moderation processes, enforcement of Terms of Service, advertising systems, and data-related practices. They are recorded separately in the Risk Driver Register.

The value of the two-layer structure is that it allows the assessment to clearly distinguish between the question "what harm may arise?" and "through which operational mechanisms could that harm arise or intensify?" This makes the analysis more causal, transparent, and easier to connect to mitigation measures.

In practice, a single systemic risk may be influenced by several drivers. For example, dissemination of illegal content may be shaped by moderation delays, recommender-system amplification, weak language coverage, or insufficient account-verification measures. Separating the two analytical levels makes these relationships easier to identify and document.

Overview of the process

Risk identification was carried out through an evidence-led process combining desk research, internal preparatory work, cross-functional validation, and structured documentation.

Phase 1 consisted of **external research and benchmarking**. The Compliance Function reviewed the European Commission's preliminary findings on systemic risks and mitigation[2], Ofcom guidance on illegal harms and children's risk assessment[3], academic and NGO material on image-based abuse, discrimination, public-health impacts, and algorithmic effects. It also examined relevant enforcement actions and case law concerning adult-content services. The aim was to ensure the assessment remained aligned with Article 34(1) DSA and reflected emerging risks such as AI-generated sexual imagery, grooming vectors, and harmful advertising practices.

Phase 2 comprised **internal preparatory work** by the Compliance Function. Using external evidence and operational experience from prior assessment cycles, draft lists of systemic risks and risk drivers were prepared with preliminary justifications, evidence tags, and references to observed platform conditions.

Phase 3 involved **cross-functional workshops**. In these workshops, draft lists of risks and measures were reviewed with representatives from Content Moderation, Notice and Complaint, Tech, Advertising and Support teams, along with input from the Regulation Director and Legal Team. The workshop format identifies and evaluates risk scenarios by fostering

[2] European Commission. (2025, March 18). Digital Services Act – Study on systemic risks and their mitigation: First workshop presentation. Directorate-General for Communications Networks, Content and Technology

[3] Ofcom. (2024, December 16). *Risk Assessment Guidance and Risk Profiles: Protecting people from illegal harms online* (Guidance). https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/risk-assessment-guidance-and-risk-profiles.pdf?v=390984

open communication. It allows participants to share thoughts and concerns based on their experience and expertise. This approach helps uncover threats and risk scenarios that might have gone unnoticed.

Despite the formal planning and documentation, the workshops used a flexible, open-discussion format that encouraged active participation and brainstorming. Each workshop had unique aspects but typically included discussions on new categories and specific risk scenarios. Participants collectively identified and assessed potential risk scenarios that were not detected in the previous iteration.

The risk identification and assessment 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 that could adapt to evolving regulatory requirements and the platform's operational specifics.

Phase 4 consisted of **consolidation and documentation**. The outputs from earlier phases were converted into the Risk Driver Register and the Systemic Risk Identification Catalogue. Each entry was documented with the relevant category, explanatory narrative, evidence base, ownership, and connections to other parts of the assessment.

The resulting registers form the foundation for later stages of the report. They are not static appendices but living governance tools intended to support subsequent mitigation mapping and follow-up actions.

4.2. Update of the Risk Register

Following long-term internal compliance and risk management efforts and constructive supervisory engagement with regulatory authorities, such as the European Commission or Ofcom, including requests for information, and as part of WGCZ's ongoing enhancement of its risk-management framework, the Risk Register from the previous assessment cycle was substantially revised. The earlier version used a mainly single-layer approach, under which adverse outcomes, operational weaknesses, product features, and compliance-process deficiencies were often recorded together as "risks". In the current cycle, the methodology was restructured into a two-layer model that distinguishes between systemic risks under Article 34(1) DSA and the service-level factors referred to in Article 34(2) DSA that may contribute to those risks.

As a result, the Risk Register was not just expanded but re-identified and rebuilt on a different conceptual basis. This explains why the wording, scope, grouping, and number of risk entries changed significantly from the previous cycle. The updated risk aligns more closely with the DSA logic. The Systemic Risk Register based on Article 34 focuses on harmful outcomes and societal effects from the service. The Risk Driver Register based on Article 34(2) focuses on underlying drivers such as recommender systems, moderation design, advertising systems, interface choices, age-assurance weaknesses, reporting flows, language coverage, monetisation logic, and other structural platform features.

In the previous Risk Register, several items framed as standalone "risks" are more accurately understood as drivers or control weaknesses. This especially applies to entries about inadequate moderation tooling, slow review times, ineffective reporting mechanisms, lack of regional-language adaptation, weaknesses in access verification, insufficient transparency in complaint processes, recommender-system design choices, and advertising-system governance. Under the revised methodology, these elements are no longer treated as Article 34 DSA systemic risks unless they directly describe a negative effect. Instead, they are assessed as factors that may cause or amplify systemic risks such as dissemination of illegal content, harm to minors, discrimination, privacy violations, gender-based violence, or restrictions on freedom of expression.

The revised methodology also led to a more granular and platform-specific identification of harms. Several scenarios previously grouped at a high level were split into more precise manifestations reflecting how harm materialises on an adult-content platform. For example, illegal-content risks are now differentiated across uploads, comments, advertising surfaces, and private messaging. Child-safety risks are distinguished by access, discovery, contact, exposure, and upload functionalities. Gender-based harms are separated into coercive sexual content, harassment, sextortion, deepfake abuse, and advertising-related abuse. This creates a more operationally useful register by linking systemic harms to the concrete service surfaces through which they arise.

At the same time, some former categories were reclassified or redistributed to better reflect their legal and methodological role. For example, earlier categories like "Transparency and Reporting Obligations," "User Rights and Platform Policies," "Recommender Systems and Algorithmic Design," and "Advertiser and Commercial Content Integrity" contained many items better understood as Article 35 DSA risk drivers, compliance dimensions, or mitigation domains rather than Article 34 DSA end-state harms. Their substance remains in the assessment but has been repositioned into the layer explaining

why systemic risks arise or worsen. Similarly, earlier fundamental rights entries were reorganized into more coherent outcome categories, such as privacy and family life, data privacy and protection, human dignity, freedom of expression and information, and non-discrimination.

Due to the methodological shift to a two-layered risk assessment, some items classified as systemic risks in the second XVideos risk assessment have been reassigned to the Risk Driver category; as this category did not exist in the previous report, direct comparison with the earlier register is necessarily limited and the changes should be understood as a structural refinement of the assessment rather than an expansion of the platform's risk profile.

For this reason, a direct one-to-one comparison between the former and current registers is only possible to a limited extent. Some previous risks were consolidated, some split into more specific scenarios, some reworded to better reflect adult-platform realities, and some reclassified from "risk" to "driver." The significant changes in the register should be understood as the result of an improved regulatory and methodological approach, not as evidence that the platform's risk profile has suddenly or artificially expanded.

4.3. Systemic Risks Overview

This section provides an overview of the platform's main systemic risk categories. The categorization reflects the structure of the *Systemic Risks Identification Catalogue* and groups scenarios arising from the platform's core functionalities, including user uploads, comments, titles, advertising, discovery features, monetization, account governance, and private messaging.

The risks were identified through a combined methodology. First, the assessment drew on the systemic-risk areas recognized in the DSA context and in the external materials referenced in the *Systemic Risks Identification Catalogue*, including regulatory and academic sources cited in the justifications. Second, the identification process incorporated internal workshops and operational input from internal teams. Third, risks were derived from platform-specific misuse patterns observed or reasonably foreseeable in adult-content environments, particularly where explicit content, anonymity, user-to-user interaction, and monetization create heightened exposure to harm. As a result, the categories below reflect both external evidence and internal service knowledge.

1. **Dissemination of illegal content**

For a large adult-content platform, dissemination of illegal content remains one of the most significant risk areas under the DSA. The category covers unlawful material that may be uploaded, distributed, linked, promoted or recirculated through the service in breach of applicable law. The scenarios captured in this category include:

- non-consensual intimate imagery, deepfakes, leaked paywalled content, and doxxing links (SR-IC-01);
- prohibited sexual or exploitative material, including coercive content, trafficking, or extreme pornography (SR-IC-02);
- advertising or solicitation of sexual services through comments or private messages, including exploitative or trafficked services (SR-IC-03, SR-IC-04, SR-IC-20, SR-IC-21);
- dissemination of CSAM, including by upload, external links, QR codes, shortened links, or private messages (SR-IC-05, SR-IC-06, SR-IC-22);
- grooming, coercion of minors, or arranging contact through comments or private messages (SR-IC-07, SR-IC-23);
- animal cruelty or abusive sexualised material involving animals (SR-IC-08);
- malware, phishing, or malicious links shared through comments or messages (SR-IC-09, SR-IC-24);
- copyright and IP infringement, including pirated creator content (SR-IC-10);
- terrorist propaganda, extremist content, harassment, coercion, and hate speech (SR-IC-11, SR-IC-12, SR-IC-18, SR-IC-25);
- drugs, weapons, unsafe products, and other illegal or unregulated goods or instructions (SR-IC-13, SR-IC-14, SR-IC-19);

- self-harm and suicide-related content, including simulated variants where harmful normalisation may occur (SR-IC-15, SR-IC-16, SR-IC-26);
- foreign interference or coordinated manipulation of public opinion (SR-IC-17).

The main external sources were the *European Commission's 2025 DSA workshop on systemic risks* ("**EC's DSA workshop**")[4] and *Ofcom's 2024 illegal harms guidance and risk profiles* ("**Ofcom's Risk Guidance**")[5]. Together, they provided the core typologies for illegal content relevant to adult-content services. These sources were used to map platform functionalities such as uploads, comments, advertising, link sharing, and private messaging against recognised risk areas including CSAM, grooming, trafficking, non-consensual imagery, illegal goods and services, hate speech, terrorist content, phishing, self-harm, and foreign interference.

The identification was further supported by sector-specific enforcement materials, including the Federal Trade Commission settlement with Pornhub[6] (2025) ("**FTC Settlement**")[7] and the U.S. Department of Justice resolution[8] concerning proceeds linked to sex trafficking. These confirmed that adult-content platforms face heightened exposure to NCII, exploitative material, and governance failures in preventing illegal sexual content.

Internal workshops with the Content Moderation and Advertising teams helped refine the category by identifying recurring misuse patterns observed in practice. These included solicitation of sexual services, unsafe product advertising, self-harm content, and misuse of comments, ads, and messaging features. The category was identified by combining recognised external illegal-content typologies with the platform's operational experience.

2. Gender-Based Violence

Gender-based violence involves systemic risks where the platform may host, amplify, monetise, or facilitate abuse targeting women and girls or reproducing misogynistic and coercive norms. In the adult-content environment, this includes direct victimisation and broader structural harms from recommendation, advertising, and community features. Included risks are:

- violent, coercive, or abusive sexual depictions that normalise harmful attitudes (SR-GB-01);
- misogynistic, sexist, or toxic content that reinforces harmful stereotypes (SR-GB-02);
- threats, stalking, brigading, doxxing, or harassment targeting women creators or public figures (SR-GB-03, SR-GB-09);
- extortion using stolen, coerced, or fabricated intimate material (SR-GB-04);
- non-consensual pornography and AI-generated deepfake abuse targeting women (SR-GB-05);
- sexualised disinformation used to damage the credibility of women in public or leadership roles (SR-GB-06);
- advertisements that monetise, promote, or normalise gender-based abuse or degrading portrayals of women (SR-GB-07, SR-GB-08).

[4] European Commission. (2025, March 18). *Digital Services Act – Study on systemic risks and their mitigation: First workshop presentation* [Workshop presentation]. Directorate-General for Communications Networks, Content and Technology

[5] Ofcom. (2024, December 16). *Risk assessment guidance and risk profiles* [PDF]. Ofcom. https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/risk-assessment-guidance-and-risk-profiles.pdf?v=390984

[6] Pornhub is an online platform that provides user-generated and professionally produced adult video content, allowing users to upload, view, and share explicit material. It is one of the largest adult-content websites globally and is operated by Aylo (formerly MindGeek).

[7] Federal Trade Commission. (2025, September 3). *FTC takes action against operators of Pornhub and other pornographic sites for deceiving users about efforts to crack down on child sexual abuse material and nonconsensual sexual content*. https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-takes-action-against-operators-pornhub-other-pornographic-sites-deceiving-users-about-efforts

[8] U.S. Attorney's Office, Eastern District of New York. (2023, December 21). *Pornhub parent company admits to receiving proceeds of sex trafficking and agrees to three-year deferred prosecution agreement*. U.S. Department of Justice. https://www.justice.gov/usao-edny/pr/pornhub-parent-company-admits-receiving-proceeds-sex-trafficking-and-agrees-three-year

The analysis relied on Lim et al. (2015)[9] and Wright et al. (2015)[10] to identify the risk that violent or coercive pornography may reinforce attitudes supportive of violence against women and sexually aggressive behaviour. Risks linked to the amplification of misogyny, sexist stereotypes, harassment, stalking, and sextortion were further informed by the EC's DSA workshop, by Soneji et al. (2024)[11], which documents harassment, doxxing, stalking, and content leakage affecting creators on adult platforms.

The category was also shaped by case-based and enforcement evidence showing how gender-based abuse manifests in practice. This included the Associated Press (2025)[12] reporting on the *GirlsDoPorn case*[13] as evidence of coercive sexual exploitation, Greenberg (2025)[14] on sextortion spyware, the FTC Settlement (2025) concerning failures to prevent NCII, Ajder et al. (2019)[15] on the prevalence of deepfake pornography targeting women, and Koltai (2024)[16] on networks circulating and monetising deepfake NCII. Risks linked to sexualised disinformation and the silencing of women in public life were identified with reference to the Khan (2023)[17] and material from the U.S. Department of State (2023)[18].

Internal input from the Advertising team helped identify recurring risks linked to ad submissions, including attempts to promote or monetise non-consensual sexual content, degrading portrayals of women, or material normalising coercion and abuse. This category reflects both recognised external risk patterns and platform-specific operational experience.

3. Protection of Minors

Protection of minors is one of the most sensitive systemic-risk categories for an adult-content platform. The European Commission's Article 28 DSA Guidelines (2025) ("**EC's Art. 28 Guidelines**")[19] require platforms accessible to minors to ensure a high level of privacy, safety, and security, and make clear that a provider cannot rely only on terms prohibiting minors if it does not implement effective measures to prevent access. Ofcom's 2025 Children's Risk Assessment Guidance (2025) ("**Ofcom's Children's RA**")[20] similarly requires services likely to be accessed by children to carry out a dedicated

[9] Lim, M. S. C., 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 & Community Health*. Advance online publication. https://doi.org/10.1136/jech-2015-205453

[10] Wright, P. J., Tokunaga, R. S., & Kraus, A. (2016). A meta-analysis of pornography consumption and actual acts of sexual aggression in general population studies. *Journal of Communication, 66*(1), 183–205. https://doi.org/10.1111/jcom.12201

[11] Soneji, A., Hamilton, V., Doupé, A., McDonald, A., & Redmiles, E. M. (2024). *"I feel physically safe but not politically safe": Understanding the digital threats and safety practices of OnlyFans creators*. In *33rd USENIX Security Symposium (USENIX Security 24)*. USENIX Association. https://www.usenix.org/system/files/usenixsecurity24-soneji.pdf

[12] Associated Press. (2025, May 13). *Founder of a California-based porn empire sentenced to 27 years in federal prison*. AP News. https://www.apnews.com/article/porn-site-sex-trafficking-guilty-d4d4f3069e15ad3ea7ada372986aec10

[13] GirlsDoPorn was a pornography production company involved in a major legal case in the United States, where operators were found to have coerced and deceived women into participating in explicit videos under false pretenses. Victims were misled about how the content would be distributed, and the material was later widely published online without their informed consent. The case resulted in criminal convictions for sex trafficking, fraud, and related offenses, as well as significant civil judgments in favor of the victims.

[14] Greenberg, A. (2025, September 3). *Automated sextortion spyware takes webcam pics of victims watching porn*. Wired. https://www.wired.com/story/stealerium-infostealer-porn-sextortion/

[15] Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The state of deepfakes: Landscape, threats, and impact* [PDF]. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

[16] Koltai, K. (2024, February 23). *Behind a secretive global network of non-consensual deepfake pornography*. Bellingcat. https://www.bellingcat.com/news/2024/02/23/behind-a-secretive-global-network-of-non-consensual-deepfake-pornography/

[17] Khan, I. (2023, August 7). *Gendered disinformation and its implications for the right to freedom of expression* (A/78/288). United Nations, Office of the High Commissioner for Human Rights. https://www.ohchr.org/en/documents/thematic-reports/a78288-gendered-disinformation-and-its-implications-right-freedom

[18] U.S. Department of State. (2023, March 27). *Gendered disinformation: Tactics, themes, and trends by foreign malign actors*. https://2021-2025.state.gov/gendered-disinformation-tactics-themes-and-trends-by-foreign-malign-actors/

[19] European Commission. (2025, October 10). Guidelines on measures to ensure a high level of privacy, safety and security for minors online, pursuant to Article 28(4) of Regulation (EU) 2022/2065 (OJ C/2025/5519). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:C_202505519

[20] Ofcom. (2025, April 24). *Children's risk assessment guidance and children's risk profiles* [PDF]. Ofcom. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/statement-protecting-children-from-harms-online/main-document/childrens-risk-assessment-guidance-and-childrens-risk-profiles.pdf?v=396653

children's risk assessment and identify adult services as a key space in which children search for and view pornographic content.

For the initial identification exercise, the assessment used the **5C lens from Annex I to the EC's Art. 28 Guidelines (2025)**: Content, Conduct, Contact, Consumer, and Cross-cutting risks. The materials explain that minors may be harmed not only by content but also by interactive features, commercial practices, platform design, and governance choices. It includes recommender systems, search autocomplete, and search results within discovery and recommendation risks. The **Ofcom's Children's RA** identifies direct messaging, commenting, posting images or videos, user-generated-content search, content tagging, and content recommender systems as functionalities that can increase risk to children.

Table 1: Mapping systemic risk scenarios related to the protection of minors to the 5C typology

Risk ID	Systemic Risk Scenario	Primary 5C category	Possible secondary	Justification[21]
SR-PM-01	Minors access pornographic content on the service, resulting in exposure to harmful or illegal sexual material	Content	Cross-cutting	This risk is best classified primarily as Content because the immediate risk is minors encountering pornographic material, which Ofcom's Children's RA treats as primary priority content harmful to children. It should also be classified secondarily as Cross-cutting because the likelihood of exposure depends heavily on the effectiveness of the platform's age-assurance, access-control, and governance arrangements. The EC's Art. 28 Guidelines recommend age verification for adult-content platforms and specify that age-assurance methods should be accurate, reliable, robust, non-intrusive, and non-discriminatory. The systemic relevance of this risk is supported by evidence that child access to pornography is widespread: Arcom (2024)[22] reported 2.3 million minors per month visiting pornographic sites in France. Of those who have seen online pornography, the average age they first encounter it is 13 – although more than a quarter come across it by age 11 (27%), and one in ten as young as 9 (10%) (Children's Commissioner for England, 2023)[23]. In the UK, Ofcom's child-safety work under the Online Safety Act is premised on the fact that children are encountering online pornography (Ofcom, 2025)[24].
SR-PM-02	Users upload or share CSAM or images of minors in sexualised contexts	Conduct	Content	This risk is best classified primarily as Conduct because the harm depends on users actively uploading, sharing, redistributing, or forwarding the material. The systemic relevance of this risk is reinforced by the following reporting data. Thorn et al. (2024)[25] received more than 21.7 million reports of suspected child sexual exploitation in 2020, including 21.4 million from electronic service providers; reporting based on those data recorded 20,307,216 reports from Meta-owned platforms and 13,229 from MindGeek, of which 4,171 were described as unique. Happ et al. (2024)[26] analysis of adult online content likewise identifies CSAM as a core illegal-content problem on pornographic websites and notes that, absent external pressure, moderation had previously been insufficient.

[21] The justification column is based on WGCZ's internal document: *Systemic Risks Identification Catalogue*

[22] Arcom. (2024, October 11). *Access by minors to pornographic content: Arcom publishes its guidelines*. https://www.arcom.fr/en/press/access-minors-pornographic-content-arcom-publishes-its-guidelines

[23] de Souza, R. (2023, January 31). *'A lot of it is actually just abuse'- Young people and pornography*. Children's Commissioner for England. https://www.childrenscommissioner.gov.uk/resource/a-lot-of-it-is-actually-just-abuse-young-people-and-pornography/

[24] Ofcom. (2025, January 16). *Age checks to protect children online*. https://www.ofcom.org.uk/online-safety/protecting-children/age-checks-to-protect-children-online

[25] Thorn & National Center for Missing & Exploited Children. (2024, June). *Trends in financial sextortion: An investigation of sextortion reports in NCMEC CyberTipline data* [PDF]. https://info.thorn.org/hubfs/Research/Thorn_TrendsInFinancialSextortion_June2024.pdf

[26] Happ, M., Harpenau, F., & Wiewiorra, L. (2024). Economics and regulation of adult online content (WIK Working Paper No. 9). WIK Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste. https://www.econstor.eu/bitstream/10419/308077/1/1913368831.pdf

				It is appropriately classified secondarily as Content because the immediate risk is the presence and circulation of illegal sexual material involving minors. Ofcom's Children's RA states that images or videos showing a child engaged in sexual activity, or appearing to be engaged in sexual activity, constitute CSAM and illegal content.
SR-PM-03	Minors are systematically exposed to explicit pornographic content through the platform's content delivery systems	Content	Cross-cutting	This risk is classified primarily as Content because the direct harm is minors' exposure to pornographic content, which Ofcom's Children's RA identifies as primary priority content harmful to children. Cross-cutting (advanced algorithms and recommender systems) is the secondary category because the risk is intensified by platform systems that prioritize and repeatedly serve content to minors. EC's Art. 28 Guidelines explain that recommender systems determine how information is displayed to minors and may amplify harmful content. Ofcom's Children's RA states that personalized recommendations are the main pathway for children to harmful content online. It shows that recommender systems are a major source of cumulative harm because they can repeatedly expose children to harmful material, prolong use, and serve harmful content even when it is not actively sought out.
SR-PM-04	Minors encounter harmful sexual content through discovery features such as search suggestions, autocomplete, or trending tags	Content	Cross-cutting	This risk is classified primarily as Content because the direct harm is minors' exposure to harmful sexual content. Cross-cutting is the right secondary category since discovery features can shape exposure even without deliberate searching. Ofcom's Children's RA identifies user-generated-content searching, predictive or suggestive search, and content tagging as relevant risk factors. It states that search tools can increase the risk of children encountering pornographic content and explains that tagging can improve the discoverability of harmful content. The EC's Art. 28 Guidelines reinforce this with a broader safety-by-design approach and recommend changes to recommender systems to reduce children's exposure to harmful content.
SR-PM-05	Minors who use comments are exposed to unwanted contact from unknown users, which may escalate into grooming	Contact	Conduct, Cross-cutting	Features such as direct messaging, friend requests, and live chat significantly increase the risk of unwanted contact, which may result in grooming or coercion. Contact risks constitute a central category of harm within EC's Art. 28 Guidelines and Ofcom's Children's RA, as they threaten children's privacy and safety. A Cross-cutting element is also present because the likelihood and severity of the risk depend on platform design choices, including whether minors' accounts are private by default, whether strangers can interact with them, and whether blocking, muting, reporting, and comment-control tools are available and effective.
SR-PM-06	Minors are extorted through threats to publish intimate images or are coerced into producing CSAM	Contact	Conduct, Content, Cross-cutting	This scenario is best classified as Contact because the primary harm results from a coercive interaction where a minor is pressured or blackmailed into producing intimate material. Such risks may arise from features like live chat, image or video sharing, and anonymous messaging. Conduct is a secondary category since the perpetrator's actions are criminal. Content is also relevant because producing sexual material involving a minor may constitute CSAM. Cross-cutting categories such as privacy and safety are implicated due to blackmail, repeated victimisation, and significant loss of control over intimate material. Data from Thorn et al. (2024)[27] show that sextortion of minors is a significant and growing concern. It received an average of 812 reports per week between August 2022

[27] Thorn & National Center for Missing & Exploited Children. (2024, June). *Trends in financial sextortion: An investigation of sextortion reports in NCMEC CyberTipline data* [PDF]. https://info.thorn.org/hubfs/Research/Thorn_TrendsInFinancialSextortion_June2024.pdf

				and August 2023 and nearly 100 reports of financial sextortion per day in 2024.
SR-PM-07	Minors are exposed to sexually explicit or pornographic ads (e.g., live sex cams, hookup services, sex toys, and sexual enhancement products).	Consumer	Content	This risk is a Consumer category because the harm arises from minors' exposure to commercial communications for adult products or services, including sexually explicit advertisements. EC's Art. 28 Guidelines identify consumer risks as including minors' exposure to advertisements intended for adults. Content is an appropriate secondary category because the advertisements may contain or promote sexually explicit material harmful or age-inappropriate for minors. This classification aligns with Ofcom's Children's RA, which treats pornographic material as harmful to children and requires providers to assess how service design and functionalities affect children's exposure to such material. The systemic relevance of the risk is strengthened by the fact that adult-content platforms commonly rely on advertising and cross-promotion as part of their business model (e.g., TrafficJunky[28] across major properties).
SR-PM-08	Exposure to pornography harms minors' well-being, contributing to anxiety, confusion, and distorted views of relationships and sexuality.	Content	Cross-cutting	Exposure to pornography can negatively affect minors' well-being, contributing to anxiety, confusion, distress, and distorted views of relationships, sexuality, and body image. However, the severity of harm varies among children. A six-country European study found that online pornography exposure is common and linked to externalising problems, particularly rule-breaking and aggressive behaviour (Andrie et al., 2021)[29]. Common Sense Media (2022)[30] also reported frequent, often accidental, exposure among U.S. teenagers, including violent or aggressive content and stereotyped portrayals, which many associated with negative emotions such as disgust and self-consciousness. Maheux et al. (2021)[31] identified links between adolescent pornography use and increased self-objectification and body comparison, suggesting effects on body-related self-perception, though not on body shame. The evidence remains mixed: Smahel (2020)[32] notes that children's reactions to sexual images differ by context, age, and individual experience, and Owens et al. (2012)[33] highlight inconsistencies in the literature. For 5C purposes, this scenario is best classified as Content, as pornographic material is the triggering exposure. The secondary classification is Cross-cutting due to the broader emotional, developmental, and psychological risks to minors.

[28] TrafficJunky's Adult Traffic Network: Increase Website Traffic

[29] Andrie, E. K., Sakou, I. I., Tzavela, E. C., Richardson, C., & Tsitsika, A. K. (2021). Adolescents' online pornography exposure and its relationship to sociodemographic and psychopathological correlates: A cross-sectional study in six European countries. *Children, 8*(10), Article 925. https://doi.org/10.3390/children8100925

[30] Common Sense Media. (2022). *Teens and pornography: A survey of U.S. teens ages 13–17* [PDF]. https://www.commonsensemedia.org/sites/default/files/research/report/2022-teens-and-pornography-final-web.pdf

[31] Maheux, A. J., Roberts, S. R., Evans, R., Widman, L., & Choukas-Bradley, S. (2021). Associations between adolescents' pornography consumption and self-objectification, body comparison, and body shame. *Body Image, 37*, 89–93. https://doi.org/10.1016/j.bodyim.2021.02.004

[32] Smahel, D., Machackova, H., Mascheroni, G., Dedkova, L., Staksrud, E., Ólafsson, K., Livingstone, S., & Hasebrink, U. (2020). *EU Kids Online 2020: Survey results from 19 countries*. EU Kids Online. https://doi.org/10.21953/lse.47fdeqj01ofo

[33] Owens, E. W., Behun, R. J., Manning, J. C., & Reid, R. C. (2012). The impact of internet pornography on adolescents: A review of the research. *Sexual Addiction & Compulsivity, 19*(1–2), 99–122.

SR-PM-09	Minors are exposed to direct contact with unverified or unsafe users through private messaging	Contact	Cross-cutting, Content	For this risk, the primary category is Contact, because the core risk is that minors are exposed to direct interpersonal interaction with unknown, unverified, or unsafe users through private messaging. Under the EC's Art. 28 Guidelines, contact risks include situations in which minors are harmed through interactions with other users. The secondary category of Cross-cutting is also appropriate, because the likelihood and severity of this risk depend heavily on service design and governance choices: whether the platform allows private messaging, whether minors can be contacted by default, whether unknown users can initiate chats, and whether safety-by-design protections are in place. The secondary category of Content can also be justified, but only as a secondary one. It fits because private messages are not only a channel for contact; they may also be used to transmit pornographic material, links, or other harmful content. Ofcom's Children's RA identifies direct messaging as a risk factor because it enables harmful material to be shared in a targeted and less visible way, often without the recipient's permission, including pornographic links or sexual content sent directly to children.
SR-PM-10	Minors are exposed to ads promoting false standards and facts about the human body, which can lead to self-doubt, mental health problems, and body dysmorphia	Consumer	Cross-cutting	This scenario is best classified primarily as Consumer, because the immediate harm arises through commercial content and promotional messaging directed at or encountered by minors. The EC's Art. 28 Guidelines require platforms to ensure that minors are not exposed to harmful, unethical, or unlawful advertising and that minors limited commercial literacy is not exploited. Cross-cutting (health and well-being) is the appropriate secondary category because advertising that promotes unrealistic bodily ideals or misleading appearance claims may contribute to body dissatisfaction, low self-esteem, anxiety, and broader mental-health harms.
SR-PM-11	Minors with regular accounts can upload content on the platform, leading to confusion or the unintentional or intentional dissemination of CSAM	Conduct	Content, Cross-cutting	This risk is best classified primarily as Conduct. The EC's Art. 28 Guidelines define conduct risks as behaviours minors may actively adopt online and explicitly includes minors posting or sending violent or pornographic content, as well as illegal behaviour such as posting or sending CSAM. Content is the appropriate secondary category because once uploaded, such material may constitute or become CSAM, creating a separate illegal-content risk. Cross-cutting is also appropriate because the likelihood of this harm depends on service-design and governance choices, including age assurance, account settings, upload permissions, and moderation capacity. Ofcom's Children's RA supports this approach by noting that weak age assurance can let children create false-age profiles and access adult-only functions, while "posting images or videos" is itself a recognised risk factor in the user-to-user risk profiles.

4. Data Privacy and Protection

This category captures risks arising from the collection, processing, sharing, or misuse of personal and highly sensitive data. In the context of the platform, this category is particularly significant because the platform may process intimate, sexual, and otherwise highly personal information that, if mishandled, can result in serious harm to individuals. The following risks are included:

- use of sensitive data concerning sexual identity or preferences for profiling, resulting in discrimination or unequal access to the platform (SR-DP-01);
- large-scale processing of sensitive data without valid consent, undermining autonomy and enabling exploitation or unlawful secondary use (SR-DP-02);
- linking intimate material or identifying information to victims' families or communities, causing harm to dignity, relationships, and social standing (SR-DP-03).

This category was identified mainly through the EC's DSA workshop, which confirms that inappropriate and excessive use of personal data is a recognised systemic risk, especially when personal data practices harm the right to data protection. It also identifies systemic misuse of personal data, lack of valid consent, and weak transparency as core threats to users' autonomy and rights. These findings informed the identification of risks related to profiling based on sexual identity or preferences and the large-scale processing of sensitive data without meaningful consent.

The same source highlights sharing highly personal information and doxxing as severe privacy violations that can cause reputational damage, social exclusion, and withdrawal from public spaces. This informed the identification of risks where intimate material or identifying information is linked not only to the individual but also to their family or wider community, creating secondary harm, stigma, and relational damage. Thus, the category reflects both direct privacy infringements and broader social consequences when sensitive sexual data is mishandled, disclosed, or exploited.

5. Civic and Electoral Impact

Civic and electoral impact risks concern systemic situations in which the platform may be used to distort public discourse, manipulate political perceptions, or interfere with democratic participation and electoral integrity. In the adult-content environment, these risks can emerge through uploads, titles, comments, community interactions, or advertising systems that are exploited to circulate political disinformation, propaganda, or targeted manipulation under the cover of non-political or explicit content. Included risks are:

- political disinformation or election-related falsehoods appearing in pornographic videos, images, titles, comments, or advertisements (SR-CE-01);
- coordinated groups using porn-related communities, comments, or titles to launder mass disinformation and manipulated narratives (SR-CE-02);
- political or state-linked actors buying or placing advertisements on the platform to spread election manipulation narratives or propaganda (SR-CE-03);
- targeted advertising disproportionately exposing vulnerable or marginalised groups to political disinformation and manipulation campaigns (SR-CE-04);
- political candidates, activists, and journalists being targeted with pornographic disinformation, including fake pornographic deepfakes or fabricated narratives, in ways that intimidate or silence their participation in public debate or elections (SR-CE-05).

This category was identified primarily through the EC's DSA workshop. The workshop identifies false election-related narratives, foreign information manipulation and interference, coordinated manipulation of public discourse, and the systemic difficulty of moderating such content in real time as significant risks to civic discourse and electoral integrity. These findings informed the identification of risks linked to political disinformation appearing within pornographic videos, titles, comments, and advertisements. They also include the use of porn-related communities or engagement features to circulate manipulated narratives under non-political cover.

The same source highlights the role of opaque advertising, micro-targeting, and dark advertising in distorting voter perceptions and shaping unequal exposure to political messaging. This informed the identification of risks linked to political or state-linked actors placing covert propaganda through the platform's advertising systems. It also includes vulnerable or marginalised groups being disproportionately targeted through behavioural profiling and targeted advertising. In the adult-platform context, these risks increase where advertising systems or recommendation features enable covert influence operations without meaningful transparency.

The category also reflects risks to democratic participation from abuse directed at public figures and participants in civic life. The workshop specifically notes harassment, deepfake abuse, and cyber-violence targeting women leaders as practices that undermine credibility and discourage participation in public debate. This informed the identification of risks involving pornographic disinformation, including fabricated sexual content or fake pornographic deepfakes targeting candidates, activists, and journalists to intimidate, silence, or discredit them. Accordingly, this category captures both the manipulation of political information flows and the use of sexualised abuse to suppress civic participation and undermine electoral processes.

6. Freedom of Expression and Information

Freedom of expression and information risks concern systemic situations in which the platform's rules, moderation systems, ranking, demonetisation, or advertising practices unduly restrict lawful speech, reduce access to information, or suppress pluralism in sexual identities and viewpoints. In the adult-content environment, these risks are particularly acute because lawful sexual expression may be subject to heightened stigma, broad contractual restrictions, or inconsistent moderation standards. Included risks are:

- excessively aggressive or inconsistent removal of lawful content, creating systemic restrictions on freedom of expression and limiting the pluralism of sexual identities and legitimate practices (SR-FE-01);
- restriction of LGBTQ+, kink, fetish, or other minority sexual content, thereby reducing pluralism of identities and viewpoints (SR-FE-02);
- advertising policies or practices that disproportionately restrict or demonetise LGBTQ+ sexual content compared with heterosexual content (SR-FE-03);
- hidden or inconsistent decisions relating to suspension, demonetisation, de-indexing, or channel status that suppress lawful creator expression and remove access to audience and revenue without meaningful explanation or redress (SR-FE-04).

The Council of Europe (2025)[34] stresses that content moderation frameworks must not create a chilling effect on lawful speech, and warns against overzealous removal, hidden demotion, and insufficient transparency or redress in moderation systems. These principles directly informed the identification of risks associated with aggressive takedowns, opaque enforcement, shadow banning, de-indexing, suspension, and the demonetization of lawful adult content. The EC's DSA workshop likewise identifies excessive, inappropriate, or arbitrary content removal, weak transparency, and insufficient safeguards for pluralism as relevant systemic concerns in protecting freedom of expression and information.

The category was also shaped by evidence showing how abrupt, externally pressured moderation changes can suppress lawful adult expression at scale. In December 2020, Pornhub removed about 10 million videos and sharply restricted uploads after payment processors suspended services. This reduced the platform's catalogue from roughly 13.5 million to under 3 million videos (Axios, 2020)[35]. Although the purge was framed as a response to abuse risks, it shows how large-scale enforcement shifts can sweep lawful content into removal decisions and drastically narrow access to lawful expression. Likewise, in August 2021, OnlyFans announced it would prohibit "sexually explicit" content following pressure from banking and payout partners, only to reverse the policy days later after obtaining assurances from those partners (Brus, 2021)[36]. This episode strongly exemplifies opaque, unstable rulemaking that chills lawful adult expression, especially when creators depend on the platform for visibility and income.

Risks related to restricting LGBTQ+, kink, fetish, and other minority sexual expression were also identified in the freedom of expression systemic-risk discussion at the EC's DSA workshop. It notes that systemic moderation and algorithmic bias hinder media and opinion pluralism and reduce diverse opinions, perspectives, and ideologies. In the adult-content sector, discriminatory ad placement and monetisation practices form a structural expression-suppression: LGBTQ+ or non-mainstream sexual content receives less commercial visibility and revenue, reinforcing inequality in representation and voice.

7. Consumer Protection

In the adult-content environment, these risks may arise not only through misleading offers or manipulative design, but also through opaque terms, uneven redress, discriminatory abuse of platform systems, and enforcement processes that can be strategically weaponised against lawful users. Included risks are:

---

[34] Council of Europe, Parliamentary Assembly, Committee on Culture, Science, Education and Media. (2025, January 8). *Regulating content moderation on social media to safeguard freedom of expression* (Doc. 16089). https://pace.coe.int/en/files/33939/htm

[35] Axios. (2020, December 17). *Pornhub's video purge poses a legal riddle*. https://www.axios.com/2020/12/17/pornhubs-video-purge-legal-riddle

[36] Brus, D. (2021, August 19). *OnlyFans to prohibit "sexually explicit" content on platform*. *Axios*. https://www.axios.com/2021/08/19/onlyfans-prohibit-sexually-explicit-content

- users being left uncertain about their rights, protections, or obligations on the platform, resulting in diminished trust and a reduced ability to exercise consumer rights (SR-CP-01);
- manipulative interface designs or dark patterns pressuring users into subscriptions or into sharing personal data (SR-CP-03);
- deceptive or misleading advertisements or promotions, such as fake premium offers or hidden fees, that exploit users (SR-CP-04);
- unequal enforcement of consumer rights and protections across Member States, creating systemic disparities in exposure to harm and access to redress (SR-CP-05);
- coordinated mass reporting against a lawful channel, video, or comment as an act of discrimination, hate, or unfair competition (SR-CP-06);
- coordinated mass reporting targeting LGBTQ+ persons, racial minorities, or other groups as an act of discrimination, hate, or unfair competition (SR-CP-07);
- automated or AI-driven coordinated reporting attacks targeting specific users through the flagging system, causing psychological, reputational, and monetisation harm (SR-CP-08).
- Weak cybersecurity enables cyberattacks or hacking incidents to breach the website database and collect sensitive data from channel owners, leading to privacy breaches, mental harm, and possible blackmail (SR-CP-09).

This category was identified primarily through the EC's DSA workshop, which highlights lack of transparency, manipulative design, deceptive practices, and unfair commercial conduct as key consumer-protection risks. These findings informed the identification of risks linked to unclear user rights, dark patterns, and misleading promotions.

The category was also shaped by the risk that platform enforcement and complaint mechanisms are applied inconsistently or abused. Coordinated mass-reporting campaigns and automated flagging attacks can distort complaint handling, trigger unjustified enforcement against lawful users, and undermine fair treatment and effective redress. This category reflects both established consumer-protection concerns and platform-specific risks related to misuse of reporting systems.

8. Non-Discrimination

Non-discrimination risks concern systemic situations in which the platform's systems, policies, or commercial practices disproportionately disadvantage certain groups and result in unequal treatment, visibility, protection, or economic opportunity. In the adult-content environment, these risks may arise through biased monetisation, discriminatory advertising, unequal regional protections, and platform features that reinforce harmful stereotypes. Included risks are:

- minority or LGBTQ+ creators being unfairly excluded from ad revenue opportunities, undermining equal participation (SR-ND-01);
- discriminatory ads that amplify hate speech or harmful stereotypes, including racist, sexist, or homophobic content (SR-ND-02);
- users in non-English-speaking regions receiving weaker protections and greater exposure to harm than English-speaking users (SR-ND-03);
- racialised or fetishising categories disproportionately exposing certain groups to degrading stereotypes (SR-ND-04);
- visibility, monetisation, support decisions, or account-status upgrades disproportionately disadvantaging LGBTQ+, racialised, non-English-speaking, or other marginalised creators (SR-ND-05);
- non-English-speaking users receiving different treatment from the Notice and Complaint team, resulting in slower processing or inappropriate handling (SR-ND-06).

This category was identified primarily through the EC's DSA workshop, which confirms that bias in recommendation algorithms, moderation processes, and ad targeting leads to structural exclusion of minorities and unequal access to visibility or revenue. On adult platforms, these effects manifest as reduced monetisation opportunities for LGBTQ+ or minority creators, amplification of racialised or fetishised stereotypes, and weaker protection against harmful content in non-English regions.

9. Public Security Concerns

This risk category involves systemic situations in which the platform may be misused to facilitate criminal activity, extremist mobilisation, violent incitement, trafficking, or targeted intimidation. In the adult-content environment, these risks may arise through uploads, comments, creator accounts, advertising tools, or other platform features that are exploited by bad actors. Included risks are:

- extremist groups using adult content as cover to recruit members, spread propaganda, or radicalise users (SR-PS-01);
- criminal networks using creator accounts, advertising, or comments to facilitate human trafficking and sexual exploitation disguised as adult services (SR-PS-02);
- offenders uploading violent or hate-driven sexualised content that incites hostility, fuels extremist narratives, or glorifies violence (SR-PS-03);
- coordinated actors targeting journalists, activists, or civil society with harassment campaigns using explicit material, thereby undermining public trust and civic discourse (SR-PS-04);
- extremist groups use the platform's encrypted private messaging system to recruit members, spread propaganda, or radicalise users (SR-PS-05).

This category was identified primarily through the EC's DSA workshop, which identify such risks as propaganda, recruit members, or normalise violence under the guise of explicit content. These findings informed the identification of risks linked to radicalisation, trafficking, sexual exploitation, and violent or hate-driven sexualised content on the platform.

The category was also shaped by the compliance team's review of external research on image-based sexual abuse, which showed that explicit material can be weaponised to intimidate or silence journalists, activists, and women in public life. This category therefore reflects both broader public-security threats and platform-specific risks linked to harassment and abuse through sexualised content.

10. Human Dignity

Human dignity risks involve systemic situations in which the platform hosts, amplifies, or monetizes content that degrades, humiliates, or dehumanizes individuals or groups, undermining respect for their inherent worth and equal status. In the adult-content environment, these risks are especially acute when sexual content portrays humiliation, coercion, racist abuse, misogyny, or other degrading treatment as normal or acceptable. One risk was identified in this category:

- users posting degrading or dehumanizing sexual content, such as extreme humiliation or racist or misogynistic abuse, that normalizes violations of human dignity and reduces protection for affected groups (SR-HD-01).

This category was identified during the Compliance and Content Moderation teams' workshop. It reflects the risk that repeated exposure to such material may normalize coercive or abusive behavior, desensitize users to violence, and weaken social respect for human dignity.

11. Public Health

Public health risks concern systemic situations in which the platform may normalise unsafe sexual practices, amplify misleading health information, or expose users to harmful products and behaviours with broader health consequences. In the adult-content environment, these risks may arise through both user content and advertising. Included risks are:

- promotion of unprotected sex or risky sexual behaviour without appropriate warnings or context (SR-PH-01);
- misleading sexual-health information or unsafe advice (SR-PH-02);
- unrealistic body standards contributing to self-esteem harm or disordered behaviour (SR-PH-03);
- explicit content involving drug use or unregulated substances, including the normalisation of chemsex-related practices (SR-PH-04);
- promotion of unsafe sexual-performance drugs, supplements, or treatments (SR-PH-05).

This category was identified through public-health policy materials, peer-reviewed studies and health-authority guidance. American Public Health Association (2010)[37] and Grudzen et al. (2009)[38] identified the risk that repeated depictions of unprotected sex may normalise unsafe sexual practices, as both document low condom uses and portray high-risk sexual acts in adult films. Rothman et al. (2021)[39] and Miller and Stubbings-Laverty (2022)[40] informed the identification of misleading sexual-health information, showing that pornography can serve as a source of sexual learning and may shape inaccurate beliefs about what is normal or safe.

The risk related to unrealistic body standards was identified through Paslakis et al. (2022)[41], which found associations between pornography exposure and negative body image outcomes, and Dubinskaya and Anger (2023)[42], which links pornographic representations of female genitalia to labiaplasty trends. Risks involving chemsex-related practices were identified through Holden (2021)[43], Bolmont et al. (2022)[44], and Mundy et al. (2025)[45], which document chemsex pornography and its links to sexual-health, dependency, and mental-health harms. The risk of unsafe sexual-performance products was identified through Tucker et al. (2018)[46] and NCCIH (2025)[47], which show that sexual-enhancement supplements may contain unapproved pharmaceutical ingredients and pose significant health risks.

12. Physical and Mental Well-being

Physical and mental well-being risks may arise through exposure to violent or extreme material, prolonged use patterns, non-consensual intimate content, and advertising or AI-generated material that encourages unhealthy consumption or unrealistic expectations. Included risks are:

- exposure to violent or extreme sexual material that may shape harmful perceptions of sexuality and relationships (SR-PW-01);
- extended platforms use that may contribute to compulsive consumption patterns harmful to users' health (SR-PW-02);
- sharing or distribution of non-consensual intimate content causing serious psychological harm to victims (SR-PW-03);

[37] American Public Health Association. (2010, November 9). *Prevention and control of sexually transmitted infections and HIV in the adult film industry* [Policy brief].

[38] Grudzen, C. R., Elliott, M. N., Kerndt, P. R., Shuster, M. A., & Brook, R. H. (2009). Condom use and high-risk sexual acts in adult films: A comparison of heterosexual and homosexual films. *American Journal of Public Health, 99*(Suppl. 1), S152–S156. https://doi.org/10.2105/AJPH.2007.127035

[39] Rothman, E. F., Beckmeyer, J. J., Herbenick, D., Fu, T.-C., Dodge, B., & Fortenberry, J. D. (2021). The prevalence of using pornography for information about how to have sex: Findings from a nationally representative survey of U.S. adolescents and young adults. *Archives of Sexual Behavior, 50*(2), 629–646. https://doi.org/10.1007/s10508-020-01877-7

[40] Miller, D. J., & Stubbings-Laverty, R. (2022). Does pornography misinform consumers? The association between pornography use and porn-congruent sexual health beliefs. *Sexes, 3*(4), 578–592. https://doi.org/10.3390/sexes3040042

[41] Paslakis, G., Chiclana Actis, C., & Mestre-Bach, G. (2022). Associations between pornography exposure, body image and sexual body image: A systematic review. *Journal of Health Psychology, 27*(3), 743–760. https://doi.org/10.1177/1359105320967085

[42] Dubinskaya, A., & Anger, J. T. (2023). Female genitalia in pornography: A source of labiaplasty trends? *The Journal of Sexual Medicine, 20*(2), 124–125. https://doi.org/10.1093/jsxmed/qdac037

[43] Holden, G. (2021). The involuntary confession of euphoria: 'Chemsex' porn and the paradox of embodiment. *Sexualities*. Advance online publication. https://doi.org/10.1177/13634607211008070

[44] Bolmont, M., Tshikung, O. N., & Trellu, L. M. (2022). Chemsex, a contemporary challenge for public health. *The Journal of Sexual Medicine, 19*(8), 1210–1213. https://doi.org/10.1016/j.jsxm.2022.03.616

[45] Mundy, E., Carter, A., Nadarzynski, T., Whiteley, C., de Visser, R. O., & Llewellyn, C. D. (2025). The complex social, cultural and psychological drivers of the 'chemsex' experiences of men who have sex with men: A systematic review and conceptual thematic synthesis of qualitative studies. *Frontiers in Public Health, 13*. https://doi.org/10.3389/fpubh.2025.1422775

[46] Tucker, J., Fischer, T., Upjohn, L., Mazzera, D., & Kumar, M. (2018). Unapproved pharmaceutical ingredients included in dietary supplements associated with U.S. Food and Drug Administration warnings. *JAMA Network Open, 1*(6), e183337. https://doi.org/10.1001/jamanetworkopen.2018.3337

[47] National Center for Complementary and Integrative Health. (2025, September 12). *4 things to know about dietary supplements marketed for sexual enhancement*. U.S. Department of Health and Human Services.

- push-style ads, such as “live now” or “exclusive access,” that may encourage compulsive consumption and harm users’ mental health (SR-PW-04);
- exposure to AI-generated videos, ads, or images that promote unrealistic body standards or sexual practices (SR-PW-05).

This category was identified through a combination of the EC’s DSA workshop, peer-reviewed academic literature, and input from the internal advertising team. Wright et al. (2016)[48] and Hald and Malamuth (2015)[49] informed the identification of risks linked to violent or extreme sexual material, as both sources indicate that exposure to such content may be associated with pro-aggression attitudes, acceptance of violence, or harmful perceptions of sexuality, although the effects are moderated by individual and contextual factors. Hellevik et al. (2025)[50] informed the identification of harms linked to non-consensual intimate content, showing that image-based sexual abuse is associated with severe outcomes, including anxiety, depression, and social withdrawal.

The category was also shaped by evidence that excessive online service use may contribute to compulsive behavior and health-related harms, as reflected in research from the University of Cambridge (2014)[51] on compulsive sexual behavior. In addition, the advertising team workshop held in the second risk assessment cycle helped identify platform-specific risks linked to urgency-based advertising and AI-generated sexualised material, including their potential to intensify compulsive consumption and reinforce unrealistic body or sexual norms.

13. Privacy & Family Life

This risk category concerns systemic situations in which the platform may enable the disclosure, misuse, or weaponisation of intimate material or personal information in ways that seriously interfere with an individual’s private life, family relationships, safety, and dignity. In the adult-content environment, these risks are especially acute because the platform hosts sexual content that may be uploaded, shared, monetised, copied, or manipulated without consent. Included risks are:

- users uploading or sharing non-consensual intimate images or videos, including leaked paywalled content, with the intention of harming a person’s private life (SR-PF-01);
- users doxxing creators or victims by posting personal information, such as addresses, contact details, or family information, connected to explicit content (SR-PF-02);
- users sharing private content or information to threaten or blackmail individuals, deliberately harming their private and family life (SR-PF-03);
- users publishing manipulated images of real individuals, including deepfakes or photoshopped material, to damage reputations and family relationships (SR-PF-04).

This category was identified through the EC’s DSA workshop, which highlights intimate image abuse, doxxing, harassment, extortion, and deepfake abuse as key risks in digital environments, particularly on adult-content services. These findings informed the identification of risks linked to NCII, exposure of personal data, blackmail using private sexual material, and manipulated intimate content involving real individuals. The category reflects the risk that such practices may seriously interfere with privacy, personal security, dignity, and family life.

The category was also shaped by additional legal and enforcement sources. Europol (2024)[52] informed the identification of sextortion and blackmail risks by showing the growing relevance of extortion involving intimate content in organised online crime. Directive (EU) 2024/1385[53] further supported the identification of deepfake and AI-generated intimate-image

[48] Wright, P. J., Tokunaga, R. S., & Kraus, A. (2016). Meta-analysis of pornography consumption and actual acts of sexual aggression in general population studies. *Journal of Communication, 66*(1), 183–205. https://doi.org/10.1111/joc.12201

[49] Hald, G. M., & Malamuth, N. M. (2015). Experimental effects of exposure to pornography: The moderating effect of personality and mediating effect of sexual arousal. *Archives of Sexual Behavior, 44*(1), 99–109. https://doi.org/10.1007/s10508-014-0291-5

[50] Hellevik, P. M., Haugen, L.-E. A., & Överlien, C. (2025). Outcomes of image-based sexual abuse among young people: A systematic review. *Frontiers in Psychology, 16*. https://doi.org/10.3389/fpsyg.2025.1599087

[51] University of Cambridge. (2014, July 11). *Brain activity in sex addiction mirrors that of drug addiction*.

[52] Europol. (2024). *IOCTA 2024: Internet organised crime threat assessment*.

[53] Directive (EU) 2024/1385 of the European Parliament and of the Council on combating violence against women and domestic violence.

abuse as a relevant and recognised harm within the EU legal framework. Together, these sources show that privacy and family life may be harmed not only through direct disclosure of real intimate material, but also through manipulation, coercion, and threats involving sexualised content.

4.4. Risk Drivers Overview

The risk drivers described in this section identify exposure factors assessed for prudential and compliance purposes. They should not be read as findings that each issue has materialised in practice in all cases, but as areas where service design, operational choices, or governance arrangements may influence systemic-risk outcomes.

1. Content Moderation Drivers

Content moderation processes form the front line of defence against the dissemination of illegal or otherwise incompatible materials on the platform. They are central to compliance with Articles 16, 17, 20, and 22 DSA, including notice-and-action mechanisms, statements of reasons, complaint-handling systems, and the handling of trusted flagger notices. In an adult-content environment, weaknesses in moderation can directly increase exposure to illegal sexual material, non-consensual content, child-safety risks, discrimination, and failures in user protection. At the same time, overly restrictive or inaccurate moderation may suppress lawful expression and disproportionately affect legitimate creators or minority groups.

The key risk drivers are:

- Ineffective, delayed, or inconsistent moderation enforcement: moderation systems may fail to detect and remove illegal or harmful content quickly enough; rules may be applied inconsistently; and high-severity material may not be escalated in time to specialist teams or law enforcement. At the same time, automated tools or overly cautious decision-making may wrongly remove lawful content, including legitimate sexual expression or educational material. This creates a combined risk of both under-enforcement and over-enforcement.
- Weak controls against repeat abuse and limited moderation capacity: inadequate verification and account-linking mechanisms may allow offenders to return after enforcement and continue re-uploading harmful material. These risks are further amplified where moderation teams lack sufficient staffing, specialist expertise, or resilience due to burnout and repeated exposure to distressing material.
- Inaccessible or poorly functioning notice, appeal, and transparency mechanisms: reporting options for illegal or harmful content may be difficult to find or use; notices may not be handled promptly, diligently, or consistently; statements of reasons may be missing or unclear; and appeal pathways may be confusing, inaccessible, or poorly reviewed. Similar weaknesses may affect both image/comment moderation and video moderation, undermining transparency and user redress rights.
- Weak external reporting and oversight integration: notices submitted by trusted flaggers may not be prioritised or processed consistently, reducing the effectiveness of an important external safeguard and weakening cooperation with qualified reporters.
- Weak channel governance and ownership enforcement: insufficient verification or re-verification of channel ownership may allow bad actors to monetise pirated, leaked, stolen, or non-consensual content through privileged channel status. At the same time, ownership-claim or piracy-reporting tools may themselves be abused to suppress legitimate creators, remove lawful content, or divert monetisation and reach.
- Minor-safety risks from platform design and interaction pathways: minors may interpret verified, model, or channel accounts as safer or platform-endorsed; they may be redirected from adult creator profiles or videos to external adult destinations; and they may also be exposed to unwanted or unknown friend requests that create pressure or grooming-related risks.
- Weak age assurance, age-gating, and minor upload controls: ineffective age verification may allow minors to access the site, accept tracking technologies, and become exposed to explicit material before age confirmation. These weaknesses may be worsened by unofficial access channels, VPN-based circumvention, insufficient checks for verified accounts, and permissive upload controls that allow minors to upload sexual or otherwise illegal material, including in extreme cases content that may qualify as CSAM.

These drivers show that content-moderation risk arises across the full moderation chain: detection, review, escalation, reporting, transparency, appeals, channel governance, age assurance, and user-safety design. They also demonstrate the close interdependence between policy enforcement, technical controls, and operational capacity.

2. Age Assurance Drivers

Age assurance is a critical protective and compliance function for a platform hosting adult or explicit content. Under Article 35 of the DSA, the provider must take, where applicable, targeted measures to protect the rights of the child, including age verification and parental control tools, tools aimed at helping minors signal abuse or obtain support, as appropriate. At the same time, any age-assurance measures must remain consistent with privacy and data protection requirements. In this context, ineffective age assurance can expose minors to explicit content, adult advertising, and other higher-risk interactions, while overly intrusive controls may create separate privacy, security, and trust concerns.

The following age-assurance risk drivers were identified through an internal assessment of the platform's age-verification controls, account-creation flows, age-gating mechanisms, access restrictions for adult content and advertising, and feature-level safeguards for comments, messaging, and other interactive functions.

The key risk drivers are:

- Limitations in age-verification and account-access controls: superficial or insufficient age checks may allow minors to obtain regular access to the platform, use higher-risk features, and interact in spaces where malicious actors can contact or groom them. This risk is particularly acute where commenting, friend requests, private messaging, or encrypted contact features are accessible without robust age assurance.
- Inadequate blocking of adult content and advertising before age confirmation: where age-gating is not effectively enforced before access, minors may be exposed to pornographic advertisements or adult content elements such as images, titles, or categories before confirming age and acceptance of the relevant access conditions.
- Privacy and data-protection risks in age-verification methods: certain age-assurance solutions may rely on intrusive biometric checks or government-issued ID data without adequate safeguards. This creates heightened risks around data minimisation, security, misuse of sensitive personal data, and overall proportionality of the measure.

These drivers show that age-assurance risk sits at the intersection of minor protection and access control. Systems that are too weak may allow underage access to explicit content and risky interactions.

3. Advertising Systems Drivers

This section summarises the main categories of risk drivers identified in the current assessment cycle. Risk drivers refer to the operational, technical, and governance factors that can affect the likelihood or impact of systemic risks. Each driver category is documented in the Risk Driver Register and was analysed through cross-functional workshops and operational evidence.

Advertising systems represent a critical intersection between commercial operations and user protection. Under Articles 26 and 39 DSA, the platform must ensure that advertising is transparent, accountable, and does not contribute to the dissemination of illegal or harmful content. In an adult-content environment, these risks are heightened by the nature of the content, the potential exposure of vulnerable users, and the possibility of monetising exploitative or unsafe material.

The following advertising-related risk drivers were identified through an internal assessment of the platform's advertising systems, in particular the ad delivery and targeting systems, ad-ranking and optimisation logic, advertiser onboarding and verification controls, ad-category screening mechanisms, and monetisation controls.

The key risk drivers are:

- Transparency risks in advertiser disclosure: the system does not clearly disclose advertiser identity, sponsorship, or targeting parameters, allowing hidden actors to influence users without accountability.
- Opaque engagement-based ranking: ad ranking may amplify degrading or fetishised categories because such content performs better on engagement metrics.

- Revenue-driven optimisation: algorithms prioritising click-through rate and revenue may systematically promote extreme, risky, or unlawful ads.
- Weak advertiser verification: insufficient onboarding and verification controls may allow traffickers, exploiters, or fraudulent actors to purchase ad space anonymously.
- Insufficient ad-category screening: the system may allow unsafe sexual-enhancement products, unverified treatments, or similar harmful products to be advertised.
- No effective safeguards against covert influence advertising: the ad ecosystem may be misused for covert political or state-linked advertising in explicit-content contexts.
- Monetisation of abusive or non-consensual material: failures in ad eligibility controls may enable monetisation of non-consensual, coercive, or abusive pornographic content.

These drivers show that weaknesses in advertising governance can directly increase systemic risks, including manipulation, discrimination, consumer harm, public-health risks, and the monetisation of illegal or abusive content. They also highlight the structural tension between revenue optimisation and compliance obligations.

4. Data-Related Practice Drivers

Data-related practices underpin nearly every aspect of the platform's operation, including content hosting, account management, advertising, creator onboarding, support processes, and regulatory compliance. For an adult-content platform, these risks are particularly acute because both content and associated metadata may reveal intimate behaviour, sexual preferences, identity-linked activity, or other highly sensitive personal information. Weaknesses in data governance therefore create not only cybersecurity exposure, but also systemic risks under the DSA and GDPR, including privacy harm, unlawful profiling, loss of user trust, and regulatory sanctions.

The key risk drivers are:

- Weak security and internal access controls for sensitive data: inadequate technical and organisational safeguards may enable breaches of intimate content, private user information, or creator data. These risks are increased where internal permissions are too broad, auditability is weak, or staff and contractors can access, copy, or leak highly sensitive user material without sufficient restriction.
- Excessive tracking and weak consent controls: users may be continuously tracked through cookies, device fingerprints, or ad-tech identifiers that reveal intimate browsing behaviour, while lacking clear opt-in or opt-out choices, granular settings, or effective consent flows. This creates risks around unlawful profiling, insufficient user autonomy, and non-compliance with data-protection principles.
- Insufficient protection, sharing, and retention controls for sensitive data and metadata: metadata such as IP addresses, geolocation, and contact details may be insufficiently protected and capable of being linked to explicit content; sensitive data may be shared with advertisers or external vendors without adequate transparency or safeguards; and intimate content or related metadata may be retained for excessive periods without effective erasure rights. Together, these practices increase the risk of de-anonymisation, secondary misuse, and disproportionate harm in the event of a breach.
- Weak governance for DSA researcher access obligations: the absence of a designated point of contact, combined with insufficient staffing or operational capacity, may delay or prevent responses to vetted researcher requests for access to publicly available data. This creates a distinct compliance risk under the DSA, including potential financial penalties and reputational harm.

These drivers show that data-related risk is not only classic cybersecurity failures, but also the broader governance of consent, access, retention, transparency, third-party sharing, and legally mandated data access.

5. Manipulation and Inauthentic Use

Manipulation and inauthentic behaviour represent a cross-cutting operational risk for user-generated content platforms. Such behaviour can exploit recommendation systems, reporting channels, creator tools, in ways that distort visibility, undermine user trust, and weaken the effectiveness of moderation and complaint-handling processes. In an adult-content environment, these risks are particularly acute because the same features that support creator reach and user engagement

can also be abused to amplify exploitative material, coordinate harassment, spread illegal links, or mislead users through false signals of legitimacy.

The key risk drivers are:

- Coordinated manipulation of platform systems: insufficient safeguards against bots, fake accounts, engagement farms, coordinated harassment, or abusive reporting may allow actors to game recommender systems, artificially amplify harmful or illegal content, mob targeted users in comment sections, or overwhelm moderation and complaint-handling capacity. These behaviours can distort visibility, burden enforcement teams, and interfere with the platform's ability to respond effectively to genuine harms.

- Abuse of links, creator tools, and dispute mechanisms: weak detection of URLs in comments may allow actors to distribute illegal-content links, phishing redirects, or other malicious destinations. At the same time, piracy-claim or ownership-dispute tools may be misused to exclude competitors, intimidate smaller creators, or unfairly reduce the visibility of legitimate content. These weaknesses turn platform governance tools into channels for manipulation rather than protection.

- False trust signals and ineffective user remedies: verified, model, or channel statuses may create a misleading impression of authenticity, safety, or platform endorsement, increasing user susceptibility to scams, deceptive promotions, premium-content upselling, or manipulative off-platform solicitation. These risks are compounded where users face persistent barriers to reporting harmful content, appealing decisions, or resolving disputes, resulting in a broader denial of effective remedy.

These drivers show that manipulation and inauthentic use are not limited to fake accounts or spam, but extend across the platform's visibility systems, reporting channels, dispute tools, creator governance features, and user trust architecture.

6. Regional and Linguistic Drivers

Regional and linguistic disparities represent a structural driver of systemic risk for platforms operating across multiple EU jurisdictions. Detection tools, moderation workflows, reporting channels, and recommendation settings are often developed first for high-volume or major-language markets. As a result, smaller linguistic or regional environments may receive weaker protection, slower enforcement, and less accessible redress.

The key risk drivers are:

- Uneven moderation and detection coverage across languages: moderation may be disproportionately focused on English or other major languages, while smaller language markets receive fewer human moderation resources and weaker automated detection. As a result, illegal or harmful content in regional languages, dialects, slang, or coded terminology may remain undetected for longer or be removed less consistently.

- Exploitation of weaker protections in non-major language markets: gaps in detection and enforcement for smaller languages may enable manipulation campaigns, including foreign interference, extremist propaganda, disinformation, grooming, trafficking, or other illegal activity, to target users in environments with weaker safeguards.

- Insufficient localisation of legal and redress frameworks: moderation processes may not be fully aligned with stricter regional legal requirements, and reporting or appeal mechanisms may not be adequately adapted for smaller or regional languages. This can lead to compliance failures, delayed enforcement, reduced user trust, and barriers to exercising notice and redress rights in a user's native language.

- Inconsistent country-level recommender settings: differences in recommender system configurations across EU countries may result in uneven exposure to harmful content, creating geographic disparities in user protection and content risk.

These drivers show that regional and linguistic risks are not limited to translation quality, but affect the full governance chain: detection, human review, legal compliance, user redress, and recommendation systems.

7. Recommender and Algorithmic Systems

Recommender and algorithmic systems are among the most influential parts of the platform's risk architecture. They determine what content is surfaced, ranked, promoted, or repeatedly shown to users, significantly amplifying both legitimate engagement and harmful dynamics. Under the DSA, a platform must ensure that recommender systems are transparent, offer meaningful user choice, and do not systematically contribute to the spread of illegal or harmful content. In an adult-content environment, these risks are especially acute because algorithmic ranking directly shapes exposure to explicit material, exploitative content, discriminatory narratives, harassment, and manipulative off-platform behaviour.

The key risk drivers are:

- Engagement-driven ranking may amplify illegal, harmful, or exploitative content: optimisation for clicks, engagement, or virality may increase the visibility of illegal sexual exploitation material, highly engaging pirated content, risky new uploads, sensationalist sexual content, or harmful body-image content before adequate safeguards are applied. In practice, ranking systems may reward precisely the content categories that create the highest user reaction, even where they present elevated legal, safety, or rights-based risks.

- Interactive ranking and virality features may intensify abuse and unsafe user exposure: "most engaging" sorting of comments or chats may elevate grooming attempts, phishing links, and CSAM redirects, while trending signals, subscriptions, and activity feeds may amplify coordinated harassment, dogpiling, repeat abuse, and repeated exposure to risky creators. These features can reduce the coordination cost of harmful behaviour and extend the reach of abusive actors.

- Algorithmic clustering and profiling may reinforce discrimination, extremity, and civic harm: recommender logic may reinforce degrading or discriminatory sexualised categories, suppress minority content, or create echo chambers around extremist, trafficking-related, or violent material. In parallel, the use of sensitive-category profiling or the surfacing of disinformation and polarising election-related narratives embedded within adult content may create discrimination, privacy, and civic-integrity risks.

- Insufficient transparency and lack of meaningful user choice weaken user autonomy: where users lack visibility into the key inputs, objectives, and signals of recommender systems, they cannot make informed choices about how content is prioritised. This risk is compounded where the Platform does not provide a meaningful non-profiling recommendation option, such as a chronological or otherwise non-personalised feed, limiting compliance with user-choice obligations.

These drivers show that recommender and algorithmic systems are not neutral discovery tools, but core amplifiers of systemic risk. When governed primarily by engagement or growth metrics, they can intensify exposure to illegal content, discriminatory narratives, manipulative behaviour, and unsafe interactions.

8. TOS Obligations

The ToS form the central legal and operational framework governing platform enforcement, and user rights. Under Article 14 DSA, the platform must ensure that these terms are clear, accessible, and applied in a diligent, non-arbitrary, and transparent manner. Weaknesses in the design or implementation of the ToS can therefore create systemic risks across moderation, redress, discrimination, and regulatory compliance. In practice, vague rules, unclear procedures, or inconsistent application may undermine user understanding, chill lawful expression, and weaken the platform's ability to enforce its own standards fairly and effectively.

The key risk drivers are:

- Vague, inconsistent, or unevenly applied rules: definitions of prohibited illegal or otherwise incompatible content may be vague, contradictory, or inconsistent across jurisdictions, and the ToS may be applied unevenly across countries, languages, or user groups. This creates risks of unpredictability, discriminatory enforcement, and reduced fairness in moderation outcomes.

- Insufficient transparency about enforcement and user rights: the ToS and related user-facing flows may not clearly explain moderation measures, the use of automation, notice-and-action processes, statements of reasons, complaint-handling mechanisms, or appeal rights. As a result, users may not understand how decisions are made or how to challenge them effectively.

- Overly narrow or overly broad substantive rules: concern about regulatory liability may lead to excessively broad removals under the Terms of Service, chilling lawful expression and reducing pluralism, while an overly narrow focus on strictly illegal material may leave harmful but lawful content, such as misogynistic abuse, body-image harm, or encouragement of self-harm, insufficiently addressed. In parallel, weak escalation rules for repeat violators may allow persistent harmful actors to return and continue causing harm.
- Insufficient clarity, accessibility, and localisation of rights information: ToS and rights-related information may not be provided in clear, localised, or age-appropriate language, limiting user understanding, especially for users in non-dominant language groups. Similarly, the absence of clear and accessible contact points for user-protection matters may reduce the effectiveness of consumer protection and user assistance.
- Weak provisions for official notices and authority contact channels: where the Platform lacks a designated contact point for official notices, or where verification of such notices is insufficiently secure, there is a risk that lawful removal orders or information requests will not be received, not be actioned, or be processed incorrectly, including in response to spoofed or fraudulent requests. Although closely linked to the separate contact-point risk category, these weaknesses also reflect deficiencies in ToS and governance design.

These drivers show that ToS risk is not only a matter of legal drafting, but also of governance quality, user comprehension, consistency of enforcement, and procedural fairness.

9. Contact point for authorities

Ensuring a reliable, continuously monitored contact point for authorities is a key compliance requirement under Article 11 of the DSA. This function allows competent authorities across EU Member States to transmit lawful orders, including removal orders and requests for information, through a designated and dependable communication channel.

The key risk drivers are:

- Risk of insufficient reliability or monitoring of the authority contact channel: where no designated and effectively monitored contact point exists, official notices such as removal orders or information requests may not be received, routed, or processed at all, creating a direct risk of non-compliance with lawful authority requests.
- Language barriers in processing official notices: where official notices are issued only in national languages that relevant staff do not understand, the Platform may face delays, misinterpretation, or failure to comply in a timely and accurate manner with legitimate orders.

These drivers show that the contact-point function, although administrative in nature, is an important part of the platform's regulatory interface.

5. Risk Assessment

5.1. Inherent Risk Assessment

In ISO 31000 risk-management standards, inherent risk is defined as the level of risk 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):

- **Rare (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 or other protected objectives. Impact scores range in the following categories:

- **Insignificant (Score 1):** Minimal to no potential harm to users and society. 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).
- **Moderate (Score 3):** Increased potential for harm to users (e.g., safety risks, privacy breaches) or society (e.g., manipulation, erosion of trust).
- **Major (Score 4):** Significant potential for harm to users (e.g., widespread exposure to harmful content) or society (e.g., manipulation of elections).
- **Critical (Score 5):** Severe potential for harm to users (e.g., exploitation, psychological harm) or society (e.g., major disruption of democratic processes).

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:

Inherent Risk Score = Likelihood Score × 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

ASSESSMENT CRITERIA

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 limited and the likelihood of occurrence remains relatively remote. These risks are generally narrower in scope, more isolated in nature, or less central to the platform's broader systemic-risk profile.

Medium Risk (**Score 6–15, Yellow**):
Medium risks represent a meaningful level of concern. These scenarios have either a moderate probability of occurring, a more material potential impact, or both. They often reflect identifiable vulnerabilities, recurring operational challenges, or areas where exposure could become more serious if left insufficiently addressed over time.

High Risk (**Score 16–25**, **Red**):
High risks are the most significant inherent-risk scenarios and indicate elevated baseline exposure before mitigation. These scenarios typically concern either serious user harm, high-likelihood recurring patterns, or a combination of both, particularly in areas involving minors, intimate content, privacy, coercion, and other severe forms of abuse.

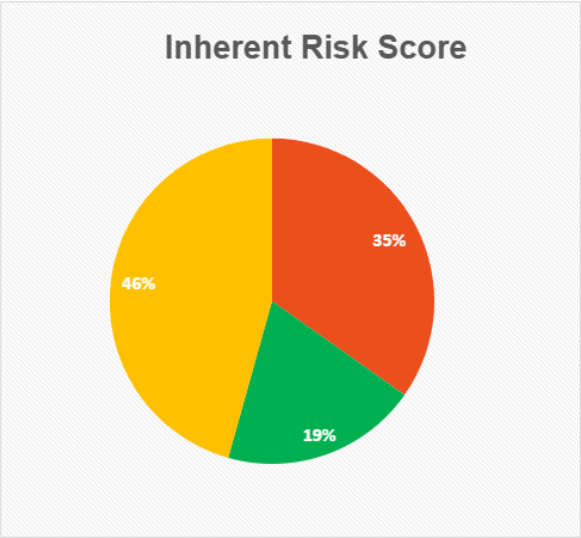
Systemic Risks Assessment Outcomes

WGCZ evaluated **92 risk scenarios** across **13 defined risk categories** to determine the inherent risk score. The average inherent risk score across all assessed scenarios is approximately 12.4, placing the platform's overall inherent risk exposure within the medium range. This indicates a meaningful but manageable baseline level of exposure before the effect of mitigating measures is taken into account. Compared with last year's report, the average inherent risk score has increased from 11 to approximately 12.4, suggesting a somewhat broader and more granular identification of relevant systemic-risk scenarios in the current cycle. While this reflects a moderately higher baseline assessment, it also shows that the platform is applying a more structured and mature approach to identifying risk exposure.

The assessment results shown in Figure 2 indicate that the platform's inherent risk profile remains centred primarily in the **medium-risk range**, while also including several high-risk scenarios in areas that are inherently sensitive for an adult-content service. Of all assessed risk scenarios, 42 out of 92 fall within the medium-risk range, 32 out of 92 are classified as high risk, and 18 out of 92 fall within the low-risk range. This means that approximately 46% of scenarios are medium risk, 35% are high risk, and 20% are low risk. Overall, this distribution suggests a diverse risk landscape, with the largest share of scenarios concentrated in the medium range rather than at the most severe end of the scale.

The highest average inherent risk scores are recorded in Data Privacy and Protection Risks (average score 20) and Privacy & Family Life (approx. 20), followed by Protection of Minors (approx. 17), Gender-Based Violence (approx. 15), Physical and Mental Well-being (approx.14), and Dissemination of Illegal Content (approx. 13). These categories stand out because they concern areas where adult-content services naturally face the highest baseline sensitivity, including minors, intimate data, coercion-related harms, and non-consensual content. The results are broadly aligned with the areas most likely to require attention in an inherent-risk assessment conducted before controls are considered.

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 some areas present greater concentrations of risk than others. At the same time, in systemic-risk assessment it is important to consider not only the severity of individual risks, but also the number of scenarios identified within each category. The volume of scenarios provides useful insight into the breadth of exposure. A category with a high number of medium-scoring risks may represent a level of exposure comparable to, or in some cases greater than, a category with fewer but individually higher-scoring risks. In this sense, systemic significance is shaped both by how severe risks are and by how broadly they are distributed across platform functions, user journeys, and types of interaction.

When considering both the inherent risk score and the number of risks in each category, **Dissemination of Illegal Content** remains the category with the broadest overall exposure. It comprises **26 distinct risk scenarios** and has an average inherent risk score of approx. 13. The scale of this category reflects the platform's broad exposure to legality and enforcement-related threats, including non-consensual intimate content, CSAM, grooming through private messaging, malware or phishing links, illegal commercial offers, and other forms of prohibited or abusive material. Although its average score is lower than in some more concentrated categories, the number of scenarios means it remains one of the platform's most operationally important areas of focus and an area where scalable controls and procedures are especially valuable.

Protection of Minors and Gender-Based Violence also remain categories of elevated concern. Protection of Minors includes 11 risk scenarios with an average inherent score of approx. 17, while Gender-Based Violence contains 9 scenarios with an average score of approx. 15. These categories include several of the platform's higher-scoring individual risks, such as minors accessing explicit content, minors being exposed to private messaging and grooming-related contact, the dissemination of non-consensual sexual material, and the use of intimate material for coercion, humiliation, or extortion. While these categories do not match Dissemination of Illegal Content in overall volume, their consistently elevated scores confirm that they remain among the most important priority areas for continued mitigation and oversight.

The categories with the highest average inherent risk scores are Data Privacy and Protection Risks and Privacy & Family Life, with average scores of 20 and 19.75 respectively. These categories contain only 3 and 4 scenarios, but the seriousness of those scenarios places them at the top of the category comparison. They include risks relating to profiling based on sensitive sexual data, processing without valid consent, linkage of intimate material to victims' families or communities, blackmail, and deliberate harm to a person's private or family life through the dissemination of intimate content. Although smaller in number, these risks are clearly concentrated in areas where the platform's mitigation framework and compliance safeguards are particularly important.

A second group of categories falls within **a moderate-to-elevated** concern range. Consumer Protection contains **8** scenarios with an average inherent score of approx. 12, while Human Dignity, although represented by only 1 identified scenario, also carries an average score of 12. Public Health contains 5 scenarios with an average score of 9, while Public

Security Concerns contains 5 scenarios with an average score of approx.8. These categories remain relevant because they capture harms that may be less dominant in aggregate but still warrant continued monitoring and proportionate mitigation, particularly where they intersect with broader patterns of exploitation, misinformation, manipulative conduct, or unsafe user experiences.

At the lower end of the spectrum, Non-Discrimination contains 6 scenarios with an average inherent score of approx. 5, while Freedom of Expression and Information contain 4 scenarios with an average score of approx. 4. Civic and Electoral Impact includes 5 scenarios and has the lowest average score, at 4. These categories are therefore currently assessed as presenting comparatively lower inherent risk in platform-specific terms, generally because the likelihood, scale, or service-specific relevance of the identified harms is more limited. Even so, their inclusion in the assessment remains important, as it demonstrates that the platform is also reviewing areas that may be less central operationally but still legally and societally sensitive.

Overall, the inherent risk assessment results reflect a risk landscape in which both **risk severity** and **risk volume** shape the platform's priorities. Categories with very high scores but fewer scenarios, such as **Data Privacy and Protection Risks** and **Privacy & Family Life**, call for focused attention on particularly severe harms. By contrast, higher-volume categories such as **Dissemination of Illegal Content** and **Protection of Minors** require scalable governance structures, moderation processes, compliance systems, and technical controls capable of remaining effective as platform operations expand in volume and complexity.

5.2. Risk Driver Assessment

Risk Driver Assessment Methodology

The assessment of risk drivers was carried out on a residual-risk basis. This approach was adopted because risk drivers represent the operational, technical, and governance factors referred to in Article 34(2) DSA, factors that may increase the likelihood or severity of systemic risks within systems and processes already operating on the platform. The assessment focuses on the level of exposure remaining after existing controls are considered, rather than on a hypothetical control-free situation.

For risk drivers, residual risk was assessed using a structured scoring model based on likelihood and impact, each rated from 1 to 5, with the resulting score expressed as ***Likelihood × Impact***. These scores were assigned qualitatively through expert assessment, considering the presence and maturity of existing safeguards. In other words, the model is score-based, but the underlying judgments remain qualitative.

Each driver was assigned an influence type to indicate whether it primarily affects the likelihood of systemic harm or the severity once it occurs. A driver-influence scale was then used to show whether the driver retained low, medium, or high residual relevance after current controls were considered. This structure supports a transparent and comparable assessment of how different operational factors continue to shape systemic-risk exposure.

To reflect the relative influence strength of each driver on the underlying systemic risks, a driver-modifier scale was applied. This scale allows small, evidence-based adjustments to the final residual score of systemic risks in cases where the driver demonstrably shifts either the likelihood or impact dimension despite the presence of established controls. The following adjustment bands were used:

- **Low influence:** The driver exerts negligible residual effect after controls; no adjustment applied.
- **Medium influence:** The driver retains a limited but measurable influence on either likelihood or impact, as supported by workshop evidence or observed performance patterns.
- **High influence:** The driver exerts a strong and persistent effect, either by significantly raising the probability of occurrence or by materially worsening potential consequences, even after mitigations are in place.

These modifiers were applied qualitatively and judgement-based, reflecting expert assessment. The purpose was to acknowledge observable differentials across drivers without implying precision beyond the level of organisational risk

tolerance or data availability. All adjustments were discussed during cross-functional workshops and recorded transparently in the Risk Drivers Register to ensure traceability and methodological consistency.

Risk Drivers Assessment Approach

A residual-risk assessment approach was used to evaluate the risk drivers, applying the same structure and process as the systemic inherent risk assessment (see Section 5.1, *Inherent Risk Assessment Approach*). In practice, both assessments were carried out together during the same cross-functional workshops. The workshops followed a two-part analytical sequence. In the first stage, teams reviewed systemic risk scenarios to align on the nature of potential harms. In the second stage, the same participants proceeded to the risk driver assessment. Teams based their judgements on direct operational experience.

Each driver was reviewed in terms of:

- The existing controls and measures already in place,
- The residual likelihood and impact of failure or underperformance,
- The influence type (whether the driver primarily affects likelihood or impact).

Following the workshops, a structured validation process was undertaken. The Compliance Team verified that each driver's rationale and influence type were logically consistent with the internal processes and related systemic risks.

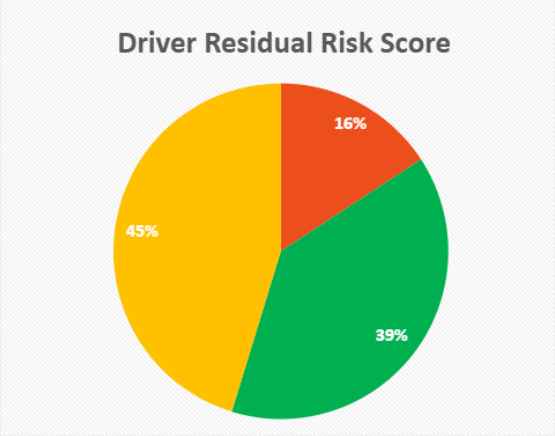
The Compliance Team also ensured that each risk driver was connected to the corresponding systemic risk entry in the *Systemic Risks Identification Catalogue*, which serves as a detailed analytical foundation for this assessment. This catalogue provides for every systemic risk linked risk drivers and related influence factors aligned with Art. 34(2) DSA.

Risk Driver Assessment Outcomes

Overall Distribution

The evaluation of 95 Risk Drivers was completed across all operational and governance domains. Across all categories, the majority of Risk Drivers were rated as medium (43 drivers; 45%) or low (37 drivers; 39%), while 15 drivers (16%) were rated as high (see Figure 4). This distribution indicates that most residual exposure remains within manageable low-to-medium bands, while a smaller group of higher-rated drivers remains concentrated in areas that are inherently more sensitive or operationally demanding. In practical terms, the Platform's control environment appears broadly functional and capable of containing a large share of residual exposure, although a targeted subset of drivers still warrants focused mitigation and continued attention.

Figure 4: Overview of Risk Drivers value



Source: WGCZ Risk Register, RA DASHBOARD

Concentration of Drivers

The distribution of residual risk across driver categories provides a useful indication of where residual exposure remains most concentrated within the control environment.

The largest concentration appears within Content Moderation, which accounts for 29 drivers in total, including 9 high, 12 medium, and 8 low. Recommender and Algorithmic Systems form the next-largest cluster with 13 drivers, all of which fall within the medium or low bands. TOS Obligations account for 11 drivers, again concentrated in the medium and low ranges. By contrast, Age-Assurance, although smaller in volume (6 drivers), shows a more elevated profile, with 5 of its 6 drivers rated high. Taken together, this means that the high-risk tail is concentrated primarily in Content Moderation and Age-Assurance, with only one additional high-rated driver appearing in Manipulation and Inauthentic Use.

The medium-rated drivers, which form the operational core of residual exposure, are concentrated primarily in Content Moderation, Recommender and Algorithmic Systems, and TOS Obligations. These categories point to areas where controls are already in place and functioning, but where implementation is not yet uniformly strong enough to reduce exposure fully into the low band. They therefore represent the main body of risks that should remain under active mitigation and monitoring, as improvements in these areas are likely to contribute most to a further reduction in overall residual exposure.

The low-rated drivers are more widely distributed across categories, including Advertising Systems, Regional and Linguistic Risks, Data-Related Practices, and parts of Content Moderation and Recommender and Algorithmic Systems. These drivers indicate that layered controls are already producing a stabilising effect across several parts of the framework. At the same time, their continued inclusion in the register remains important, as low residual risk in this context reflects effective control performance rather than the absence of relevant downside if those controls were to weaken.

Influence Dynamics

The analysis of influence type, namely whether a driver primarily affects the **likelihood** of an adverse event or the **impact** once such an event occurs, reveals a clear structural pattern in the platform's residual-risk profile.

Out of **95 total risk drivers**, 64 (67%) were classified as likelihood-oriented, while 31 (33%) were identified as impact-oriented. This distribution shows that the platform's residual-risk profile is shaped predominantly by operational and technical conditions that affect the probability of harmful events occurring, rather than by an equal split between probability and consequence.

Impact-type drivers include 2 high, 15 medium, and 14 low items. These drivers remain important because they influence the severity, visibility, and rights-related consequences of harmful events once they occur. Even though they are fewer in number, they are concentrated in areas where failures can intensify user harm, reduce procedural fairness, or increase regulatory sensitivity. Their profile suggests that consequence management remains a meaningful component of the framework, even if it is not the dominant source of residual exposure.

Likelihood-type drivers, comprising 13 high, 28 medium, and 23 low items, form the larger part of the risk landscape. This indicates that the principal source of residual exposure lies in the ongoing performance of operational mechanisms and technical controls. In other words, the main challenge is not only how the platform responds when harm materialises, but also how consistently existing systems prevent those events from arising in the first place. The fact that **13 of the 15 high-rated drivers** are likelihood-oriented further underlines that the current residual-risk profile is shaped primarily by drivers that increase the frequency or plausibility of adverse events.

Taken together, the analysis shows that **impact amplification** remains present but comparatively more contained, whereas **likelihood-related exposure** is broader and more structurally embedded across the control environment. In effect, the platform's residual-risk profile is defined less by an absence of safeguards in principle than by the varying strength and consistency with which existing safeguards reduce the probability of harm across key operational domains.

5.3. Mitigation Measures Assessment

Mitigation Measures Register and Assessment Framework

In 2025, WGCZ adopted a new component of its risk management system, the **Mitigation Measures Register**. It is designed to provide a cohesive, measure-centric view of how the platform addresses identified risk scenarios. Its key elements include:

- a systematic listing of mitigation measures, with each entry specifying the measure's name, type, purpose, and assigned owner;
- categorisation of each measure as **preventive** (aimed at preventing 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 in order to reduce 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 typically part of a broader approach), or **low** (limited impact due to technical or contextual constraints, narrow scope, or reliance on inconsistent user action); and
- explicit mapping of measures to specific risk scenarios, such as illegal content or data breaches, clearly demonstrating how each measure addresses the relevant risk.

Maintaining a single source of record for all mitigation measures ensures updates, like refining age-assurance approaches or deploying new moderation controls, are documented clearly and in a standardised format. This reduces duplication, inconsistent references, and mismatches that arose when the same measure was described differently across multiple risk entries.

Rather than focusing mainly on the probability and impact scores of individual risks, the measure-centric approach keeps operational teams focused on how effectively each measure performs in practice. For example, if moderation capacity expands but response times to flagged content do not improve, that discrepancy becomes immediately visible. This supports deeper evaluation of real-world performance and encourages continuous refinement.

The Mitigation Measures Register also improves transparency for regulators, auditors, and other stakeholders, who can assess WGCZ's mitigation framework through a single consolidated record rather than having to reconstruct the control environment from multiple separate documents. This is particularly beneficial for responses to formal information requests, transparency reporting, and external audit processes.

Mitigation Measure Assessment Methodology

The mitigation measures assessment focused on evaluating the controls currently in place. Each measure was assessed across three dimensions: **reasonableness, proportionality, and effectiveness.**

The reasonableness and proportionality ratings determined whether a measure was implemented appropriately, timely, and in a balanced way. These ratings were not directly included in the residual-risk calculation. In contrast, the effectiveness rating assessed the practical risk-mitigating capacity of each measure, and its results were used as input for the residual-risk assessment. Section 5.5 explains how these effectiveness results were reflected in the next stage of the analysis.

The assessment criteria were based on a combination of information derived from the EU preliminary findings on systemic risks and their mitigation, together with an established qualitative assessment framework. For each assessment dimension, a qualitative rating scale (**High / Moderate / Low**) was applied, supported by predefined guidance criteria. The following subsections outline the relevant assessment dimensions and the principles used to assign ratings.

Control effectiveness assessment

The assessment of control effectiveness was conducted qualitatively, taking into account a range of factors likely to influence a measure's 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 set out in Table 2.

Table 2: 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

Following the European Commission's expectations as formulated in a request for information, WGCZ is also working to gradually introduce quantitative assessment of mitigation measures based on predefined KPIs. However, at the time of this assessment, such an approach was feasible only for a limited number of measures, namely *Vercury, Advertisements review before publication, Advertisements review after publication, and the Advertisers certification program*, as these were the only measures for which relevant KPIs had already been established.

For that reason, the predefined effectiveness levels were used as the primary assessment framework, including when a formal KPI-based evaluation was not yet available.

During the workshop phase, potential KPI options were discussed with relevant teams to expand the range of measures. Where indicative or estimated performance data were available, they were taken into account; however, these estimates were used only as part of the qualitative assessment and were not treated as formal quantitative evidence.

WGCZ acknowledges that the formal KPI-based evaluation of mitigation measures is still being developed. This point is addressed in the Action Plan (see Section 6.2), which includes a dedicated action aimed at defining measurable KPIs for relevant mitigation measures in future assessment cycles. The progressive development of KPI frameworks should therefore be understood as an enhancement of the platform's evidential and quantitative assessment capacity, not as an indication that existing qualitative assessments or mitigation measures are defective.

Control Reasonableness assessment

The assessment of control reasonableness was conducted qualitatively, taking into account factors likely to influence whether a given measure is applied in a timely, structured, and predictable manner.

To ensure consistency and comparability across all assessed measures, a predefined rating scale was applied. This scale **categorises each measure's reasonableness as high, moderate, or low**, based on qualitative criteria reflecting the extent to which the measure is implemented without unnecessary delay and supported by established internal rules, relevant knowledge, and defined processes. The definitions are set out in Table 3.

Table 3: Control reasonableness scale

Reasonableness Level	Description
High	The measure is applied promptly and systematically, without unnecessary delays. Measure application relies on established internal rules, expert knowledge, and/or predefined processes rather than ad hoc judgment/random execution.
Moderate	The measure is generally applied in a timely manner, but with occasional delays or inconsistencies. Some steps or decisions rely on informal practices or limited documentation. Existing knowledge and rules are used, though not always comprehensively.
Low	The measure is applied irregularly or with significant delays. There are no clear procedures or consistent use of existing knowledge. Decisions depend mainly on individual discretion, leading to unpredictable outcomes.

Source: WGCZ Mitigation Measure Register, Dashboard

Control Proportionality assessment

The assessment of control proportionality was conducted qualitatively, taking into account factors likely to influence whether a given measure is appropriately calibrated to the platform's risk environment, operational capacity, and the need to respect fundamental rights.

To ensure consistency and comparability across all assessed measures, a predefined rating scale was applied. This scale categorises each measure's proportionality as high, moderate, or low, based on qualitative criteria reflecting whether the measure protects users and platform integrity while remaining appropriately balanced in relation to privacy, freedom of expression, and the platform's size and available resources. The definitions are set out in Table 4.

Table 4: Control proportionality scale

Proportionality Level	Description
High	The measure effectively protects users and platform integrity while fully respecting fundamental rights such as privacy and freedom of expression. Implementation is scaled appropriately to the platform's size and operational capacity.
Moderate	The measure contributes to user safety but may introduce minor unnecessary restrictions or gaps in rights protection. Resource allocation is adequate but not fully proportionate to the platform's scale.
Low	The measure may unduly restrict user rights or lack necessary resources for consistent application. Implementation is misaligned with the platform's size or operational needs.

Source: WGCZ Mitigation Measure Register, Dashboard

For each individual measure, two targeted assessment questions were developed for reasonableness and two for proportionality, based on the predefined rating levels described above. These questions were tailored to the specific purpose, scope, and operational context of the relevant measure.

Each question was assessed using a simple three-point response scale:

- **Yes** - 2 points
- **Partly** - 1 point
- **No** - 0 points

The combined score was then used to determine the final rating for the relevant assessment dimension, as follows:

- **4 points** – High Reasonableness / High Proportionality
- **2-3 points** – Moderate Reasonableness / Moderate Proportionality
- **0-1 points** – Low Reasonableness / Low Proportionality

This scoring method helped ensure that the assessment was carried out in a consistent, structured, and transparent manner across all evaluated measures.

Assessment Process

The assessment of mitigation measures was carried out to evaluate whether the platform's existing controls are appropriately designed, operationally embedded, and capable of reducing the inherent systemic risks identified in the assessment. For this purpose, the evaluation combined two complementary sources of input:

- desk research and data review; and
- cross-functional workshops.

The desk research and data review consisted of an analysis of relevant internal documentation, supporting operational materials, transparency reporting data, information gathered during the workshop phase, and, where relevant, complementary external studies. This review served as the evidentiary basis for the qualitative assessment of the identified measures.

The workshop phase was used to validate the identified mitigation measures, confirm their practical implementation, and assess them against the three evaluation dimensions outlined in this report: reasonableness, proportionality, and effectiveness. Workshops were conducted with the key operational and governance functions involved in the design, implementation, or oversight of the relevant controls, including:

- Content Moderation and Advanced Review Team;
- Notice and Complaint Team;
- Tech Team;
- Support Team;
- Advertising Team,
- Legal Team.

Within each workshop, the participating teams reviewed the measures falling within their respective areas of responsibility. With the Compliance Team acting as facilitator, the review process followed three main steps:

- first, the teams assessed whether the identified measures were operationally relevant and accurately reflected the controls currently in place;
- second, the adequacy of each measure was evaluated through targeted assessment questions tailored to the specific control, with particular focus on reasonableness and proportionality; and
- third, the teams assessed the effectiveness of each measure and discussed both existing and potential KPIs that may support a more quantitative assessment in future cycles.

The findings from these workstreams were subsequently consolidated by the Compliance Team and incorporated into the Mitigation Measures Register, which now serves as the central repository of the platform's mitigation measures together with their assessment results.

Results of the Mitigation Measure Assessment

Proportionality Assessment

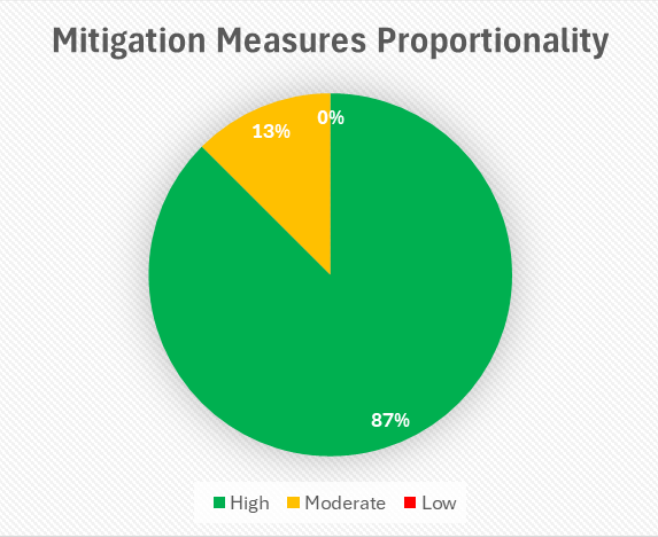
Similarly, the assessment outcomes in Figure 5 show that the platform's mitigation framework is overall strongly proportionate. **Most mitigation measures (49 out of 56) were assessed as highly proportionate.** These measures protect users and platform integrity while remaining appropriately calibrated to the platform's nature, and operational capacity, and while generally respecting fundamental rights such as privacy and freedom of expression. Measures in this category include, for example, Vercury, moderation team training, abuse reporting form for illegal or incompatible content, advertiser eligibility and certification controls, layered access protection for sensitive systems, and the ToS. In practice, these controls are not only relevant to the risks identified but are also implemented in a way that is balanced and operationally sustainable.

A smaller group of measures (7 out of 56) was assessed as having moderate proportionality. These measures still contribute meaningfully to user protection and risk reduction, but some practical limitations or trade-offs remain visible. Measures in this category include MD5 hash, reporting mechanism for comments, dedicated content removal request form, default content blurring on entry, network firewalls and access restrictions, cookie consent banner and preference controls, and account activation email and email validation.

Low proportionality was recorded for 0 measures. This is a positive indication that the platform's mitigation framework is broadly well balanced and that no currently assessed measure was found to create a clearly disproportionate burden in relation to the platform's operational context or the rights of users.

Overall, the results suggest that the platform's mitigation framework is appropriately calibrated. Relevant action points are included in the Action Plan (see Section 6.2) for measures where further refinement would strengthen the platform's risk-management framework, while taking into account both the assessment results and the platform's operational capabilities.

Figure 3: Overview of mitigation measures by proportionality



Source: WGCZ Mitigation Measure Register, Dashboard

Reasonableness Assessment

The results of the assessment, presented in Figure 6, indicate that **the majority of mitigation measures are highly reasonable (51 out of 56).** These measures are generally applied promptly and systematically, without unnecessary delay, and their implementation relies on established internal rules, expert knowledge, or predefined processes rather than ad hoc decision-making. Measures in this category include, for example, manual review of uploaded videos, moderation team training, structured moderation procedures, priority treatment of trusted flagger notices, strong password requirements for

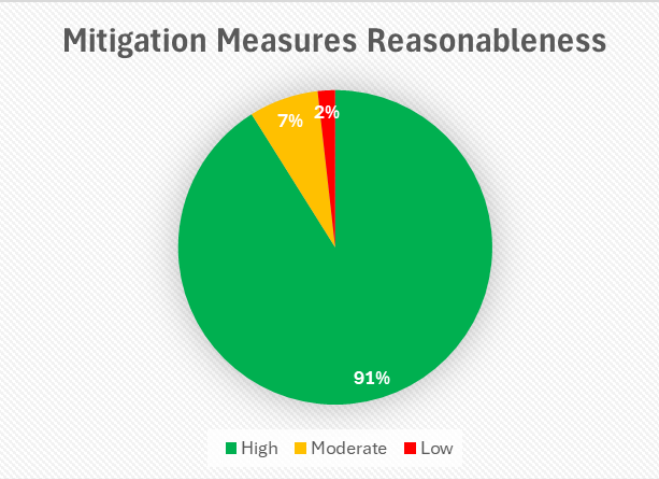
internal access, and pre-publication advertising review. In practice, this indicates that most of the platform's controls are operationally embedded and implemented in a structured and predictable manner.

Moderate reasonableness was identified for 4 measures. These measures are still generally applied in a timely and functional manner, but certain aspects of implementation are not yet fully standardized across all relevant processes. Measures in this category include reporting mechanism for comments, dedicated content removal request form, internal code review for security risk prevention, and guidance on parental controls and access restriction tools. In these cases, the relevant controls remain operationally relevant and broadly reliable, but their implementation still leaves some room for greater formalisation or consistency.

Low reasonableness was recorded for 1 measure, specifically the XVideos user verification process. Although this measure still performs an important control function, its implementation was assessed as less structured or predictable than that of many the platform's other measures.

Overall, the results show that the platform's risk-management framework is well structured and that most mitigation measures are implemented in a manner broadly consistent with regulatory expectations and operational feasibility. Relevant action points are included in the Action Plan (see Section 6.2) for measures that would further strengthen risk management, considering both the assessment results and the platform's operational capabilities.

Figure 6: Overview of mitigation measures by reasonableness



Source: WGCZ Mitigation Measure Register, Dashboard

Effectiveness Assessment

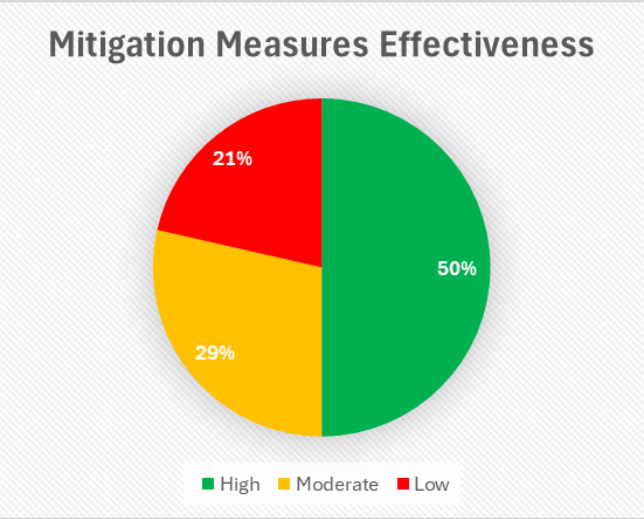
The assessment outcomes in Figure 7 show that **a substantial proportion of the platform's mitigation measures are highly effective**. Of the 56 measures assessed, 28 were rated as highly effective. These include, for example, Vercury, Hive, manual review of uploaded videos, moderation team training, structured moderation procedures, abuse reporting form for illegal or incompatible content, human review and decision-making for reported videos, pre-publication review of advertisements, ongoing post-publication ad monitoring and enforcement, advertiser eligibility and certification controls, and layered access protection for sensitive systems. Many of these measures are technically robust, consistently implemented, and supported either by automated enforcement or structured human oversight. As a result, they provide strong and reliable mitigation across the relevant risk areas.

Moderate effectiveness was recorded for 16 measures. These measures include, for example, Google SafetyNet API, keyword filtering and flagged-text review, reporting mechanism for comments, dedicated content removal request form, in-product reporting tool for suspected illegal or harmful content, cookie use and consent information, privacy policy and GDPR compliance framework, user controls over tracking and viewing-history settings, external vulnerability disclosure through HackerOne, adult-content access warning, and default content blurring on entry. These measures contribute meaningfully to risk reduction and are relevant parts of the overall control environment. At the same time, their effectiveness

is more conditional in nature, often because it depends on user engagement, limited scope, contextual factors, or the support of additional layers of control.

Low effectiveness was recorded only for 12 measures, representing only a smaller part of the overall mitigation framework. These include, for example, MD5 hash, Safer, ToS, recommender system criteria, review of recommender system algorithms, user-controlled content preferences, user-controlled content selection, and several measures within the Protection of Minors category, such as the age self-confirmation gate, ToS acceptance before access, and automated RTA metadata tagging. In general, these measures remain relevant, but their standalone mitigating effect is more limited due to their narrow scope, indirect impact, dependence on user behaviour, or reliance on other supporting measures to achieve stronger outcomes. This is particularly visible in the Protection of Minors area, where certain measures are designed primarily as warning, signalling, or supporting safeguards rather than as robust age-assurance mechanisms in themselves. Their lower effectiveness rating should therefore not be understood as meaning that they are without value; rather, they function as supporting or complementary components within the platform's broader layered control environment.

Figure 7: Overview of Mitigation Measures by Effectiveness



Source: WGCZ Mitigation Measure Register, Dashboard

5.4. Overview of Mitigation Measures in Place

a) Automated tools

Automated tools form the first technical detection layer within the mitigation framework. They are used to identify exact matches, risk indicators, and problematic content patterns at scale, thereby supporting early detection and routing higher-risk material for further review.

MD5 hash

Control Description: The platform uses MD5 hashing to generate a fixed-size digital fingerprint of files for integrity verification, duplicate detection, and change tracking. This mechanism is used to detect exact duplicates of known files.

Control Type: Detective

Vercury

Control Description: The platform uses Vercury as a fingerprint-based detection tool relying on an internal signature database. Video and image signatures are compared against the database and assigned a match score. Where a relevant match is identified, the content is routed for review, primarily in connection with potentially illegal material.

Control Type: Detective

Hive

Control Description: Hive is used to analyse images and videos through AI-based detection of behaviour, context, and potentially harmful activity within the scene. The tool is used mainly to identify content that may breach the Terms of Service, including categories such as alcohol, violence, or firearms. Depending on the result, the content may be blocked or flagged for review.

Control Type: Detective

Google SafetyNet API

Control Description: Google SafetyNet API is used to analyse images and assign a risk score, including with regard to potential underage indicators identified in the content. The tool is used to flag potentially harmful material for review, including content associated with malware or abusive material.

Control Type: Detective

Safer

Control Description: Safer maintains a fingerprint database of known content and is used to identify and flag matches against that database. The tool functions in a comparable manner to other automated risk-detection solutions and is used to flag potentially harmful or otherwise inappropriate content for review.

Control Type: Detective

Keyword filtering and flagged-text review

Control Description: Text content is screened through blacklist and grey list logic to identify problematic keywords and phrases. Blacklisted terms are blocked entirely, and the relevant content is not saved. Grey-listed terms trigger further review and analysis. Both lists are maintained and updated on an ongoing basis in real time.

Control Type: Detective

Table 5: Results of the effectiveness assessment of controls in place for the *Automated Tools* category

Control ID	Control Name	Effectiveness	Proportionality	Reasonableness	Justification
AT-01	MD5 hash	Low	Moderate	High	While MD5 hashing provides reliable identification of exact duplicates with minimal false positives, its detection scope is limited to identical files. Therefore, although technically precise, the measure's overall preventive impact is low due to its narrow detection range.
AT-02	Vercury	High	High	High	Although the system may generate false positives, it detects most of the known and derivative illegal content with a high degree of reliability. The combination of automated fingerprint comparison and human review ensures strong overall performance.
AT-03	Hive	High	High	High	According to information from Transparency reporting, data are currently not quantifiable, however there is approximately 90% error rate for problematic content and 10% of highly problematic Hive reports are confirmed to be correct upon review. Although false positives occur, Hive successfully identifies a wide range of incompatible content categories. Combined with human review, the system maintains a high overall effectiveness and reliability in preventing publication of prohibited material.
AT-04	Google SafetyNet API	Moderate	High	High	While SafetyNet provides reliable automated detection, its accuracy is limited in certain cases. Despite this, the measure contributes

					meaningfully to overall risk mitigation by flagging harmful uploads for human verification.
AT-05	Safer	Low	High	High	While Safer provides automated detection and coverage of known content, its accuracy remains limited. Nonetheless, it still serves as a supplementary control layer in the broader detection framework.
AT-06	Keyword filtering and flagged-text review	Moderate	High	High	Although the keyword search system operates automatically and is regularly updated, its effectiveness is limited by the potential for circumvention. Users may avoid detection by altering spelling or using alternative expressions. While the system provides valuable preliminary screening and supports manual review, its detection capability is inherently constrained to exact or closely matching terms.

b) Content moderation

Content moderation controls are designed to detect, assess, and respond to illegal, harmful, or policy-violating content through a combination of human review, reporting channels, procedural safeguards, and escalation mechanisms. Together, these measures support consistent enforcement while balancing user safety, legality, and procedural fairness.

Manual review of uploaded videos

Control Description: All videos selected by automated tools for manual human review are assessed by the review team. As part of that assessment, moderators verify, where relevant, the age of the user by examining physical characteristics and contextual details, as well as the legality of the content, copyright status, and the existence of participant consent. Depending on the outcome of the review, the content may be validated, ghosted, taken down, or deleted.

Control Type: Detective

Reporting mechanism for comments

Control Description: Users and third parties can report illegal content or content violating the ToS through the Report button available next to each comment. Once a comment report is submitted, the reporter receives an automatic email acknowledgment confirming successful submission of the report. The report is then handled through the relevant internal review process, without any subsequent follow-up communication on the outcome of the review.

Control Type: Reactive

Dedicated content removal request form

Control Description: Users and third parties can report illegal content or content violating the ToS through the dedicated Takedown amateur form accessible via the Privacy takedown form in the Content removal section of the webpage. After a report is submitted, the reporter receives an automatic email acknowledgment confirming that the submission has been successfully received. The report is then reviewed internally, and no further follow-up communication on the outcome of that review is provided afterwards.

Control Type: Reactive

Moderation team training

Control Description: Content moderators undergo regular training led by the team leader and experienced moderators, ensuring that both newly recruited and existing team members are guided by experienced professionals. Newly hired moderators complete a mentorship-based onboarding process under which they first observe experienced colleagues and then gradually take on moderation tasks under supervision. The training focuses on the responsible handling of sensitive content, the identification and treatment of illegal material, and familiarity with moderation procedures and internal classification methods.

Control Type: Preventive

Internal classification of manifestly illegal content and cooperation with reporting channels

Control Description: The platform applies an internal classification system to identify content considered undoubtedly illegal. Where such content is identified and flagged by a reviewer, the related data is automatically saved by the system. This database is made available to law enforcement agencies. The platform also cooperates with the official Czech reporting portal stoponline.cz operated by CZ.NIC, through which illegal content located outside the platform environment, and content that the review team has no authority or practical means to address directly, is reported by means of an established reporting form.

Control Type: Reactive

Translator tools

Control Description: To extend language coverage, the platform uses an integrated translation tool for languages that are not directly supported within the group. In more complex or nuanced cases, moderators verify translations through contextual checks, cross-referencing with online sources, and consultation with colleagues who are native speakers or who have sufficient proficiency in English.

Control Type: Preventive

Multilingual moderation coverage

Control Description: Moderators collectively provide coverage across a broad range of official EU and non-EU languages, allowing content, notices, and complaints from diverse groups of users to be reviewed effectively. As a minimum requirement, all moderators must demonstrate advanced proficiency in English.

Control Type: Preventive

Structured moderation procedures

Control Description: The platform applies structured moderation procedures to ensure that user-generated content is handled in a consistent, fair, and legally compliant manner. These procedures define the specific steps to be followed for content review, classification, and moderation decision-making.

Control Type: Preventive

Escalation policy for repeat illegal content offenders

Control Description: Users who repeatedly upload illegal content, including copyright-infringing material, are subject to a strike-based enforcement process following prior warnings. As a general rule, accounts are terminated after three strikes, unless a strike is later shown to have been issued in error. In particularly serious cases, including CSAM, NCII, terrorist content, or threats to life, the platform applies immediate deletion of the content and permanent account bans without reliance on the standard strike sequence.

Control Type: Preventive Reactive

Abuse reporting form for illegal or incompatible content

Control Description: Users and third parties may report content considered illegal or contrary to the ToS through the Abuse Reporting Form available in the Content Removal section. The form requires the reporting party to provide the exact electronic location of the content, the relevant category of the reported material, an explanation of why the content is believed to be illegal or in breach of the Terms of Service, contact details of the reporting party, and a statement confirming the reporting party's bona fide belief that the submitted information and allegations are accurate and complete. After submission, the reporter receives both an on-screen confirmation and an email confirmation including a link to the ticketing system and later receives a reasoned decision setting out the outcome, rationale, and available redress options.

Control Type: Reactive

Copyright reporting form for infringement notices

Control Description: Users and third parties may submit copyright-related notices through the Copyright Infringement Reporting Form. The form requires the URL of the reported content together with contextual information describing the

copyrighted material allegedly being infringed. Once the notice is submitted, the reporter receives on-screen and email confirmation of receipt and is subsequently provided with a reasoned decision including the outcome, the underlying rationale, and available redress options.

Control Type: Reactive

In-product reporting tool for suspected illegal or harmful content

Control Description: Each video includes a Report button allowing users to flag content suspected to be illegal or otherwise harmful without delay. Following submission, the reporter receives an automatic email acknowledgment confirming receipt of the report and containing a link to the reported video. A further email is then sent communicating the decision taken, the reasons for that decision, and access to available appeal mechanisms.

Control Type: Reactive

Priority treatment of trusted flagger notices

Control Description: Notices submitted by Trusted Flaggers are treated on a priority basis and reviewed with urgency. Such notices are specifically flagged within the workflow and are automatically moved to the top of the moderation queue for accelerated handling.

Control Type: Preventive

Dedicated registration process for Trusted Flaggers

Control Description: Trusted Flaggers may apply through a dedicated Trusted Flaggers registration page. Access to this registration process is limited to organisations established in the European Union, formally designated as trusted flaggers by the competent Digital Services Coordinator, and listed in the public database maintained by the European Commission. As part of the registration form, applicants must provide the organisation's name, registered office address, country of establishment, and the official email address recorded in the European Commission's Trusted Flaggers register. Once the application is submitted, the organisation receives both on-screen and email confirmation acknowledging receipt. After the registration request has been verified and approved, the applicant receives an activation email confirming account approval and containing a link for password creation.

Control Type: Preventive

Human review and decision-making for reported videos

Control Description: Videos reported by users or Trusted Flaggers are reviewed by the Notice and Complaint Team. As part of the review, the team assesses the notice and adopts an enforcement decision based on the outcome of the assessment. Where the uploader contests the decision, the team may request further clarifications from either the reporter or the uploader, including copyright-related documentation where relevant. Possible outcomes include validation, ghosting, takedown, or deletion. No automated tool is used in this process.

Control Type: Detective

Appeals and complaint-handling workflow

Control Description: The platform operates a complaint-handling system enabling recipients of moderation decisions affecting their content or accounts to challenge those decisions. Complaints may be submitted through a ticketing system available in the user dashboard or by email through the link included in the moderation notice. The ticketing system is used to log, track, and manage all communications in a transparent and organised manner. Each appeal is reviewed by a dedicated team, which evaluates the justification for the original moderation action in accordance with internal guidelines. Where an error is identified, corrective action is taken, including restoration of content or reversal of an account suspension. Information derived from appeals is also used to improve moderation systems over time.

Control Type: Corrective Reactive

Contact channel for Member States' authorities

Control Description: The platform has established a dedicated contact point for the receipt and processing of official removal orders concerning terrorist content issued by competent state authorities. This contact point is made available through a link in the ToS and through a dedicated platform page for authorities. Communications submitted through this channel may be made in English or Czech.

Control Type: Preventive

Anti-spam protection

Control Description: Mechanisms are applied to reduce automated spam activity, including the creation of spam accounts and the posting of unwanted or manipulative content. These controls are intended to limit abuse volume and reduce downstream moderation burden.

Control Type: Preventive

Table 6: Results of the effectiveness assessment of controls in place for the *Content Moderation* category

Control ID	Control Name	Effectiveness	Proportionality	Reasonableness	Justification
CM-01	Manual review of uploaded videos	High	High	High	Although no formal KPIs are defined, internal information shows that moderator error rates remain within a very low margin, reviewing content volumes that exceed daily uploads. The structured workflow, combining automated detection with human review demonstrates a high degree of effectiveness.
CM-02	Reporting mechanism for comments	Moderate	Moderate	Moderate	Although no KPIs are defined, the process is technically stable and proportionate, allowing all users to report freely and ensuring acknowledgment of submissions, but it does not yet close the feedback loop. The notice mechanism for reported comments provides only partial communication with users. While automated email acknowledgments confirm receipt of every report, reporters do not receive a follow-up message with the review outcome. This limits transparency and user feedback.
CM-03	Dedicated content removal request form	Moderate	Moderate	Moderate	Although no KPIs are defined, the process is technically stable and proportionate, allowing all users to report freely and ensuring acknowledgment of submissions, but it does not yet close the feedback loop. The notice mechanism for reported pictures provides only partial communication with users. While automated email acknowledgments confirm receipt of every report, reporters do not receive a follow-up message with the review outcome. This limits transparency and user feedback.
CM-04	Moderation team training	High	High	High	Although no formal KPIs are defined, the training program has demonstrated strong results. All moderators complete introductory training and supervised onboarding, which typically enables them to operate independently within a week. Sharing of uncertain cases within the team further strengthens consistency and reduces errors.
CM-05	Internal classification of manifestly illegal content and cooperation with reporting channels	High	High	High	The platform applies an internal classification system to identify content considered undoubtedly illegal. Where such content is identified and flagged by a reviewer, the related data is automatically saved by the system. This database is made available to law enforcement agencies. The platform also cooperates with the official Czech reporting portal stoponline.cz operated by CZ.NIC, through which illegal content located outside the platform environment, and content that the review team has no authority or practical

					means to address directly, is reported by means of an established reporting form.
CM-06	Translator tools	High	High	High	Although no formal KPIs are defined, translation tools are actively used daily. Integrated function makes the tool practical and reliable. Combined with cross-checking and consultation practices, this approach ensures that reports and content in multiple languages are handled effectively, maintaining accuracy and responsiveness in moderation.
CM-07	Multilingual moderation coverage	High	High	High	Although no formal KPIs are in place, the multilingual capacities of the team ensure that most European languages are covered. This setup enables accurate and timely moderation of multilingual content and user reports, making the measure highly effective in practice.
CM-08	Structured moderation procedures	High	High	High	Although no formal KPIs are in place, the structured and consistently applied moderation procedures ensure uniformity, fairness, and compliance. Reflection of legal and practical experience maintain their relevance and prevent procedural gaps. The system's predictability and consistency reduce the probability and severity of erroneous moderation actions effectively.
CM-09	Escalation policy for repeat illegal content offenders	High	High	High	Although no formal KPIs are set, the measure combines immediate strike notifications, automated account termination after the third strike, and permanent bans for severe violations. Users are directly informed when strikes are issued, which provides deterrence, while automation ensures consistent enforcement. These factors make the policy highly effective in addressing repeat violations.
CM-10	Abuse reporting form for illegal or incompatible content	High	High	High	Although no formal KPIs are set, mandatory data fields produce high-quality, actionable notices, while continuous availability enables timely reporting. Reviewers can gather additional information where needed, and the average 24-hour handling time supports rapid remediation. The reasoned-decision feedback loop strengthens process discipline.
CM-11	Copyright reporting form for infringement notices	High	High	High	Although no formal KPIs are in place, the combination of mandatory reporting fields, automated confirmations, and structured review ensures that copyright infringement reports are processed effectively. The reasoned-decision process with outcome and redress options provides transparency and consistency, making the measure highly effective in addressing copyright violations.
CM-12	In-product reporting tool for suspected illegal or harmful content	Moderate	High	High	The in-product report function is easily accessible and supports timely, structured notice submission. It materially contributes to risk reduction and to prompt review. Its effectiveness is nevertheless assessed as moderate because the current workflow does not yet provide users with full progress visibility through the ticketing system.

CM-13	Priority treatment of trusted flagger notices	High	High	High	Although no formal KPIs are set, the automated prioritisation mechanism guarantees that Trusted Flaggers' notices are placed at the top of the queue and promptly reviewed under established rules.
CM-14	Dedicated registration process for Trusted Flaggers	High	High	High	Although no KPIs are set, the registration system effectively ensures that only officially designated Trusted Flaggers gain access. Verification against the EC register and the requirement to use official contact details prevent misuse. Automated and manual elements of the process together ensure accuracy, transparency, and trustworthiness, making the measure highly effective in practice.
CM-15	Human review and decision-making for reported videos	High	High	High	Although no KPIs are set, immediate suspension of reported videos limits exposure, while a structured, guideline-driven review, typically completed in about 24 hours, supports accurate outcomes. The ability to seek clarifications improves decision quality, reducing both the probability of wrongful removals and the severity of harm from illegal content remaining online.
CM-16	Appeals and complaint-handling workflow	High	High	High	Although no KPIs are set, the combination of immediate access to appeal channels, a dedicated review team, comprehensive logging, and corrective actions where errors are found results in accurate outcomes. Peer consultation reduces the probability of persistent false positives. Feedback gathered through appeals is also used to refine moderation systems.
CM-17	Contact channel for Member States' authorities	High	High	High	Although no KPIs are set, the dedicated and prioritised channel ensures that removal orders from authorities are promptly received, verified, and enforced. The combination of continuous availability, verification procedures, and resource allocation makes this measure highly effective in practice.
CM-18	Anti-spam protection	High	High	High	Although no formal KPIs are defined, anti-spam controls are applied consistently in practice, are specifically targeted at automated abusive behaviour, and rely on repeatable operational processes. The measure is tailored to the relevant risk scenario and helps reduce abuse volume without unnecessarily affecting lawful users. On that basis, overall effectiveness is assessed as high, even though the supporting evidence is primarily operational rather than metric based.

ToS and Policies

ToS and policy controls provide the formal legal and informational framework governing access to the service, content rules, privacy, and consent-related practices. These measures support clarity, transparency, and consistent application of platform rules across the user base.

Platform's Terms of Service

Control Description: The platform applies ToS that establish the legal framework governing access to the service, account creation, user submissions, and content-related rules. The ToS set clear age restrictions, refer users to parental control tools, and expressly prohibit the submission of illegal, harmful, non-consensual, or terrorist content. They also regulate intellectual property matters, require transparency in sponsorship disclosures, and set out dispute resolution procedures. The ToS are available to all users through a dedicated link on the website, are translated into all official languages of the European Union and are preceded by a summary intended to help users understand their content. The ToS apply to all users without exception.

Control Type: Preventive

Cookie Use and Consent Information

Control Description: The platform provides a Cookie Policy explaining how cookies are used to ensure proper website functionality, enhance user experience, and support security and performance. The Policy distinguishes between essential cookies, which are necessary for the operation and security of the site, and non-essential cookies, which require user consent and are used for purposes such as personalization, advertising, and service improvement. Users are provided with clear information on the purposes, duration, and categories of cookies, as well as on how consent can be managed or withdrawn. The Policy also states that the data collected is not shared with unrelated third parties and is not transferred outside the EEA without adequate safeguards.

Control Type: Preventive

Privacy Policy and GDPR Compliance Framework

Control Description: The platform applies a Privacy Policy to defined user categories, namely users who create an account. The Policy describes how user information, including personal data, is collected, used, processed, and disclosed in connection with access to and use of the website. Personal data is processed only in accordance with the Privacy Policy and applicable legislation, particularly the GDPR. By accessing the website, the user acknowledges that they have read the Privacy Policy.

Control Type: Preventive

Table 7: Results of the effectiveness assessment of controls in place for the *ToS* category

Control ID	Control Name	Effectiveness	Proportionality	Reasonableness	Justification
LE-01	Platform's Terms of Service	Low	High	High	The ToS provide a clear legal basis for restricting access to adults only, requiring users to confirm that they are 18+ or of the age of majority and setting out core platform rules and responsibilities. This makes the measure highly reasonable and highly proportionate as a foundational, low-burden compliance tool. However, its standalone effect is limited because it depends mainly on user declarations and parental-control measures, rather than direct age-verification or other active enforcement at the point of access.
LE-02	Cookie Use and Consent Information	Moderate	High	High	The Cookie Policy clearly explains the use of cookies and similar technologies, requires prior express consent for non-essential cookies, allows users to manage or withdraw consent at any time, and is updated when cookie practices change. This supports a high assessment of reasonableness and proportionality, as the measure is transparent, standard, and non-intrusive. Its effectiveness is moderate because it gives users meaningful information and consent control, but its practical impact still depends on users actively engaging with the consent mechanism.

LE-03	Privacy Policy and GDPR Compliance Framework	Moderate	High	High	The Privacy Policy clearly explains the platform's data-processing framework, including the categories of data processed, purposes, legal bases, retention periods, security measures, and user rights. This supports high reasonableness and proportionality because the measure is transparent, structured, and provides users with practical control options. Its effectiveness is moderate because it improves transparency and user protection, but as a governance measure it still depends on users actively exercising the rights and controls made available to them.

c) Recommender and Algorithmic systems

Recommender system controls are intended to make content-discovery logic more transparent, more reviewable, and more responsive to user choice. These measures help reduce the risk of unwanted amplification, repetitive recommendation patterns, and insufficient user influence over suggested content.

Recommender system criteria

Control Description: The platform uses content recommendation logic based on factors such as geographic location, browsing history, and user-selected preferences. The criteria influence the ordering or display of suggested content and are applied to improve relevance while preserving user control over recommendation-related settings.

Control Type: Preventive

Review of recommender system algorithms

Control Description: The recommender system is subject to periodic internal review to assess transparency, fairness, consistency, and alignment with applicable regulatory requirements. The review examines the criteria used by the system, potential bias, and the impact of changes to recommendation logic.

Control Type: Detective

User-controlled content preferences

Control Description: Users can adjust certain recommendation-related preferences, such as location or preferred content categories, so that suggested content better reflects their selected settings.

Control Type: Preventive

User-controlled content selection

Control Description: Users can refine the content displayed to them by selecting categories, tags, performers, or keywords, allowing them to actively shape their browsing and recommendation experience.

Control Type: Preventive

Table 8: Results of the effectiveness assessment of controls in place for the *Recommender and Algorithmic systems* category

Control ID	Control Name	Effectiveness	Proportionality	Reasonableness	Justification
RS-01	Recommender system criteria	Low	High	High	The measure is consistently embedded in the recommendation logic and gives users a degree of control over how suggestions are shaped. However, its mitigation effect is indirect, and no specific indicators or case-based evidence have been identified to demonstrate measurable risk reduction. The measure is therefore considered of low overall effectiveness for mitigation purposes. This low rating reflects the current absence of a formal KPI framework or documented measurable evidence for this control, rather than a finding that the measure is non-functional or ineffective as part of the platform's recommender-system governance.
RS-02	Review of recommender system algorithms	Low	High	High	The measure provides an important governance and oversight function by supporting periodic assessment of recommendation logic. However, no specific indicators or documented follow-up outcomes have been identified to show measurable risk reduction over time. The measure is therefore considered of low overall effectiveness for mitigation purposes. This low rating reflects the current absence of a formal KPI framework or documented measurable evidence for this control, rather than a finding that the measure is non-functional or ineffective as part of the platform's recommender-system governance.
RS-04	User-controlled content preferences	Low	High	High	The measure gives users direct control over certain recommendation-related preferences and is clearly linked to how suggested content is personalized. However, its mitigation effect depends on user uptake, and no specific indicators or user-side evidence have been identified to demonstrate measurable risk reduction. The measure is therefore considered of low overall effectiveness for mitigation purposes. This low rating reflects the current absence of a formal KPI framework or documented user-side evidence demonstrating measurable risk reduction, rather than a finding that the measure is non-functional or ineffective as part of the platform's broader user-choice and recommender-system governance.

RS-05	User-controlled content selection	Low	High	High	The measure gives users active and granular control over the types of content shown to them and directly affects their individual browsing and recommendation experience. However, its mitigation effect depends on whether users actively use the feature, and no specific indicators or user-side evidence have been identified to demonstrate measurable risk reduction. The measure is therefore considered of low overall effectiveness for mitigation purposes. This low rating reflects the current absence of a formal KPI framework or documented user-side evidence demonstrating measurable risk reduction, rather than a finding that the measure is non-functional or ineffective as part of the platform's broader user-choice and recommender-system governance.

d) Data Privacy and Security

Data protection controls are intended to reduce privacy, security, and retention-related risks associated with user data and internal access to sensitive systems. These measures combine user-facing privacy controls with technical safeguards, access restrictions, and vulnerability-detection mechanisms.

User controls over tracking and viewing-history settings

Control Description: Users are provided with control over the collection and use of their data. Before first access to the homepage, users are prompted to set their cookie preferences, which can later be changed at any time through a dedicated link available at the bottom of the website. Users are also given the option to enable or disable viewing history. Where viewing history is deactivated, past activity is not used to influence future content recommendations.

Control Type: Preventive

Internal code review for security risk prevention

Control Description: The platform performs internal reviews of its codebase to identify vulnerabilities, implementation errors, and other issues that may create security risks. These reviews are carried out particularly before new features or updates are deployed and focus on detecting security-relevant bugs even where the code appears to function correctly.

Control Type: Detective

Layered access protection for sensitive systems

Control Description: A layered access control framework is applied to ensure that only authorised users can access sensitive systems, data, or functionalities. This includes VPN-based protection against unauthorised external network access, web-server password protection for server-level environments, and, for selected applications, additional application-level passwords or two-factor authentication. Access monitoring logs both successful and failed login attempts. Where failed or unusual access attempts are detected, automatic email notifications are sent so that administrators can verify or report potential unauthorised activity.

Control Type: Detective

Strong password requirements for internal access

Control Description: The platform applies an internal password policy requiring strong passwords for all accounts with access to systems and data.

Control Type: Preventive

Real-time warnings for weak password creation

Control Description: Users are warned when attempting to create a weak or easily guessable password. These warnings are displayed in real time during password creation to encourage the use of stronger credentials meeting the applicable security standard.

Control Type: Detective

Network firewalls and access restrictions

Control Description: Firewalls and related access control mechanisms are deployed to protect the platform's network and systems against unauthorised connections, malicious traffic, and external threats.

Control Type: Detective

External vulnerability disclosure through HackerOne

Control Description: The platform operates an external vulnerability reporting programme through which independent ethical hackers may identify and report weaknesses in the platform's systems, including vulnerabilities that could enable unauthorised access to data. The programme is conducted externally and does not provide participants with direct access to internal systems.

Control Type: Detective

Limited retention and deletion of IP address data

Control Description: The platform applies retention rules under which IP address data is stored only for a limited period. Standard IP address data is retained for one year and is then automatically deleted. Certain IP records may be retained for longer where they are linked to specific processes, including takedown-related matters. Automatic deletion routines are run weekly to remove outdated records.

Control Type: Preventive Reactive

Cookie consent banner and preference controls

Control Description: Before interacting with content on the platform, users are presented with a cookie banner offering three options: accept all cookies, manage cookie preferences, or reject all cookies. The banner also includes a brief explanation of how cookies are used and managed, together with a direct link to the full Cookie Policy.

Control Type: Preventive

Table 9: Results of the effectiveness assessment of controls in place for the *Data Privacy and Security* category

Control ID	Control Name	Effectiveness	Proportionality	Reasonableness	Justification
DP-01	User controls over tracking and viewing-history settings	Moderate	High	High	While users can easily set and modify cookie and tracking preferences at any time, practical effectiveness depends on user engagement with these settings.
DP-02	Internal code review for security risk prevention	Moderate	High	Moderate	The code review process is well established and prevents most security-related issues from reaching production. Occasional bugs identified post-launch indicate room for improvement.
DP-03	Layered access protection for sensitive systems	High	High	High	The multi-layered design provides strong protection against unauthorized access. The multiple verification steps reduce the probability of intrusion, making the measure highly effective in safeguarding sensitive internal systems.
DP-04	Strong password requirements for internal access	High	High	High	The fully automated enforcement of strong password rules ensures consistent application across all internal accounts. By enforcing standardized criteria, the measure

					effectively minimizes the probability of unauthorized access due to weak passwords.
DP-05	Real-time warnings for weak password creation	Low	High	High	Although the mechanism promotes stronger passwords through awareness, it does not enforce minimum complexity requirements beyond basic length. Users may still choose weak passwords despite warnings. The feature's preventive potential is therefore limited.
DP-06	Network firewalls and access restrictions	High	Moderate	High	The firewall and access control systems provide strong, automated protection against unauthorized access and external threats.
DP-07	External vulnerability disclosure through HackerOne	Moderate	High	High	The HackerOne program adds an important external layer of security review, increasing the probability of identifying overlooked vulnerabilities. However, because external testers cannot access internal systems or source code, the scope of findings is inherently limited.
DP-08	Limited retention and deletion of IP address data	High	High	High	The measure is technically robust, automated, and operates on a predictable schedule. The clear retention framework and weekly deletion ensure that no outdated IP data remains stored unnecessarily.
DP-09	Cookie consent banner and preference controls	Moderate	Moderate	High	The internal data show that around 53 % of users actively made a choice (38.6 % accepted cookies and 14.2 % rejected them), while roughly one-third closed the banner without selection. This indicates that the banner is displayed and interacted with. The cookie banner provides meaningful choice and transparency while maintaining site functionality for users who opt out of non-essential cookies. Minor technical limitations on outdated devices have minimal overall impact. The measure is therefore considered moderate effective.

e) Protection of Minors

Protection of minors' controls are designed to reduce the risk of underage access to adult content and to support parents or guardians in implementing access restrictions outside the platform environment. These measures primarily rely on entry warnings, age-related access gates, content labelling, and user guidance.

Guidance on parental controls and access restriction tools

Control Description: The platform provides a dedicated page containing practical information intended to help guardians prevent minors from accessing adult online material. A dedicated link to these instructions is also made available upon initial entry to the site, before content is displayed. The guidance covers the activation of technical restrictions at device and network level. It also includes links to filtering tools available in common search engines, including Google SafeSearch, Microsoft Bing search settings, and Yahoo SafeSearch, to reduce the appearance of adult services in search results. In addition, users are informed about kid-safe visual search tools, device-level content control options available in standard operating systems such as Windows, Android, and Apple environments, and the availability of network-level content filtering.

Control Type: Preventive

Adult-content access warning

Control Description: Upon a user's first visit, the platform displays a clear warning stating that the service contains content restricted to adults. Until the user explicitly confirms their age, access to the website remains blocked and the underlying content remains blurred.

Control Type: Preventive

Age self-confirmation gate

Control Description: Upon initial access, users are presented with a clear notice stating that entry to the website confirms that they are above the applicable legal age threshold. Users must actively confirm this statement before access is granted. Until such confirmation is provided, access remains blocked and the underlying content stays blurred.

Control Type: Preventive

ToS acceptance before access

Control Description: Upon first entry to the website, users are informed that access to the service is subject to acceptance of the Terms of Service. The page content remains blurred until the user provides explicit confirmation by selecting the designated acceptance button.

Control Type: Preventive

Default content blurring on entry

Control Description: The platform applies a default blurring effect to the front page so that explicit content is not immediately visible upon entry. Full access to the site and its content is granted only after the user takes an additional action, including age confirmation and acceptance of the applicable terms and conditions.

Control Type: Preventive

Automated RTA metadata tagging

Control Description: The platform applies the RTA label automatically across all relevant pages and domains. The RTA tag is embedded in the page metadata and applied uniformly to content made available through the service.

Control Type: Preventive

The Protection of Minors measures assessed in this section should be understood as components of a layered minor-protection architecture, not as individually complete age-assurance mechanisms. Some measures, in particular the age self-confirmation gate, Terms of Service acceptance before access, and automated RTA metadata tagging, are designed primarily to create entry friction, provide notice, establish an adult-only access perimeter, and support external filtering or parental-control ecosystems. Their lower standalone effectiveness ratings therefore reflect the limited ability of each individual measure to prevent determined circumvention on its own. They should not be read as a finding that the measures are without value or ineffective within the broader multi-layered mitigation system.

Table 10: Results of the effectiveness assessment of controls in place for the *Protection of Minors* category

Control ID	Control Name	Effectiveness	Proportionality	Reasonableness	Justification
PM-01	Guidance on parental controls and access restriction tools	Moderate	High	Moderate	External evidence indicates that parental control tools can help reduce minors' exposure to inappropriate online content, but their effectiveness depends heavily on parental awareness, digital skills, and consistent activation. Parental controls are not a stand-alone solution and work best when combined with active parental mediation. The measure is therefore justified as an enabling safeguard: it provides guardians with accessible, practical instructions and directs them to widely used device-, browser-, and network-level protections, thereby increasing the likelihood that such controls are used in practice. On that basis, the measure is reasonably assessed as moderately effective and proportionate, while recognising that its real-world impact depends on user uptake and configuration.[54]
PM-02	Adult-content access warning	Moderate	High	High	This measure is best characterised as a friction-based and anti-accidental-exposure safeguard rather than a standalone age-assurance mechanism. By keeping explicit material blurred until the user actively confirms entry, it can reduce immediate or accidental exposure and creates a deliberate click-through step before adult content becomes visible. This supports a high assessment of proportionality and reasonableness, as the measure is standard, low-burden and non-intrusive. Its effectiveness is moderate, not high, because warning and notice mechanisms may help reduce accidental exposure, but they do not reliably prevent intentional access by minors without stronger age-assurance controls.[55]

[54] Stoilova, M., Bulger, M., & Livingstone, S. (2024). *Do parental control tools fulfil family expectations for child protection? A rapid evidence review of the contexts and outcomes of use*. *Journal of Children and Media, 18*(1), 29–49. Full article: Do parental control tools fulfil family expectations for child protection? A rapid evidence review of the contexts and outcomes of use

[55] Turvey, J., McKay, D., Kaur, S. T., Castree, N., Chang, S., & Lim, M. S. C. (2024). *Exploring the feasibility and acceptability of technological interventions to prevent adolescents' exposure to online pornography: Qualitative research*. *JMIR Pediatrics and Parenting, 7*, e58684. JMIR Pediatrics and Parenting - Exploring the Feasibility and Acceptability of Technological Interventions to Prevent Adolescents' Exposure to Online Pornography: Qualitative Research

PM-03	Age self-confirmation gate	Low	High	High	The measure requires users to actively confirm that they are 18+ or of the age of majority before entering the platform, thereby establishing a clear adult-only access threshold and a formal user attestation at the point of entry. This supports a high assessment of reasonableness and proportionality, as the measure is standard, easy to implement, and minimally intrusive for users. Its effectiveness is nevertheless low because it relies entirely on truthful self-declaration and can be easily bypassed[56], so it operates primarily as a supporting rather than determinative safeguard. This rating reflects the limited age-assurance capacity of the measure when assessed in isolation. It should not be understood as meaning that the measure is without value within the broader layered framework. In combination with adult-content warnings, default blurring, ToS acceptance, parental-control guidance, and other safeguards, the age self-confirmation gate contributes to entry friction, formal adult-only signalling, and reduction of casual or accidental access.
PM-04	Terms of Service acceptance before access	Low	High	High	This measure functions primarily as a procedural notice and governance safeguard rather than as a standalone age-assurance mechanism. Requiring users to accept the ToS before entry improves formal notice, reinforces the adult-only nature of the service, and contributes to access friction and legal clarity. Its standalone effectiveness is low because acceptance of terms does not, by itself, verify age or reliably prevent intentional access by minors. The rating should therefore be read as a conservative assessment of the measure in isolation, not as a conclusion that it is ineffective within the platform's layered minor-protection framework. The measure remains highly proportionate and reasonable because it is standard, transparent, minimally intrusive, and supports the overall access-governance structure.[57]
PM-05	Default content blurring on entry	Moderate	Moderate	High	This measure is justified as a practical exposure-reduction control designed to prevent immediate and unintentional viewing of explicit content. Research on technologies addressing youth exposure to online pornography indicates that such tools are more useful for reducing accidental exposure than for preventing intentional access. This logic is also consistent with mainstream protective design patterns, where potentially sensitive content is blurred and requires an additional user action before being displayed. In the platform context, default

[56] Yao, Y., McCollum, S., Sun, Z., & Zhang, Y. (2025). Easy As Child's Play: An Empirical Study on Age Verification of Adult-Oriented Android Apps. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security 2025)* (pp. 21-39). USENIX Association. https://www.usenix.org/system/files/usenixsecurity25-yao-yifan.pdf

[57] European Commission, Directorate-General for Justice and Consumers. (2016). *Study on consumers' attitudes towards Terms and Conditions (T&Cs): Final report*. European Commission. https://commission.europa.eu/system/files/2018-03/terms_and_conditions_final_report_en.pdf

					blurring ensures that explicit material is not immediately visible on first entry and that users must complete further steps before viewing it. The measure is therefore appropriately described as moderately effective: it meaningfully reduces inadvertent first-view exposure, while not functioning as a complete barrier to determined access.[58]
PM-06	Automated RTA metadata tagging	Low	High	High	This measure is justified as a technical signalling safeguard that improves compatibility with parental-control and filtering ecosystems, rather than as an access barrier. The purpose of the RTA tag is to help parental-control software, device-level tools, and related filtering systems identify age-restricted content. Its standalone effectiveness is low because its real-world impact depends on whether third-party filtering tools are installed, enabled, and properly configured by parents or device administrators. The rating should therefore be understood as reflecting external dependency and limited direct enforceability, not the absence of value. Within the broader layered framework, automatic and consistent RTA tagging supports downstream filtering, reinforces the adult nature of the service, and contributes to a proportionate multi-level approach to reducing minors' exposure..[59]

f) Advertising Integrity

The platform applies advertising-related controls to ensure that commercial content is transparent, reviewable, and compliant with applicable legal and internal requirements. These measures are intended to reduce the risk of deceptive, harmful, or non-compliant advertising and to strengthen accountability across the advertising lifecycle.

Ad transparency notice for users

Control Description: Each advertisement displayed on the platform includes an information icon ("i"). When a user clicks on this icon, the advertisement is replaced by an "About This Ad" button. By selecting this button, the user is redirected to a separate window containing transparency information about the advertisement, including the advertiser's identity and the targeting parameters used for the delivery of the ad.

Control Type: Preventive

Pre-publication review

Control Description: All advertisements are subject to manual review before they can be approved and published. During this review, each element of the advertisement is checked against internal advertising rules and non-compliant ads are rejected. The review process is supported by several tools, including antivirus scanners, URL analysis tools, VPNs, and translation utilities. Internal rules define which advertising formats and categories are accepted, which must be labelled, and which are prohibited, including for example non-consensual content, violence, incest, scatophilia, malware, phishing, ransomware, prostitution, racism, and political advertising. For advertisements targeting users in the EU, the UK, or

[58] Turvey, J., McKay, D., Kaur, S. T., Castree, N., Chang, S., & Lim, M. S. C. (2024). *Exploring the feasibility and acceptability of technological interventions to prevent adolescents' exposure to online pornography: Qualitative research*. *JMIR Pediatrics and Parenting, 7*, e58684. JMIR Pediatrics and Parenting - Exploring the Feasibility and Acceptability of Technological Interventions to Prevent Adolescents' Exposure to Online Pornography: Qualitative Research

[59] Stoilova, M., Bulger, M., & Livingstone, S. (2024). *Do parental control tools fulfil family expectations for child protection? A rapid evidence review of the contexts and outcomes of use*. *Journal of Children and Media, 18*(1), 29–49. Full article: Do parental control tools fulfil family expectations for child protection? A rapid evidence review of the contexts and outcomes of use

California, additional checks are carried out to verify that the ads do not drop cookies, in order to ensure compliance with applicable data protection requirements. All submitted advertisements remain in pending status and are reviewed in chronological order.

Control Type: Detective

Ongoing post-publication ad monitoring and enforcement

Control Description: Once an advertisement has been approved, it remains subject to continuous scanning aimed at identifying policy violations. Where a potential violation is detected, an investigation is initiated. The platform applies a predefined tiered enforcement process depending on the severity of the breach. Minor violations result in labelling measures, more serious violations may lead to campaign rejection and suspension of repeat violators, and severe breaches result in immediate bans and permanent closure of the advertising account.

Control Type: Detective Corrective

Advertiser eligibility and certification controls

Control Description: The platform operates a certification programme designed to ensure that only qualified and compliant advertisers may access designated advertising zones. To obtain certification, advertisers must satisfy predefined eligibility requirements. Applications are reviewed through a formal assessment process that includes the evaluation of relevant risks and verification of compliance with applicable advertising rules. Certified advertisers are also reviewed on a recurring basis to confirm that they continue to meet the relevant standards. Where violations are identified, the platform may remove certification, suspend campaigns, or terminate the advertiser's account.

Control Type: Preventive

Advertising repository

Control Description: The platform maintains a repository of approved advertising materials and advertiser assets to support consistent review outcomes and to reduce the repeated approval of unsafe or non-compliant advertising content.

Control Type: Preventive

Impression-related advertising data

Control Description: Advertisements are accompanied by transparency information and impression-related data that support the identification, traceability, and review of advertising activity.

Control Type: Preventive

Table 11: Results of the effectiveness assessment of controls in place for the *Advertising Integrity* category

Control ID	Control Name	Effectiveness	Proportionality	Reasonableness	Justification
AD-01	Ad transparency notice for users	Moderate	High	High	The overall effectiveness of the measure is rated as moderate. The information symbol ("i") is displayed on all of advertisements, providing users with access to the "About This Ad" window, which offers transparency on advertiser's identity and the targeting parameters applied. This ensures a generally effective implementation of the transparency feature while maintaining user privacy. However, a small number of ad formats — menu links— currently don't display the information label. While this affects only a limited portion of ads, it prevents full coverage across the platform. Considering these factors, the measure achieves a high degree of implementation but not complete coverage, with an estimated performance of around 75–85%, corresponding to a moderate level of effectiveness.

AD-02	Pre-publication review	High	High	High	There are many metrics being tracked and KPIs are met. The only reserve is that some of these metrics are recorded at ad network level, instead of website level. This is not an issue while KPIs are met globally.
AD-03	Ongoing post-publication ad monitoring and enforcement	High	High	High	The measure is supported by a defined KPI framework focused on the speed of response to problematic advertisements identified after approval. Cases are expected to be investigated and resolved within 30 minutes of detection, and performance is monitored monthly together with the volume of post-approval ad violations. As the KPI targets are achieved at a level of 80% or above on a consistent monthly basis, and the overall trend indicates that post-approval violations are being kept to a minimum, the measure can be assessed as highly effective.
AD-04	Advertiser eligibility and certification controls	High	High	High	The measure is assessed as highly effective. The certification programme is intended to ensure that only compliant advertisers gain access to advertising zones and traffic, with a KPI target that at least 95% of certified advertisers remain problem-free. Over the last 12 months, only 1 out of 260 certified advertisers committed a violation and lost certification status, resulting in a compliance rate of 99.62%, which indicates strong performance in practice.
AD-05	Advertising repository	High	High	High	The measure is considered highly effective because it supports consistent review outcomes, reduces the risk of repeated approval of unsafe or non-compliant advertising content, and improves the efficiency of the review workflow. It is used as an operational reference by the Advertising Team and contributes to more reliable decision-making over time.
AD-06	Impression-related advertising data	High	High	High	The measure is considered highly effective because it improves traceability, supports advertising review, and enables better identification of advertising activity on the platform. The consistent availability of transparency and impression-related data strengthens compliance oversight and auditability. Based on internal performance observations, the measure can be considered to operate at a high level of effectiveness, with estimated performance above 80%.

g) User Support

User support controls provide a basic communication layer through which users can seek assistance, clarification, or service-related guidance. Although these measures are not primary moderation controls, they support accessibility, procedural clarity, and user communication.

Contact channels for user support

Control Description: Users may contact the Support Team either through the dedicated "Contact Us" form or through the dashboard path Contact Us → Content Support → Discussion. These channels allow users to submit both general enquiries and specific questions relating to the use of the service. The contact form is available on a dedicated page.

Control Type: Preventive

Table 12: Results of the effectiveness assessment of controls in place for the *User Support* category

Control ID	Control Name	Effectiveness	Proportionality	Reasonableness	Justification
US-01	Contact channels for user support	High	High	High	The contact channels are continuously available, easy to use, and integrated directly into the platform interface. The Support Team operates with adequate capacity.

h) User Verification

User verification controls are designed to reduce the risk of unauthorised uploads, impersonation, and publication of content without sufficient linkage to the uploader's identity, age, or rights to the material. These measures operate as an upstream safeguard before upload access is granted or maintained.

Channel verification

Control Description: Channels are subject to a verification process before being authorised to upload videos. This process requires submission of an official identity document together with content samples demonstrating that the uploader is the same person as the individual appearing in the video. Verification checks cover the uploader's identity, date of birth for age verification purposes, as well as consent and ownership in relation to the uploaded content. Only channels who successfully complete all verification checks may upload videos. The same verification logic also applies to newly registered XVideos uploader channels.

Control Type: Preventive Detective

User account verification procedure

Control Description: Users uploading videos through XVideos are required to complete verification before being allowed to upload. In the case of legacy XVideos users, verification is carried out through the submission of a profile picture together with a video upload, followed by a visual matching check intended to confirm authorship. Only users who successfully pass this verification process are authorised to upload videos.

Control Type: Preventive Detective

Account activation email and email validation

Control Description: Following registration, the user receives an account activation email which also contains the link for email validation. At the same time, users who only access the service for watching and commenting on videos are not required to complete uploader verification.

Control Type: Preventive Detective

Table 13: Results of the effectiveness assessment of controls in place for the *User Verification* category

Control ID	Control Name	Effectiveness	Proportionality	Reasonableness	Justification
UV-01	User Verification	High	High	High	The verification framework reduces the probability of illegal or non-consensual uploads by requiring ID-based verification and visual authorship matching.

UV-02	User account verification procedure	Moderate	High	High	The verification framework reduces the probability of illegal or non-consensual uploads by requiring visual authorship matching for older users. The effectiveness is limited given the more complex procedure applied on new XVideos uploaders (ID-based verification).
UV-04	Account activation email and email validation	Low	Moderate	Low	The verification mechanism provides only a minimal safeguard, as users' accounts are activated without completing the validation step.

5.5. Residual Risk Assessment

Residual risk is the level of risk remaining after applying mitigation measures and other controls. While inherent risk reflects baseline exposure with no controls, residual risk reflects the remaining exposure after considering the control environment and the ongoing influence of relevant risk drivers.

In this assessment, the residual risk of systemic-risk scenarios was determined through a structured qualitative assessment informed by quantitative inputs. Specifically, the final residual classification was based on:

- the inherent risk level of the relevant scenario;
- the assessed effectiveness of the mitigation measures linked to that scenario, including the extent to which they reduce the underlying risk exposure; and
- the continuing influence of relevant risk drivers, including whether those drivers continue primarily to affect likelihood or impact.

Where relevant, this assessment used structured analytical inputs such as the Risk Reduction Factor (RRF) and **the Driver Impact Score (DIS)** to support consistency and traceability across scenarios. However, the final residual-risk classification was not derived from a fully mechanical formula. Instead, it was determined through evidence-based qualitative judgment, supported by these inputs and validated through cross-functional review.

This approach reflects the fact that the same mitigation measure may operate with different practical effects depending on the nature of the scenario, the combination of other measures in place, and the broader operational context. Residual-risk assessment therefore remains structured and comparable while avoiding false precision in areas where expert judgment is still required.

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

Low Risk (Green):
The measures are sufficient, and the remaining exposure is limited. These risks are generally well contained under current operations, although they should continue to be reviewed periodically to ensure that the current level of control remains effective over time.

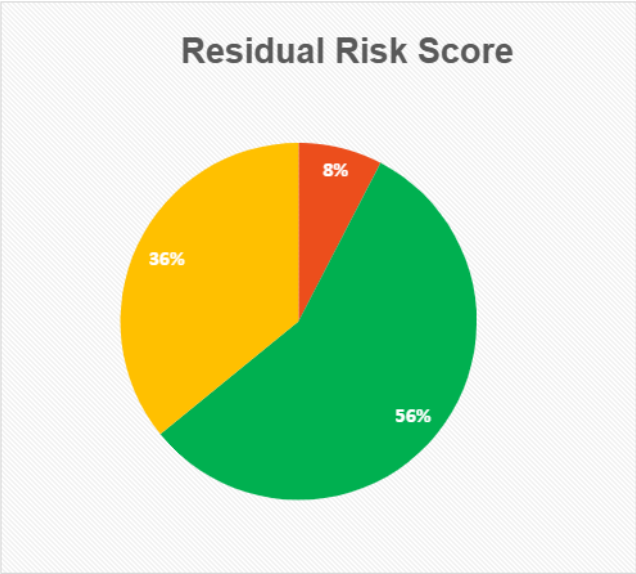
Medium Risk (Yellow):
Relevant controls have been implemented; however, a meaningful level of residual exposure persists due to practical limitations such as control coverage, scalability, responsiveness, or the specific features of the relevant scenario. These risks do not necessarily indicate control failure. In many cases, they reflect scenarios where mitigation measures already reduce exposure materially, but where some residual vulnerability remains because the harm is highly sensitive, adaptive, behavioural, or otherwise difficult to address fully through technical or procedural controls alone. A significant part of the medium residual-risk group consists of scenarios that were initially assessed as high inherent risks and were meaningfully reduced through layered mitigation measures, even if not to a low level.

High Risk (Red):

Existing controls reduce part of the exposure, but a high level of residual risk remains because the scenario involves particularly severe harm and the current measures cannot yet prevent, detect, or contain it in a sufficiently reliable and scalable way. In the current assessment, high residual-risk scenarios are concentrated in Protection of Minors and Dissemination of Illegal Content. Within the latter category, the remaining high-risk items are limited to a narrow subset of illegal-content scenarios involving private messages, especially CSAM-link sharing and grooming or coercive contact involving minors, rather than to private messaging as a feature in general. Within Protection of Minors, high residual risk remains primarily where minors bypass access safeguards and may still be exposed to explicit content, sexually explicit advertising, recommender-driven exposure, or user contact. These scenarios remain in the red zone because the remaining exposure is still material and warrants prioritised monitoring and further refinement of safeguards.

A high residual-risk classification should not be read as equivalent to a DSA infringement. It reflects the conservative outcome of the residual-risk methodology in scenarios where severe harm may remain possible despite existing safeguards, particularly because of circumvention, adversarial behaviour, private or downstream interactions, or structural limits on ex ante prevention. The classification therefore identifies priority areas for continued risk reduction and monitoring, not a conclusion that the existing control framework is non-functional or legally inadequate.

Figure 8: Results of residual risk assessment



Source: WGCZ Risk Register, RA DASHBOARD

The residual risk assessment shows a **substantial overall reduction in risk exposure** following the application of mitigation measures, even though some elevated residual exposure remains in a narrower group of more challenging scenarios. Of the **92 assessed scenarios**, **52 (57%)** were rated as **low residual risk**, **33 (36%)** as **medium residual risk**, and **7 (8%)** as **high residual risk**. This demonstrates that the measures currently in place materially reduce exposure across a large part of the overall risk landscape. At the same time, the results also show that a smaller number of scenarios continue to require heightened attention because the potential harm is particularly severe and because, in a limited subset of areas, current safeguards remain more dependent on behavioural, adversarial, or context-specific factors than in public-facing environments.

The medium residual risks can be grouped into two broad types:

Persistent medium risks (Inherent Risk = Medium → Residual Risk = Medium): Several scenarios remain at medium residual risk because they are structurally linked to the nature of the service and cannot realistically be reduced to a low level through technical or procedural controls alone. This applies to scenarios in **Public Health** and **Physical and Mental Well-being**, such as compulsive consumption patterns, harmful body-image effects, or the normalisation of unsafe sexual behaviour. It also applies to certain **consumer-protection** and **rights-related** scenarios, such as manipulative interface

practices, cybersecurity exposure affecting creators, or hidden or inconsistent account-status decisions. In these cases, mitigation measures are relevant and meaningful, but the residual exposure remains material because the harms are behavioural, cumulative, or dependent on broader social and contextual factors beyond the platform's full control.

De-escalated high risks (Inherent Risk = High → Residual Risk = Medium): A significant number of medium residual risks are de-escalated high risks. These are scenarios that were initially assessed as high at the inherent stage but were reduced to medium through layered mitigation measures. This group includes, among others, the upload of non-consensual intimate content, CSAM distribution through uploads, deepfake or exploitative sexual content targeting women, unlawful or non-consensual processing of sensitive data, doxxing, blackmail, and certain minors-related scenarios such as harmful discovery features or the possibility for minors to upload content. In these cases, the controls materially reduce likelihood or impact, but not to a level that would justify a low residual classification, given the seriousness of the potential harm and the continued possibility of circumvention, recurrence, or deliberate misuse. These scenarios therefore reflect meaningful mitigation progress while still requiring active oversight.

Unlike in the previous report, the current assessment also identifies a limited and clearly concentrated group of scenarios that remain at high residual risk. **All 7 high residual-risk scenarios** are concentrated in two categories: Protection of Minors and Dissemination of Illegal Content. In the Protection of Minors category, high residual risk remains where minors are able to access pornographic content, are systematically exposed to explicit content through content-delivery systems, are exposed to sexually explicit advertising after entering the platform, experience broader pornography-related harm to well-being, or are exposed to private messaging after accepting friend requests from other users. These scenarios remain at a high residual level because the current measures primarily reduce entry-level access, add friction, or provide warning and signalling, but cannot yet fully prevent downstream exposure once initial safeguards are bypassed.

In the **Dissemination of Illegal Content** category, the high residual scenarios are linked to the private messaging system. They concern the sharing of links leading to CSAM and the use of private messages for grooming or coercive contact with minors. As reflected in the register, the current safeguards in this area rely mainly on keyword signals, user-initiated reporting, and ex post enforcement, which are inherently more reactive in a private communication environment. For that reason, the mitigation effect for these scenarios has been assessed conservatively, and the residual risk remains high despite measures already being in place.

WGCZ also analysed residual risks across **13 predefined categories** in order to assess the concentration of residual exposure and the effectiveness of implemented controls from a category perspective. The analysis shows that three categories have been reduced entirely to low residual risk: Civic and Electoral Impact, Human Dignity, and Public Security Concerns. The remaining categories still contain medium or high residual risks, reflecting either the seriousness of the potential harm, the adaptive nature of abusive behaviour, or structural limits on how far current controls can reduce exposure given the nature of the service.

- **Consumer Protection**: This category includes 8 risk scenarios, of which 2 remain at medium residual risk. These scenarios relate to manipulative interface design or dark patterns and cybersecurity weaknesses that could expose sensitive user or creator data. Although relevant controls are in place, including transparency measures, user controls, policy documentation, and technical safeguards, residual risk remains because these issues are tied to interface design choices, user behaviour, and the continuing possibility of cyberattacks. The category therefore remains relevant from both a consumer-rights and user-trust perspective, even though most scenarios in this category have already been reduced to low residual risk.

- **Data Privacy and Protection Risks**: This category includes 3 risk scenarios, of which 2 remain at medium residual risk. The remaining elevated scenarios concern the profiling of users based on sensitive sexual data and the large-scale processing of such data without sufficiently valid consent. These risks were initially assessed as high inherent risks because of the seriousness of the rights impact and the sensitivity of the data involved. While privacy controls, consent tools, and GDPR-related compliance measures reduce the likelihood of harm, the residual exposure remains meaningful due to the scale of processing and the difficulty of fully eliminating misuse or unlawful secondary use.

- **Dissemination of Illegal Content**: This category contains the highest number of risk scenarios overall (26). 16 scenarios have been reduced to low residual risk, 8 remain at medium residual risk, and 2 remain at high residual risk. The medium residual scenarios include the upload of non-consensual intimate material, exploitative sexual content, CSAM through uploads, links in comments leading to externally hosted illegal material, and several

private-messaging scenarios where the controls still provide only partial coverage. The high residual scenarios are limited to CSAM-link sharing and grooming or coercive contact involving minors through private messages. Despite extensive controls, including automated detection, human moderation, reporting channels, and cooperation with external partners, residual exposure persists because malicious actors adapt their behaviour and these specific illegal-content scenarios are less amenable to preventive detection than public-facing content.

- **Freedom of Expression and Information**: This category includes 4 risk scenarios, of which 1 remains at medium residual risk. The remaining scenario concerns hidden or inconsistent decisions relating to suspension, demonetisation, de-indexing, or channel status, which may suppress lawful creator expression without sufficient explanation or redress. Although complaint-handling and moderation procedures are in place, residual risk remains because these decisions can still affect users' access to audiences and revenue in ways that may be difficult to predict or challenge effectively.

- **Gender-Based Violence**: This category includes 9 risk scenarios, of which 3 remain at medium residual risk. These scenarios concern extortion using stolen, coerced, or fabricated intimate material, the upload and circulation of non-consensual sexual content including deepfake pornography targeting women, and ongoing harassment or retaliatory targeting of women creators through comments, direct messages, fan interactions, impersonation, or piracy. These risks remain at medium residual level because they are highly harmful, often context-dependent, and can evolve quickly across both on-platform and off-platform channels. Existing moderation and reporting tools significantly reduce exposure but cannot fully eliminate the underlying pattern of abuse.

- **Non-Discrimination**: This category includes 6 risk scenarios, of which 1 remains at medium residual risk. The scenario concerns racialised or fetishising categories that may disproportionately expose certain groups to degrading stereotypes and reinforce systemic inequality. Although most identified risks in this category have been reduced to low residual risk, this scenario remains at medium because the harm arises not only from individual content items but also from repeated patterns of categorisation, framing, and exposure that may cumulatively reinforce discriminatory treatment or representation.

- **Physical and Mental Well-being**: This category includes 5 risk scenarios, of which all 5 remain at medium residual risk. These scenarios concern compulsive platform use, serious psychological harm resulting from the sharing of non-consensual intimate content, push-style advertising that may encourage compulsive consumption, and exposure to AI-generated content that promotes unrealistic body standards or sexual expectations. The residual risk remains medium because these harms are behavioural and psychosocial in nature and therefore cannot be fully mitigated through content moderation or policy controls alone. Even where safeguards are applied, the platform cannot fully control how users internalise or respond to sexualised content and advertising environments.

- **Privacy & Family Life**: This category includes 4 risk scenarios, of which 3 remain at medium residual risk. The medium-risk scenarios concern revenge pornography, doxxing, and the sharing of private material or information for the purpose of threats or blackmail. These were assessed as highly serious inherent risks due to their potential to cause severe harm to victims' private and family life. Existing controls, including moderation measures, reporting channels, and rules against prohibited content, reduce the likelihood of persistence and recirculation, but residual risk remains because these harms are often intentional, targeted, and difficult to reverse fully once intimate material or personal information has been disclosed.

- **Protection of Minors**: This category includes 11 risk scenarios, of which 6 remain at medium residual risk and 5 remain at high residual risk. The high residual scenarios concern minors accessing pornographic content, being systematically exposed to explicit content through platform delivery systems, being exposed to sexually explicit advertising, experiencing broader pornography-related harm to well-being, and being exposed to private messaging after accepting friend requests from other users. The medium residual scenarios include the upload or sharing of sexualised material involving minors, harmful exposure through discovery features, unwanted contact through comments, coercive or extortion-related harms, and the possibility that minors with regular accounts may upload content. This category remains one of the most difficult to mitigate fully because controls

such as warnings, access restrictions, classification measures, and account-related safeguards can reduce but not entirely prevent circumvention, downstream exposure, or misuse of communication features.

- **Public Health**: This category includes 5 risk scenarios, of which 2 remain at medium residual risk. These scenarios concern the promotion of unsafe sexual practices without appropriate disclaimers and the reinforcement of unrealistic or unhealthy body standards that may contribute to self-esteem problems or disordered behaviour. These risks remain at medium residual level because they are indirect and societal in nature. While WGCZ can apply content rules, warnings, and monitoring in some areas, it cannot fully determine how users interpret or absorb such messages, nor can it fully offset wider cultural influences that shape sexual norms and body-image expectations.

Overall, the category-level residual assessment confirms that most risk categories have been materially reduced through existing controls, while elevated residual exposure is concentrated in a relatively narrow set of especially sensitive areas. Protection of Minors and Dissemination of Illegal Content remain the only categories containing high residual risks, with the most serious scenarios linked to downstream minors' exposure and a narrow subset of CSAM- and grooming-related scenarios. The remaining medium residual risks generally reflect either severe harms that cannot currently be reduced further with confidence, or risks that are structurally difficult to eliminate fully through platform-level controls alone.

6. Risk Management Strategy

6.1. Risk Management Strategy

In accordance with the risk assessment methodology described in this report, WGCZ assigns a risk response strategy to each risk scenario based on its final residual risk classification. The available response strategies are as follows:

- Low residual risk (Score 1–5): **Accept**
- Medium residual risk (Score 6–15): **Accept / Disclose or Maintain / Monitor**
- High residual risk (Score 16–25): **Reduce**

For risk scenarios classified as **Low residual risk**, WGCZ applies the **Accept strategy**. Under this approach, the remaining exposure is considered within the platform's risk tolerance, and no additional risk-specific measures are currently required. This does not mean that the risk is disregarded. Rather, it reflects the conclusion that the existing control environment is currently sufficient. Such risks remain subject to periodic review to confirm that their likelihood and impact do not materially increase over time.

For risk scenarios classified as **Medium residual risk**, WGCZ applies either **Accept and Disclose or Maintain and Monitor**, depending on the nature of the risk and the extent to which the platform can reasonably influence it.

The Accept and Disclose strategy is used when the residual exposure arises primarily from external human behavior, broader societal dynamics, or other factors that cannot be materially reduced by additional proportionate measures by the platform. In such cases, WGCZ acknowledges the risk, discloses it where appropriate, and monitors complaints, user feedback, regulatory developments, and relevant research to identify any material change in the risk profile.

The Maintain and Monitor strategy is used where practical and proportionate mitigation measures have already been implemented, and no further reasonable reduction measures are currently available. In these cases, WGCZ maintains the existing control environment, monitors relevant indicators and developments, and reassesses the scenario where there is a material change in likelihood, impact, or operating context.

For risk scenarios classified as **High residual risk**, WGCZ applies the **Reduce strategy**. This indicates that the remaining exposure exceeds the platform's risk tolerance and requires further action. Under this strategy, WGCZ introduces or strengthens mitigation measures, including policy changes, technical safeguards, operational controls, training, and governance enhancements. Clear ownership, implementation steps, and review timelines are assigned to support effective implementation and follow-up.

Overall, this framework ensures that residual risks are managed proportionately, taking into account both the seriousness of the remaining exposure and the extent to which WGCZ can realistically influence the relevant underlying risk drivers.

6.2. Residual Risk Management Roadmap

Consistent with the previously described risk response strategies, the outcomes of the mitigation measures assessment, and the **2026-2027 Action Plan**, WGCZ has identified a set of follow-up actions for the current assessment cycle. These actions aim to address residual risks, enhance the evidential basis for future assessments, and maintain the platform's responsiveness to regulatory, technical, and operational developments.

The roadmap comprises the following elements:

- Cross-cutting actions that apply to multiple risk scenarios, including KPI development, external audits, research monitoring, and regulatory monitoring; and
- Risk-specific follow-up measures associated with individual medium- and high-residual-risk scenarios.

This approach ensures that the action plan addresses both systemic control improvements and scenario-level residual exposure in a structured and proportionate manner.

Table 14: Residual risk management roadmap actions

Risk Strategy	Description
Monitor EU guidance on interface design practices	The provider continues to monitor future EU guidance, emerging best practices, and interpretative materials that clarify the European Commission's expectations regarding interface design for various types of providers. This ongoing activity aims to ensure that the platform remains aligned with evolving supervisory expectations related to user choice architecture, consent design, and potentially manipulative interface practices. From a risk management perspective, the platform will track regulatory developments and use them to reassess whether existing measures remain sufficient or if additional actions are required.
Continuously monitor development of Inherent Risk Values	The provider continuously monitors the evolution of underlying risk trends, assessing whether the baseline likelihood or severity of specific risk scenarios changes in response to market developments, platform evolution, user behavior, or external evidence. When significant changes are detected, the platform conducts an ad hoc reassessment of the relevant scenario as appropriate. This approach ensures that risks are not regarded as static and that the risk register is updated when material developments necessitate further mitigation or a revised treatment strategy.
Continuously monitor development of Risk Drivers Values	The provider continuously monitors the development of specific factors that influence risk exposure, including scale, accessibility, discoverability, functionality, and patterns of misuse. When significant changes in these factors are detected, the platform will conduct an ad hoc reassessment of the relevant risk scenario as appropriate. This approach ensures that changes in underlying risk conditions are identified promptly and integrated into subsequent risk assessment and mitigation strategies.
Define KPIs for mitigation measures effectiveness assessment	The provider will strengthen the evidential basis for assessing residual risk and mitigation effectiveness by introducing, where appropriate, performance metrics or quantitative indicators linked to the operation of existing measures. The aim is to ensure a more robust, evidence-based assessment of mitigation effectiveness. Implementation will proceed in stages, initially prioritizing the most material risks and measures, while acknowledging that not all scenarios permit immediate or meaningful assessment based on key performance indicators.
Define KPIs framework for protection of minors	The provider will develop a dedicated KPI framework for protection of minors as a priority stream within the broader enhancement of its evidence-based risk-management approach. The purpose is to strengthen the evidential basis for assessing residual exposure and the performance of existing and future safeguards in one of the platform's most sensitive risk areas. This activity reflects the platform's intention to expand its quantitative assessment capacity in line with the evolving regulatory and technological environment for age assurance, child-safety controls, and minors' protection measures. The framework will be developed progressively, focusing first on indicators that are operationally meaningful, proportionate, and capable of supporting future reassessments.
Define KPIs framework for high-risk messaging	The platform will develop a dedicated KPI framework for high-risk illegal-content scenarios involving private communications, in particular CSAM-link sharing, grooming, coercive outreach, and comparable severe misuse patterns. This action is intended to strengthen evidence-led assessment and monitoring in a technically and legally sensitive area, where quantitative indicators must be designed in a proportionate manner and consistently with the legal framework governing private communications. The purpose is to improve the platform's ability to evaluate residual exposure and safeguard performance for clearly defined high-risk scenarios, rather than to characterise private messaging as a uniformly high-risk function in itself.
External audit of compliance with Article 25 of the DSA	The external audit under Article 37 DSA is expected to take place in March-April 2027. Upon receipt of the Independent Audit Report, the platform will prepare an Implementation Report that outlines the audit findings and planned follow-up actions. In this context, the review of compliance with Article 25 of the DSA will provide independent assurance on interface design and user-facing practices, including issues related to dark patterns and user autonomy. This process independently evaluates the adequacy of current controls and establishes a basis for additional remedial action if audit findings indicate the need for further mitigation.

External audit of compliance with Article 12, Article 16 of the DSA	The external audit under Article 37 DSA is expected to take place in March-April 2027. Upon receipt of the Independent Audit Report, the platform will prepare an Implementation Report that outlines the audit findings and planned follow-up actions. In relation to Articles 12 and 16 DSA, the audit will provide an independent assessment of the platform's notice-and-action and related procedural obligations. This helps verify whether the current governance and procedural framework remain adequate and provides a structured basis for follow-up improvements if deficiencies are identified.
External audit of content moderation practices and ToS enforcement	The external audit under Article 37 DSA is expected to take place in March-April 2027. Upon receipt of the Independent Audit Report, the platform will prepare an Implementation Report that outlines the audit findings and planned follow-up actions. In this context, the audit of content moderation practices and ToS enforcement will provide independent assurance that the provider's moderation systems, reviewer workflows, escalation mechanisms, and enforcement practices are functioning consistently and effectively. This activity is particularly relevant to risk scenarios involving illegal, harmful, gender-based violence, privacy violations, the protection of minors, and freedom of expression.
External audit of recommender systems practices	The external audit pursuant to Article 37 of the DSA is scheduled for March to April 2027. Following receipt of the Independent Audit Report, the platform will prepare an Implementation Report detailing the audit findings and proposed follow-up actions. In this context, the assessment of recommender system practices will provide independent assurance as to whether current recommendation and visibility mechanisms align with the platform's compliance and risk governance objectives. This process will validate the adequacy of existing systems and identify areas where further mitigation or design adjustments may be necessary.
External audit of compliance of advertising practices	The external audit pursuant to Article 37 of the DSA is scheduled for March to April 2027. Following receipt of the Independent Audit Report, the platform will prepare an Implementation Report detailing the audit findings and proposed follow-up actions. In relation to advertising practices, the audit will help assess whether current advertising-related processes and controls are aligned with the platform's legal and risk management obligations. This process will validate the adequacy of existing systems and identify areas where further mitigation or design adjustments may be necessary.
Research monitoring on violent content and gender-based violence	The provider will monitor research, policy developments, and emerging abuse patterns relating to violent sexual content, coercion, harassment, sexual extortion, and gender-based violence. This action supports gender-based violence risks by ensuring that risk ratings and future controls reflect current evidence and threat patterns.
Research monitoring on CSAM	The provider will monitor relevant research, threat developments, and evidence concerning CSAM, child exploitation, and related abuse patterns. This activity is part of broader monitoring of how risk trends and drivers evolve over time. Its outcomes will be reflected in the identification and assessment of risks in the next risk assessment report. This action ensures that one of the highest-severity harm areas remains under ongoing evidence-based review and that future reassessment is supported by current knowledge.
Research monitoring on impact of adult content on health and mental well-being	The provider will monitor research and evidence concerning the impact of adult content on health and mental well-being, including issues such as compulsive use, harmful norms, body image, self-esteem, and broader psychosocial effects. This activity supports ongoing risk monitoring and complements the platform's broader review of changes in inherent risk trends and drivers. Its findings will be reflected in the identification and assessment of risks in the subsequent risk assessment report.
Benchmarking of possible mitigation measures on private messaging	The provider will evaluate and compare potential technical and procedural interventions to reduce harms in private messaging. These interventions may include enhancements to reporting mechanisms, implementation of friction measures, account restrictions, or detection strategies. Such actions are particularly relevant to high-risk messaging scenarios, including solicitation, grooming, phishing, and CSAM-related contact, where direct mitigation is challenging and future design decisions necessitate systematic evaluation. Any such measures will be considered and assessed considering the currently applicable legal framework governing the monitoring of private communications.
In-depth analysis of research data and statistics to assess inherent risk	For selected risk scenarios, the provider will deepen its analysis of available research, statistics, operational evidence, and regulatory materials in order to further strengthen the evidential basis for inherent-risk scoring. This reflects the platform's continuing refinement of its assessment methodology and its intention to expand the quantitative and contextual basis for future assessments, rather than the correction of an identified defect. Where newer studies, reliable operational data, or relevant supervisory materials become available, they will be used to validate, contextualise, or recalibrate existing judgments and to support a more granular understanding of likelihood and impact.

Collaboration with NGOs focusing on age verification and protection of minors	The platform will engage with specialised NGOs and child-safety stakeholders to improve the design, proportionality, and effectiveness of age assurance and minors' protection measures. This action relates to protection of minors' risks, particularly those concerning access to adult content, harmful exposure, grooming pathways, and youth safety safeguards.
Cooperation with the EC on identifying a sufficient age assurance solution	This activity aligns with the European Commission's expectations that adult content platforms implement more robust age assurance measures to better prevent minors from accessing and being exposed to harmful content. Accordingly, the platform will cooperate with the European Commission and monitor relevant regulatory developments to identify the most appropriate age assurance solution. The objective is to establish an approach that is sufficiently effective to meet the Commission's expectations, while remaining proportionate and minimizing unnecessary interference with users' privacy and data protection rights. This initiative seeks to identify stronger structural safeguards in one of the highest-risk areas: the protection of minors.
Cooperation with law enforcement agencies	The provider will continue implementing processes to facilitate cooperation with law enforcement and competent authorities. These include appointing a dedicated point of contact and establishing internal workflows to receive, assess, and respond to notifications from competent EU authorities. These measures enable the platform to respond effectively and fulfil obligations to enforcement and regulatory bodies as part of systemic risk mitigation. From a risk management perspective, these actions reinforce existing escalation and response protocols and support procedures for high-severity scenarios requiring external intervention. This approach is critical for addressing risks involving criminal or highly harmful conduct such as CSAM, coercion, trafficking-related activity, sexual extortion, blackmail, and other situations where platform action alone is insufficient and law enforcement involvement is necessary.

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 medium and high residual risks identified in Section 5.5 (Residual Risk Assessment).

Table 15: Residual risk scenarios and action plan

		Final Risk Value (residual)	Risk Management Strategy	Measures to implement
Risk ID	**Risk Scenario**	**RR value**	**RMS**	**MTI**
SR-CP-03	The platform uses manipulative interface designs (dark patterns) that pressure users into subscriptions or into sharing personal data	Medium	Maintain / Monitor	• Monitor EU guidance on interface design practices • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • External audit of compliance with Article 25 of the DSA
SR-CP-09	Weak cybersecurity enables cyberattacks or hacking incidents to breach the website database and collect sensitive data from channel owners, leading to privacy breaches, mental harm, and possible blackmail	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment
SR-DP-01	Sensitive data concerning sexual identity or preferences is used for profiling, resulting in discrimination or unequal access to the platform	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment
SR-DP-02	Sensitive data is processed at scale without valid consent, undermining autonomy and enabling exploitation or unlawful secondary use	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment

SR-FE-04	Hidden or inconsistent decisions relating to suspension, demonetization, de-indexing, or channel status may suppress lawful creator expression and remove access to audience and revenue without meaningful explanation or redress	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • External audit of content moderation practices and ToS enforcement
SR-GB-04	Stolen, coerced, or fabricated intimate images are used to extort money or sexual acts from women, causing severe mental, emotional, and physical harm	Medium	Maintain / Monitor	• Research monitoring on violent content and gender-based violence • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-GB-05	Non-consensual sexual content, including AI-generated deepfake pornography targeting women, is uploaded, shared, and recirculated, violating privacy, dignity, and well-being	Medium	Maintain / Monitor	• Research monitoring on violent content and gender-based violence • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-GB-09	Women creators may be targeted through channel comments, direct messages, fan activity, impersonation, piracy, or retaliatory reporting linked to their profile or channel presence	Medium	Maintain / Monitor	• Research monitoring on violent content and gender-based violence • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • External audit of compliance with Article 12 , Article 16 of the DSA
SR-IC-01	Users upload intimate images or videos without consent (e.g., revenge pornography, deepfakes, or leaked paywalled content), sometimes including doxxing links	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-IC-02	Users upload prohibited sexual or exploitative material, including extreme pornography or content involving coercion or trafficking	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-IC-05	Users distribute CSAM through uploaded images or videos	Medium	Maintain / Monitor	• Research monitoring on CSAM • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for protection of minors • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-IC-06	Users post URLs, QR codes, or shortened links in comments that lead to externally hosted CSAM	Medium	Maintain / Monitor	• Research monitoring on CSAM • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for protection of minors • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-IC-20	Users advertise or solicit sexual services through private messages, including self-prostitution, self-promotion, or escort services	Medium	Maintain / Monitor	• Benchmarking of possible mitigation measures on private messaging • Define KPIs framework for high-risk messaging • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values

SR-IC-21	Traffickers or exploiters misuse private messages to advertise or solicit sexual services, including coerced postings for prostitution or escort services	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for high-risk messaging • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-IC-22	Users share URLs, QR codes, or links through private messages that lead to CSAM	High	Reduce	• Benchmarking of possible mitigation measures on private messaging • Define KPIs framework for high-risk messaging • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values
SR-IC-23	Users use private messages to groom minors, coerce explicit material, or arrange contact	High	Reduce	• Benchmarking of possible mitigation measures on private messaging • Define KPIs framework for high-risk messaging • In-depth analysis of research data and statistics to assess inherent risk
SR-IC-24	Users spread malware or phishing links through private messages, potentially compromising accounts or devices	Medium	Maintain / Monitor	• Benchmarking of possible mitigation measures on private messaging • Define KPIs framework for high-risk messaging • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values
SR-IC-25	Extremist groups distribute terrorist propaganda through private messages, glorifying or endorsing violent acts	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for high-risk messaging • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-ND-04	Racialised or fetishising categories disproportionately expose certain groups (e.g., Black or Asian performers) to degrading stereotypes, reinforcing systemic inequality	Medium	Maintain / Monitor	• External audit of recommender systems practices • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment
SR-PF-01	Users upload or share non-consensual intimate images or videos (“revenge pornography”), including leaked paywalled content, with the intention of harming a person’s private life	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-PF-02	Users doxx creators or victims by posting personal information (e.g., addresses, contact details, or family information) connected to explicit content	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-PF-03	Users share private content or information to threaten or blackmail individuals, deliberately harming their private and family life	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • External audit of content moderation practices and ToS enforcement • Cooperation with law enforcement agencies
SR-PH-01	Users or advertisers promote content or ads that encourage unprotected sex or risky sexual behaviours without appropriate disclaimers, normalising unsafe practices	Medium	Accept / Disclose	• Research monitoring on impact of adult content on health and mental well-being • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values

SR-PH-03	Exposure to content promoting unrealistic or unhealthy body standards contributes to self-esteem problems or disordered behaviours	Medium	Accept / Disclose	• Research monitoring on impact of adult content on health and mental well-being • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values
SR-PM-01	Minors access pornographic content on the service, resulting in exposure to harmful or illegal sexual material	High	Reduce	• Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution • Define KPIs framework for protection of minors
SR-PM-02	Users upload or share material depicting sexualised minors (CSAM) or images of minors in sexualised contexts	Medium	Maintain / Monitor	• Research monitoring on CSAM • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for protection of minors • External audit of content moderation practices and ToS enforcement
SR-PM-03	Minors are systematically exposed to explicit pornographic content through the platform's content delivery systems	High	Reduce	• Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution • Define KPIs framework for protection of minors
SR-PM-04	Minors encounter harmful sexual content through discovery features such as search suggestions, autocomplete, or trending tags	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for protection of minors • Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution
SR-PM-05	Minors who use comments are exposed to unwanted contact from unknown users, which may escalate into grooming	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for protection of minors • Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution
SR-PM-06	Minors are extorted through threats to publish intimate images or are coerced into producing sexual material	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for protection of minors • Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution
SR-PM-07	Minors are exposed to sexually explicit or pornographic ads (e.g., live sex cams, hookup services, sex toys, and sexual enhancement products).	High	Reduce	• Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution • Define KPIs framework for protection of minors
SR-PM-08	Exposure to pornography harms minors' well-being, contributing to anxiety, confusion, and distorted views of relationships and sexuality	High	Reduce	• Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution • Define KPIs framework for protection of minors
SR-PM-09	Minors with regular profiles, after accepting friend requests from other users, are exposed to private messaging, leading to encrypted contact with those users	High	Reduce	• Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution • Define KPIs framework for protection of minors

SR-PM-10	Minors are exposed to ads promoting false standards and facts about human body which can lead to self-doubt, mental health problems and body dysmorphia	Medium	Accept / Disclose	• Research monitoring on impact of adult content on health and mental well-being • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution
SR-PM-11	Minors with regular accounts are able to upload content on the platform, causing confusion or unintentional/intentional dissemination of CSAM	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for protection of minors • Collaboration with NGOs focusing on age verification and minor protection • Cooperation with the EC on identifying a sufficient age assurance solution
SR-PW-01	Users are exposed to violent or extreme sexual material, shaping harmful perceptions of sexuality and relationships	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs framework for protection of minors • Cooperation with law enforcement agencies
SR-PW-02	Extended platform use may lead to compulsive consumption patterns that harm users' health	Medium	Accept / Disclose	• Research monitoring on impact of adult content on health and mental well-being • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values
SR-PW-03	The sharing or distribution of non-consensual intimate content causes serious harm to victims, including psychological trauma, heightened anxiety, and depression	Medium	Maintain / Monitor	• Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • Define KPIs for mitigation measures effectiveness assessment • Cooperation with law enforcement agencies
SR-PW-04	Push-style ads (e.g., "live now", "exclusive access") encourage compulsive porn consumption and may harm users' mental health.	Medium	Accept / Disclose	• Research monitoring on impact of adult content on health and mental well-being • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values
SR-PW-05	Users are exposed to AI created videos, ads and pictures, promoting unnatural body standards and practises, leading to delusional body standards and sexual practises	Medium	Maintain / Monitor	• Research monitoring on impact of adult content on health and mental well-being • Continuously monitor development of Inherent Risk Values • Continuously monitor development of Risk Drivers Values • External audit of compliance of advertising practices