Digital Trust
& Safety Partnership

The Safe Framework Specification

June 2025

This specification is functionally identical to: ISO/IEC 25389

Digital Trust
& Safety Partnership

Table of Contents

Foreword

This document was prepared by the Digital Trust and Safety Partnership (DTSP) and drafted in accordance with its editorial rules.

Introduction

Digital services are increasingly central to our daily lives, facilitating social discourse, economic activity, and much more. These services provide powerful tools for users across the globe to engage in a wide range of valuable online activity. But like any tool, they can also be misused to facilitate harmful behaviour and content. Awareness of and action against this misuse has grown in recent years, which has led to increasing urgency in understanding, supporting, and evaluating effective ways to reduce harms associated with online content and behaviour, while also protecting people's ability to express themselves, carry out business, access information, associate, work, study, and participate in their communities through digital services.

Striking this balance presents a considerable challenge. To begin, there is no one-size-fits-all approach to handling online content and associated behavioural risks or, more generally, to organizations' trust and safety operations. Depending on the nature of the digital service, each may face unique risks relative to the various products or features they provide – different threats, different vulnerabilities, and different consequences. Products or features may engage with end users directly or indirectly, as well as with other services or businesses. What is an effective practice for one digital service may not suit another, and highly prescriptive or rigid approaches to defining trust and safety practices are likely to be too broad, too narrow or have negative unintended consequences. Further, risks change over time and so approaches to mitigating them must also have room to evolve.

Given the diversity of digital services, it is important to define an overall framework and set of aims for what constitutes a responsible approach to managing content- and conduct-related risks, to which digital services can then map their specific practices. This flexible approach has been deployed in other domains, such as information security, yet existing frameworks are not sufficiently concrete to be applied when it comes to addressing harmful behaviour and content online.

This document aims to fill this need by offering a framework of commitments to address content- and conduct-related risks. While the overarching commitments are uniform, the method by which they are fulfilled – whether by application of the illustrative practices in this document or alternatives – will vary by digital product or feature and evolve with both the challenges faced and advances made in the field of trust and safety.

This document also provides recommendations for organizations to evaluate the maturity of their implementation of these commitments through a rigorous and flexible approach to assessment.

This document is for the internal use of the organization responsible for trust and safety operations for a digital product or service. Recommendations for public reporting about the commitments and their implementation are outside the scope of this document.

This document is neither a management system standard, nor does it consider the issues of information security, privacy, and data management that are addressed by existing international standards.

Digital Trust and Safety Framework

1. Scope

This document provides a framework of recommendations for organizations that offer a public-facing digital product or service for which they conduct trust and safety operations to control or manage content- and conduct-related risks.

This document also includes recommendations for assessing the implementation of practices for addressing content- and conduct-related risks.

2. Normative references

There are no normative references in this document.

3. Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

- ISO Online browsing platform: available at https://www.iso.org/obp
- IEC Electropedia: available at https://www.electropedia.org/

3.1 abuse

use of a product or service in a way that violates the provider's product governance (3.13), generally because it creates or increases the risk of harm to a person or group or tends to undermine the purpose, function or quality of the service.

3.2 assessment

method to evaluate policies and operations for accuracy, changing user practices, emerging harms, effectiveness and process improvement

3.3 best practices

examples of practices embodying the commitments to product development, product governance, enforcement, improvement, and transparency

3.4 commitment

the actions taken by an organization to identify and address content- and conduct-related risk (3.8)

3.5 community guidelines

content policy

acceptable use policy

the set of conditions and limitations governing use of a digital service that a user must agree to as a condition of use.

Note to entry: These are generally written in plain and concrete language compared to legal language used in terms of service (3.18).

3.6 conduct

user behaviour facilitated by a digital product or service

Note to entry: behaviour may take place entirely online, or take place offline but is mediated by the use of the digital product or service.

3.7 content

text, images, audio and videos which are accessed by users via a digital product or service

Note to entry: content may be created or generated by other users, AI, publishers or other entities before being made available for a user to access.

3.8 content- and conduct-related risk(s)

the possibility of certain illegal, dangerous, or otherwise harmful content or behaviour, including risks to human rights, which are prohibited by relevant policies and terms of service

3.9 control

measure that maintains and/or modifies risk

Note 1 to entry: Risk controls include, but are not limited to, any process, policy, device, practice, or other conditions and/or actions which maintain and/or modify risk.

Note 2 to entry: Risk controls do not always exert the intended or assumed modifying effect.

[SOURCE:ISO 31073:2022, 3.3.33]

3.10 digital product

a product offered by one party to another party by means of digital hardware or software technology, or both, including communication over a network

3.11 digital service

a service offered by one party to another party by means of digital hardware or software technology, or both, including communication over a network

[SOURCE: ISO/IEC TS 5928:2023]

3.12 monthly active registered users

the number of users with a registered account who logged in or otherwise authenticated to visit the product website, mobile website, desktop or mobile application, within the last 30 days, from the date of measurement.

3.13 product governance

the set of agreements, rules, and guidelines mediating user interaction with the digital service and structuring conduct related to the product (examples include terms of service, privacy policy, community guidelines, content policy, acceptable use policy, codes of conduct, and any organizational processes by which these governing statements are created, adopted, or iterated).

3.14 question bank

set of questions that can be used by an organization to understand and identify the specific practices it uses to mitigate content- and conduct-related risks.

3.15 risk

effect of uncertainty on objectives

Note 1 to entry: In the context of this document, risk can be expressed as effect of uncertainty on implementation of commitments

Note 2 to entry: in the context of this document, risk is associated with content- and conduct-related risk (3.8)

[ISO/IEC 27000:2018, 3.61]

3.16 risk assessment

overall process of risk identification, risk analysis and risk evaluation

[SOURCE: 31073:2022, 3.3.8]

3.17 risk profile questionnaire

yes/no questions used to measure the extent to which product or service features and policies implicate content- or conduct-related risks (3.8).

3.18 terms of service

rules by which users agree to abide in order to use a service

[SOURCE ISO 32110:2023, 3.4.7]

3.19 trust and safety

The field and practices that manage challenges related to content- and conduct-related risk (3.8), including but not limited to consideration of safety-by-design, product governance (3.13), risk assessment (3.16), detection, and response, quality assurance, and transparency.

Note 1 to entry: "trust and safety" is a term used throughout the digital products and services industry that has a definition that is distinct from ISO/IEC definitions for trustworthiness and safety, which have been referenced in this document for completeness and to avoid confusion.

Note 2 to entry: for the ISO/IEC definition of trustworthiness, see ISO/IEC TS 5723.

Note 3 to entry: for the ISO/IEC definition of safety, see ISO/IEC Guide 51.

3.20 user volume

average monthly active registered users over the past twelve months.

4. Digital trust and safety

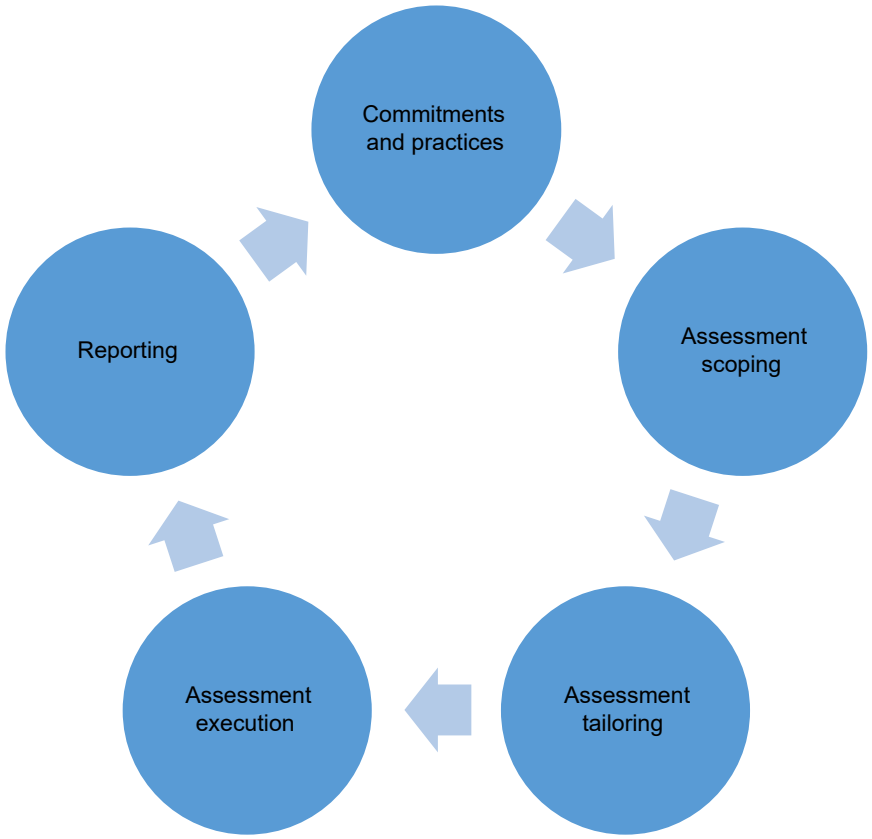
For an organization which provides a digital product or service, digital trust and safety refers to the part of that organization that focuses on understanding and addressing the harmful content or conduct potentially associated with that service.

Content- and conduct-related risks are sources of hazard and harm created by user generated content and conduct, which are distinct from other risks in digital products and services, which can be managed through existing international standards. These include but are not limited to information security, privacy, data management, and artificial intelligence.

Each organization is guided by its own values, product aims, digital tools, and human-led processes to make decisions about how to enable a broad range of human expression and conduct, while working to identify and prevent harmful content or conduct. Despite these individual approaches, a shared framework of best practices and assessments can help advance the development of industry best practices to ensure consumer trust and safety when using digital products and services.

The diagram shows how commitments, practices, and assessment are used by an organization to ensure digital trust and safety in a world of changing threats and regulatory expectations.

Figure 1 — The role of assessment in the development and ongoing maintenance of digital trust and safety



NOTE 1 International standards in the ISO/IEC 27000 series provide requirements and guidance for information security. ISO/IEC 29100 provides a high-level privacy framework. And for artificial intelligence, see ISO/IEC 22989 and the ISO/IEC 42000 series of standards relating to AI management systems.

NOTE 2: ISO/IEC Guide 51 provides guidelines for inclusion of safety aspects in standards. The Safe Framework, in the context of ISO/IEC Guide 51, is a group safety standard, "comprising safety aspects applicable to several products or systems, or a family of similar products or systems," applicable to digital products and services.

5. Commitments and practices

5.1 Overview

The organization should account for content- and conduct-related risk in the domains of product development, governance, enforcement, improvement, and transparency, and assign responsibilities and resources in each domain.

The organization should demonstrate its commitments through investment in and development of relevant personnel and technology; adoption of rights-respecting trust and safety principles and considerations in the development, governance, enforcement, and improvement of products; and the appropriate documentation of digital products and services.

Across the commitments, 35 best practices have been identified that are non-exhaustive examples of the kinds of activities and processes that an organization may have in place to mitigate risk and achieve tolerable risk for the service. These sample practices are summarized in the following Figure:

Figure 2 — DTSP Inventory of 35 Best Practices

Product Development	Product Governance	Product Enforcement	Product Improvement	Product Transparency
PD1: Abuse Pattern Analysis	PG1: Policies & Standards	PE1.1: Roles & Teams	PI1: Effectiveness Testing	PT1: Transparency Reports
PD2: Trust & Safety Consultation	PG2: User Focused Product Management	PE1.2: Operational Infrastructure	PI2: Process Alignment	PT2: Notice to Users
PD3: Accountability	PG3: Community Guidelines/Rules	PE1.3: Tooling	PI3: Resource Allocation	PT3: Complaint Intakes
PD4: Feature Evaluation	PG4: User Input	PE2: Training & Awareness	PI4: External Collaboration	PT4: Researcher & Academic Support
PD5: Risk Assessment	PG5: External Consultation	PE3: Wellness & Resilience	PI5: Remedy Mechanisms	PT5: In-Product Indicators
PD6: Pre-Launch Feedback	PG6: Document Interpretation	PE4: Advanced Detection		
PD7: Post-Launch Evaluation	PG7: Community Self Regulation	PE5: User Reporting		
PD8: User Feedback		PE6.1: Enforcement Prioritization		
PD9: User Controls		PE6.2: Appeals		
		PE6.3: External Reporting		
		PE7: Flagging Processes		
		PE8: Third Parties		
		PE9: Industry Partners		

A summary of the 35 best practices categorized by Commitment

5.2 Product development

The organization should identify, evaluate, and adjust for content- and conduct-related risks in product development.

The organization may consider the following examples of practices embodying a commitment to evaluate and adjust for content- and conduct-related risks in product development:

a. Developing insight and analysis capabilities to understand patterns of abuse and identify preventive mitigations that can be integrated into products;

b. Including trust and safety team or equivalent stakeholder in the product development process at an early stage, including through communication and meetings, soliciting and incorporating feedback as appropriate;

c. Designating a team or manager as accountable for integrating trust and safety feedback;

d. Evaluating trust and safety considerations of product features balancing useability and the ability to resist abuse;

e. Using in-house or third-party teams to conduct risk assessments to better understand potential risks;

f. Providing for ongoing pre-launch feedback related to trust and safety considerations;

g. Providing for post-launch evaluation by the team accountable for managing risks and those responsible for managing the product or in response to specific incidents;

h. Iterating product in light of trust and safety considerations including based on user feedback or other observed effects, including ensuring that the perspectives of minority and underrepresented communities are represented;

i. Adopting appropriate technical measures that help users to control their own product experience where appropriate (such as blocking or muting).

5.3 Product governance

The organization should adopt explainable processes for product governance including which team is responsible for creating rules, and how rules are evolved.

Product governance includes external and internal rules and processes by which an organization fosters certain activities and discourages others in relation to its product(s). This function exists in addition to compliance with and mitigation of risk related to applicable laws. One embodiment of product governance is an organization's terms of service (and for multi-product companies, sometimes multiple terms) — the

contract between users and the organization that sets forth underlying expectations and boundaries. Additionally, some organizations may maintain additional rules that more directly address acceptable conduct, often in more plain and concrete language. These may be called rules, community guidelines, acceptable use policies, or content policies. Their drafting and evolution may draw on user communities, or a combination of stakeholders with varied relationships to the organization.

The organization may consider the following examples of practices embodying a commitment to adopt explainable processes for product governance:

a. Establishing a team or function that develops, maintains, and updates the organization's corpus of content, conduct, or acceptable use policies;

b. Instituting processes for taking user considerations into account when drafting and updating relevant product governance;

c. Developing user-facing policy descriptions and explanations in easy-to-understand language;

d. Creating mechanisms to incorporate user community input and user research into policy rules;

e. Working with recognized third-party civil society groups and experts for input on policies;

f. Documenting for internal use the interpretation of policy rules and their application based on precedent or other forms of investigation, research, and analysis;

g. Facilitating self-regulation by the user or community to occur where appropriate, for example by providing forums for community-led governance or tools for community moderation, and finding opportunities to educate users on policies, for example, when they violate the rules.

5.4 Product enforcement

The organization should conduct enforcement operations to implement product governance.

The organization may consider the following examples of practices embodying a commitment to conduct enforcement operations to implement product governance:

a. Ensuring the organization has personnel and technological infrastructure to manage content- and conduct-related risks, to which end the organization may:

   - Constitute roles and teams within the organization accountable for policy creation, evaluation, implementation, and operations;

- Develop and review operational infrastructure facilitating the sorting of reports of violations and escalation paths for more complex issues;
- Determine how technology tools related to trust and safety will be provisioned (i.e., build, buy, adapt, collaborate);

b. Formalizing training and awareness programs to keep pace with dynamic online content and related issues, to inform the design of associated solutions;

c. Investing in wellness and resilience of teams dealing with sensitive materials, such as tools and processes to reduce exposure, employee training, rotations on/off content review, and benefits like counselling;

d. Where feasible and appropriate, identifying areas where advance detection, and potentially intervention, is warranted;

e. Implementing method(s) by which content, conduct, or a user account can be easily reported as potentially violating policy (such as in-product reporting flows, easily findable forms, or designated email address);

f. Operationalizing enforcement actions at scale where:

- Standards are set for timely response and prioritization based on factors including the context of the product, the nature, urgency, and scope of potential harm, likely efficacy of intervention, and source of report;
- Appeals of decisions or other appropriate access to remedy are available;
- Appropriate reporting is done outside the organization, such as to law enforcement, in cases of credible and imminent threat to life;

g. Ensuring relevant processes exist that enable users or others to "flag" or report content, conduct, or a user account as potentially violating policy, and enforcement options on that basis;

h. Working with recognized third parties (such as qualified fact checkers or human rights groups) to identify meaningful enforcement responses;

i. Working with industry partners and others to share useful information about risks, where consistent with legal obligations and security best practices.

5.5 Product improvement

The organization should assess and improve processes associated with content- and conduct-related risks.

The organization may consider the following examples of practices embodying a commitment to regularly assess and improve processes associated with content- and conduct-related risks:

a. Developing assessment methods to evaluate policies and operations for accuracy, changing user practices, emerging harms, effectiveness and process improvement;

b. Establishing processes to ensure policies and operations align with these commitments;

c. Using risk assessments to determine allocation of resources for emerging content- and conduct-related risks;

d. Fostering communication pathways between the organization on the one hand, and users and other stakeholders (such as civil society and human rights groups) to update on developments, and gather feedback about the social impact of product and areas to improve;

e. Establishing appropriate remedy mechanisms for users that have been directly affected by moderation decisions such as content removal, account suspension or termination.

5.6 Product transparency

The organization should ensure that appropriate trust and safety policies are published to the public, and report periodically to the public and other stakeholders regarding actions taken.

The organization may consider the following examples of practices embodying a commitment to publishing and reporting on relevant trust and safety policies:

a. Publishing periodic transparency reports including data on salient risks and relevant enforcement practices, which may cover areas including abuses reported, processed, and acted on, and data requests processed and fulfilled;

b. Providing notice to users whose content or conduct is at issue in an enforcement action (with relevant exceptions, such as legal prohibition or prevention of further harm);

c. Logging incoming complaints, decisions, and enforcement actions according to relevant data policies;

d. Creating processes for supporting academic and other researchers working on relevant subject matter (to the extent permitted by relevant law and consistent with relevant security and privacy standards, as well as business considerations, such as trade secrets);

e. Where appropriate, creating in-product indicators of enforcement actions taken, including broad public notice (e.g., icon noting removed content providing certain details), and updates to users who reported violating content and access to remedies.

6. Assessment framework

6.1 Overview

Organizations may use as necessary the assessment framework to evaluate relevant people, processes, and technology that contribute to managing content- and conduct-related risks and that reflect existing commitments and practices.

Assessment of a product or service through the framework takes place in 3 phases: scoping, tailoring, and executing.

6.2 Scoping

The organization can choose to apply the assessment framework to people, processes, and technology that contribute to managing content- and conduct-related risks and that reflect existing practices.

Each assessment should cover one public-facing digital product or service for which the organization conducts trust and safety operations, or otherwise implements controls to manage content- or conduct-related risks.

6.3 Tailoring

Given the diverse nature of the organizations and the digital services provided, there is no "one-size-fits-all" approach to conducting assessments against the commitments and best practices.

Organizations may tailor the depth of self-assessment to be proportionate to the particular risks and nuances of the product or service being assessed.

There are three proposed levels of assessment, referred to simply as "Level 1" (or L1), "Level 2" (or L2), and "Level 3" (or L3) that an organization may undertake to examine trust and safety practices in support of a particular product, digital service, or function. The Level 3 assessment is designed as the most in-depth in terms of the breadth and depth of assessment procedures, while Level 1 is less detailed and provides for more summary-level analysis.

Applying the tailoring framework, each organization can determine whether a Level 1, Level 2, or Level 3 assessment should be performed for a particular product or service during the assessment execution phase.

6.3.1 Tailoring methodology

An organization performs an assessment exercise for their practices related to the five commitments for the service or product to be assessed. Multiple assessments may be required if the organizations offer multiple products or services.

Application of the tailoring framework consists of four steps:

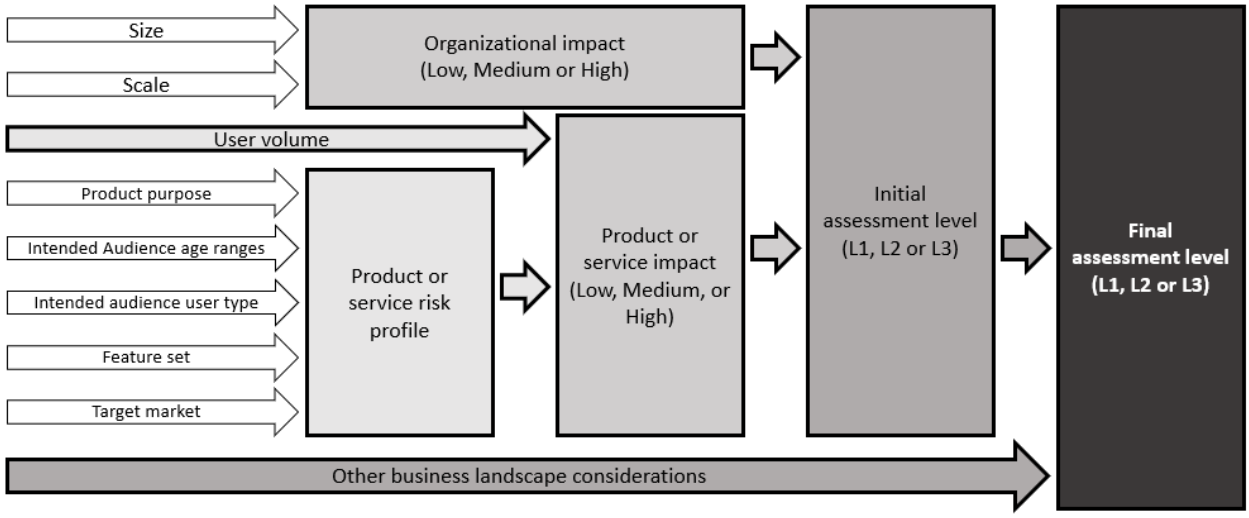
a. Evaluating the organization's size and scale;

b. Evaluating the impact of the product or digital service;

c. Determining the initial recommended assessment level;

d. Factoring additional business landscape considerations.

Figure 3 — Elements of the tailoring framework



6.3.2 Evaluating the organization's size and scale

It is important to establish a set of objective criteria for determining the size and scale of an organization. This component of the tailoring framework defines inputs for consideration that are indicative of an organization's size and scale:

a. Scale - Previous year's revenue (in Euros);

b. Size - Total number of employees for products/services in scope of assessment.

Together, these inputs are measured to categorize each organization into a "low", "medium", and "high" classification. The table below provides the proposed thresholds for defining these buckets and are subject to refinement based on the initial self-assessment process.

TABLE 1 — ORGANIZATIONAL SIZE AND SCALE INPUTS AND PROPOSED THRESHOLDS FOR ORGANIZATIONAL SIZE AND SCALE CLASSIFICATION

Input	Low	Medium	High
Scale - Previous year's total revenue (in Euros)	< €25B	€25B - €100B	> €100B
Size - Total number of employees	< 10,000	10,000 - 100,000	> 100,000

If either of the Inputs are High, then the overall categorization of the organization will be High. If both Inputs are Low, then the overall categorization of the organization will be Low. Otherwise, the overall categorization of the organization will be Medium.

See Annex A (informative) for illustrative examples of this categorization.

6.3.3 Evaluating the impact of the product or digital service

Once organizations have applied the size and scale criteria, they evaluate their product or digital service risk drivers on two axes: user volume and risk profile.

User volume is measured as the average monthly active registered users over the past twelve months. The broader the audience consuming the content or services of the product, the greater the impact of content- and conduct-related risks.

Risk profile, measured by a Risk Profile Questionnaire, is used to measure the extent to which a product or services features and policies implicate content- or conduct-related risks. Certain features, such as live streaming or video sharing or hosting of user-generated content, can expand the risk landscape. In general, the more of these features that a product makes available to users, the more complex and broader the set of risks.

The following yes/no questions are used to develop the risk profile. The questions are organized into five groupings related to

a. Product purpose;

b. Intended audience age ranges;

c. Intended audience user type;

d. Feature set;

e. Target market.

See Annex B (informative) for the risk profile questionnaire.

Each of the inputs is measured on a "low", "medium", and "high" scale. The proposed measurement thresholds for these inputs are outlined in the figure below.

TABLE 2 — THRESHOLDS FOR PRODUCT/DIGITAL SERVICE IMPACT INPUTS

Input	Low	Medium	High
User Volume Avg. monthly active registered users in the past 12 months	<100 mil	100-500 mil	>500 mil
Risk Profile Questionnaire Number of yes answers	<10	10-15	>15

The measurements for each of the inputs are aggregated to determine an overall categorization of the product/service's impact. The more inputs that are measured as "high", the higher the impact or implied risk that may be attributed to the product or service.

If either of the Inputs are High, then the overall categorization of the organization will be High. If both Inputs are Low, then the overall categorization of the organization will be Low. Otherwise, the overall categorization of the organization will be Medium.

See Annex A (informative) for illustrative examples of this categorization.

The resulting categorization derived in this step, along with the previous step (evaluating organization size and scale) are combined to determine the initial recommended assessment level (explained in the next section, Section 6.3.4).

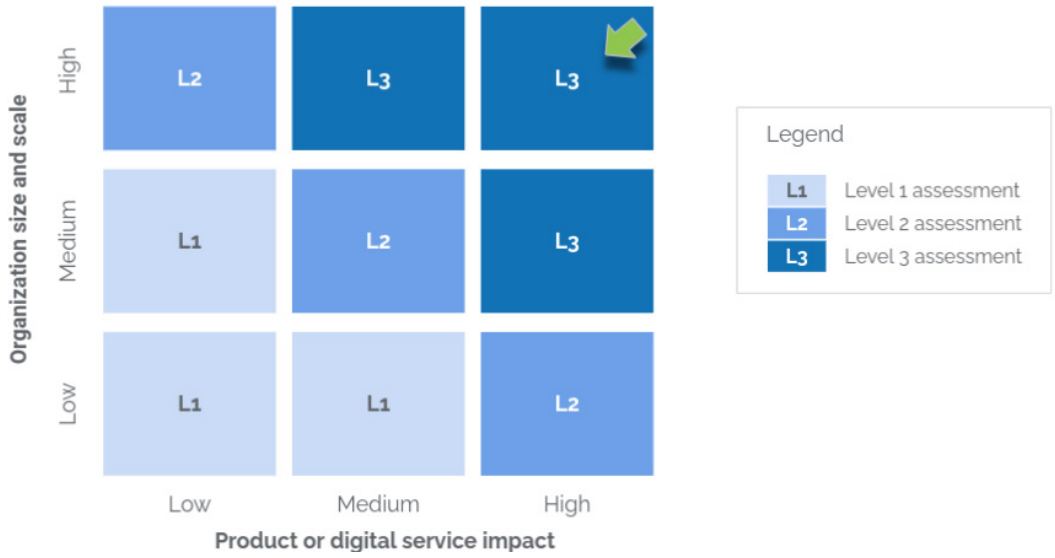
6.3.4 Determine the initial recommended assessment level

The evaluations from 6.3.2 (organizational size and scale) and 6.3.3 (product or service impact) are combined to determine the initial recommended level of assessment, using the matrix depicted in Table 4 — Assessment levelling matrix)

below to determine which of the three levels of assessments is appropriate for the relevant organization and product. Both the organizational size/scale and product impact should be factored in when contemplating a proportionate level of assessment.

For example, if an organization is determined under organizational size and scale to be 'high' and its product as "high impact", it is placed in the top-right box of the matrix, which implies that a Level 3 assessment is recommended, as shown in the following table.

Figure 4 — Assessment leveling matrix



6.3.5 Factor in additional business landscape considerations

The final step in applying the tailoring framework involves integrating considerations related to the business landscape in which an organization and product are operating. This is an optional internal measure, relating to factors that are internal/non-public or proprietary, where businesses may be aware of a factor that may justify a different level of assessment.

It is anticipated that these business landscape considerations would generally be used to increase the recommended level of assessment, rather than decrease it. The level of assessment should be informed by any unique circumstances or events that may impact the risks that a particular product or digital service must navigate.

For example, an organization may be aware of factors that may impact the likelihood of risk and therefore merit a more tailored assessment than would otherwise be indicated by the organizational size or product or service impact (e.g., if the product was due to expand into new markets). In addition, an organization may have information that does not become apparent in the initial determination of the assessment level, which may impact the appropriate assessment level.

There are several factors that may impact the level of assessment chosen for a product or service. These factors may include:

a. Service in new market (expansion): the product provides a new service or services a new geographic region for the organization;

b. Rapid product changes: significant changes to a product, or new features have been added to the product, in the past year;

c. New merger and acquisition activity, joint venture, or partnership: a merger or acquisition completed or joint venture/partnership entered into in the past year that impacts the product;

d. Prior assessments/audits: recently completed assessments/audits provide information on strength or weakness in controls similar to the practices in this document;

e. User growth trajectory: measured as the percentage in growth of registered users over the past twelve months. A higher growth rate may indicate increased exposure and a more rapidly evolving threat landscape as it relates to content;

f. Rapid social or political changes: can increase the likelihood, scope, or severity of content- and conduct-related risks, including increases risks to users, political violence, social unrest.

The specific impact and magnitude of these events can vary widely from organization to organization, and from product to product. If one or more of these circumstances or events apply, the organization makes a risk-based determination as to whether an adjustment in the level of assessment is warranted.

6.4 Assessment Execution

6.4.1 Overview

After applying the tailoring framework to determine the appropriate assessment approach (L1, L2, L3), the assessment itself is executed. The assessment that the

organization undertakes is according to the scoping in Clause 6.1 above, and is assessed for adherence to the overarching commitments with specific focus on the organization selected practices that underpin those commitment areas. The practices are then objectively evaluated across three key dimensions: people, process, and technology.

The assessment is designed to help the organization develop a deeper understanding of the implementation of selected practices to mitigate content- and conduct-related risks. The outcome of the assessment will help the organization better understand the current state of their capabilities and their dependencies with respect to people, processes, and technologies.

6.4.2 Assessment Methodology

This section details the methodology that the organization may follow to execute each assessment. There are five stages or steps that make up the assessment process, from initial information gathering or discovery, to reporting of findings and results, shown in the figure below.

Figure 5 — Assessment methodology

The corresponding activities or procedures performed within each step will differ based on the selected level of depth for the assessment (L1, L2, or L3). For example, a Level 3 assessment may include detailed testing of the effectiveness of specific process controls (e.g., are target turnaround times for user complaint reviews being met?), while a Level 1 assessment may involve a higher-level review and understanding of processes. Please reference Annex C for a summary of differentiating attributes for Level 1, Level 2, and Level 3 assessments.

TABLE 3 — ASSESSMENT STEPS

Step	Description	Objective
Discover relevant information	Engage key product stakeholders and perform initial information discovery on the organization's practices across the 5 commitments and identify the practices to be evaluated for their use in mitigating content and conduct risks.	Establish baseline understanding of the operational landscape and identify the specific practices used to mitigate content- and conduct-related risks.
Identify and prioritize relevant risk considerations	Using the artefacts and information collected during the "Discover" stage - identify, document and prioritize risks about the ways that content- and conduct-related risks are identified and mitigated.	Prioritize risks about the ways that content- and conduct-related risks are identified and mitigated to inform focus areas for the assessment.
Assess practices and risk mitigation	For the relevant risks about the ways that content- and conduct-related risks are identified and mitigated at the organization and focus areas identified in the previous step, analyse the practices employed to control for, or protect against, trust and safety risks.	Understand current-state processes, practices, and tools in relation to a common maturity scale.
Test control strength and effectiveness	Perform a control strength evaluation, including control design and effectiveness testing.	Understand, at a granular level, the operational effectiveness of risk mitigation processes, procedures, and tools.
Report results and findings	Compile all analysis results and report out on findings, observations, and future opportunities for improvement on the ways that content- and conduct-related risks are identified and mitigated at the organization moving forward.	Share key observations and findings with partners to facilitate collaborative development of industry standards and perspective.

6.4.3 Discover

The initial step in executing the assessment involves engaging the relevant stakeholders to gain an initial understanding of the operational and risk landscape in which the organization is operating. This initial information discovery provides the baseline understanding that drives downstream assessment activities.

Example activities in this stage of the assessment include:

a. Socialize assessment and objectives with stakeholders to drive general awareness of the commitments and how this assessment process fits into the broader context.

b. Hold initial workshops and collaborative discussions with cross-functional teams to gain a high-level understanding of trust and safety practices, potential challenges, and risk areas relevant to the areas of self-assessment.

c. Develop questionnaire, or information discovery form, to facilitate detailed information gathering (questions selected from the Question Bank based on the applicable practices identified by the organization). Please reference Annex D for an illustrative example of an information discovery form. Please see Annex E for the Question Bank.

d. Collect and review available documentation, such as organizational charts, process diagrams/workflows, to develop a baseline understanding of practices of the organization related to the commitments. [Level 2 and Level 3 only]

6.4.4 Identify

Grounded in the information and artifacts gathered during the "Discover" stage, the organization can identify and prioritize specific practices and focus areas to hone in on during the assessment. This prioritization may take the form of a traditional high/medium/low risk stratification, or "heatmap",

Example activities in this stage of the assessment include:

a. Analyse discovery results (e.g., information discovery form responses) to identify potential focus areas for downstream self-assessment activities.

b. Review any available documentation (e.g., previously completed product assessments or audits) to understand applicable risks related to the processes by which product content and conduct-related risks are identified and mitigated.

c. Segment or stratify identified risks related to the processes by which product content- and conduct-related risks are identified and mitigated into high,

medium, and low (or similar tranches) to facilitate prioritization [Note: Level 2 assessments against the commitments will focus on the highest risk areas, while Level 3 assessments will consider high, medium, and low risk areas in relation to the commitments].

6.4.5 Assess

Once the initial information discovery and risk identification is completed, companies can begin to evaluate the practices, processes, and tools in place to control for, or mitigate, content- or conduct-related risks. A maturity scale, grounded in the five commitments, will serve as the basis for the evaluations:

TABLE 4 — MATURITY RATING SCALE

1. Ad Hoc	2. Repeatable	3. Defined	4. Managed	5. Optimized
A rating of Ad Hoc is assigned when execution of best practices is incomplete, informal, or inconsistent.	A rating of Repeatable is assigned when execution of best practices occurs without standardized processes. Organizations aim to document more formalized practices.	A rating of Defined is assigned when execution of best practices occurs with defined and documented processes. Processes are more proactive than reactive, and implemented across the organization.	A rating of Managed is assigned when execution of best practices is defined, documented, and managed through regular reviews. Organizations use feedback to continuously mitigate process deficiencies.	A rating of Optimized is assigned when execution of best practices promotes trust and safety in every aspect. Processes are continuously improved with innovative ideas and technologies.

Depending on the level of assessment depth, there are several activities in this stage that will facilitate this evaluation.

Example activities in this stage of the assessment include:

a. Hold additional workshops with relevant stakeholders for deeper discussions regarding practices, processes, and tools, and how they are designed to control for content- or conduct-related risks.

b. Compare the practices reviewed against the common maturity scale to develop high-level observations with respect to the commitments and the exemplary best practices.

c. Perform review of any system, workflow, or procedural documentation that may be available to develop an in-depth understanding of processes, and where controls may be implemented within those processes [Level 2 and Level 3 assessments only].

d. Document risk mitigation controls in place (e.g., develop a control "inventory' for the product). [Level 2 assessments focus on higher risk areas, while Level 3 assessments consider all risk areas]

6.4.6 Test

Testing enables organizations to ensure that their content and conduct risk mitigation practices including people, process and technologies are working. The practices use a combination of people, processes and tools to prevent, detect or correct issues caused by unwanted events. Testing controls in a L2 or L3 assessment provides companies an opportunity to more thoroughly evaluate the design and the operational effectiveness of the controls. Testing for the assessment framework typically involves reviewing a data sample to help understand how the practices and the supporting controls are working. For example, selecting and analysing a random data sample enables an organization to "test" how well an organization can track policy violation rates and improvement over time.

Example activities in this stage of the assessment include:

a. Develop a testing plan for selected control including the procedures, documentation, and data/evidence required to perform the testing.

b. Review procedural, process, and technical documentation to evaluate the design of the identified control.

c. Analyse empirical evidence or representative sample data to determine the current efficacy of the control from an operational standpoint.

d. Review testing results with product stakeholders (either business or technical, or both) to validate observations and supporting evidence leveraged.

There are no standardized testing methodologies for trust and safety related practices and organizations will develop their own approaches.

6.4.7 Report

Compile all analysis results and report out on findings, observations, and future opportunities for process improvement moving forward. Create a solution roadmap for identified future opportunities, and key performance indicators (KPIs) to monitor ongoing progress. Please reference Annex 4 for an illustrative example of a potential report layout.

Sample activities include:

a. Facilitate workshops to discuss findings and observations with stakeholders

b. Issue report on summary-level observations and go-forward opportunities for better processes to identify and mitigate content- and conduct-related risks

c. Report on detailed process, risks related to existing processes, and control-level results

d. Develop solution roadmap for key improvement opportunities identified for processes

e. Align on KPIs to monitor progress against go-forward plan or solution roadmap for improved processes to identify and mitigate content- and conduct-related risks (e.g., reducing turnaround times for policy enforcement)

Annex A (informative)

Illustrative examples of the tailoring framework

A.1 Illustrative examples of size and scale classification

Input	Low	Medium	High	Resulting Categorization
Example Organization A: *Revenue of €15B and approx. 8K employees*				
Previous year's revenue	**€15B**	€25B - €100B	> €100B	**Low**
Total number of employees	**~8,000**	10,000 - 100,000	> 100,000	
Example Organization B: *Revenue of €6B and approx. 15K employees*				
Previous year's total revenue	**€6B**	€25B - €100B	> €100B	**Medium**
Total number of employees	< 10,000	**~15,000**	> 100,000	
Example Organization C: *Revenue of €150B and approx. 40K employees*				
Previous year's total revenue	< €25B	€25B - €100B	**€150B**	**High**
Total number of employees	< 10,000	**~40,000**	> 100,000	

Illustrative examples of product impact classification

Example Product A:
User volume of 50 million and 1 product feature with content/conduct risk

Input	Low	Medium	High	Resulting Categorization
User Volume	**50 mil**	100-500 mil	> 500 mil	**Low**
Product Feature Risk	**1**	5-15	>15	

Example Product B:
User volume of 400 million and 2 product features with content/conduct risk

Input	Low	Medium	High	Resulting Categorization
User Volume	< 100 mil	**400 mil**	> 500 mil	**Medium**
Product Feature Risk	**2**	5-15	>15	

Example Product C:
User volume of 70 million and 16 product features with content/conduct risk

Input	Low	Medium	High	Resulting Categorization
User Volume	**70 mil**	100-500 mil	> 500 mil	**High**
Risk Profile	< 5	5-15	**>15**	

Annex B (informative)

Risk Profile Questionnaire

B.1 Product purpose

Product purpose profiling involves asking:

a. Is the primary purpose of the product for users to interact with each other online?

b. Is the primary purpose of the product designed for users to either make or consume content, or both?

c. Is the primary purpose of the product primarily designed to facilitate offline interactions (e.g., dating apps, sharing economy)?

d. Is the primary purpose of the product is to facilitate economic transactions?

Organizations should select at least one primary purpose for the assessed product. Although many digital products or services may have more than one purpose, organizations should rely upon documentation, including public reporting and marketing materials, to determine their responses above.

B.2 Intended audience age ranges

Intended audience age ranges profiling involves asking:

a. Is the product marketed to children (under 13)?

b. Is the product marketed to minors (13-18)?

c. Is the product marketed to adults only?

B.3 Intended audience user type

Intended audience user type profiling involves asking:

a. Is the product primarily offered to consumers?

b. Is the product primarily offered to enterprises?

Organizations should use objective criteria, such as publicly available documentation, marketing materials, and recommended ages from app stores, in order to answer these questions.

B.4 Feature set

Feature set profiling involves asking:

a. Does the product host or store user-generated content?

b. Does the product enable monetization of user-generated content?

c. Does the product enable non-public-facing user-generated content?

d. Does the product enable live video or audio streaming?

e. Does the product provide an online marketplace for commercial transactions?

f. Does the product use either curation or algorithmic promotion of content, or both, to promote virality?

g. Does the product have other features that implicate content- or conduct-related risks as specified by the organization (e.g., generative AI features)?

h. Was the product released in the past year?

Organizations should consider whether the product has features that recommend content to large audiences of users who would not otherwise view the content, or has features that facilitate the sharing of content via other products or services.

B.5 Target market

Target market profiling involves asking:

a. Is the product offered in more than 20 languages?

b. Does the product specifically target more than 30 countries?

c. Does the product provide a new service in a given geographic region or service a new geographic region for the organization in the past year?

Specifically targeting a country means that the organization offering the product has taken proactive measures to market the product in the country (e.g. sales teams focused on a country).

Annex C (informative)

Summary of differences between L1, L2, and L3 Assessments

TABLE 1 — DIFFERENCES BETWEEN ASSESSMENT LEVELS

Assessment Step	Assessment Activities	Level 1	Level 2	Level 3
Discover	Hold initial workshops	Full	Full	Full
	Draft questionnaire	Limited	Targeted	Full
	Collect and review documentation	N/A	Targeted	Full
Identify	Analyze results from discovery to understand risks to practices used to implement commitments	Full	Full	Full
	Identify high risk areas for deep dive	N/A	Targeted	Full
	Segment risks as high, medium, and low for prioritized review	N/A	N/A	Full
Assess	Hold additional workshops to understand processes, tools, and operational practices	Limited	Targeted	Full
	Review detailed systems, workflows, and procedural documentation, if available	N/A	Targeted	Full
	Identify and document controls for all applicable risks (low, medium and high) for the processes identified in the previous step	N/A	N/A	Full
Test	Identify testing areas based on prioritization and evaluate design and operating effectiveness	N/A	Targeted	Full
Report	Facilitate workshops to discuss findings and observations with stakeholders and report on summary level observations	Limited	Targeted	Full
	Develop detailed process, risk, and control results; solutions roadmap; and align on key risk indicators	N/A	Targeted	Full

Annex D (informative)

Sample information discovery form

DTSP ID	Commitment	ID	Question	Answer	POC/ Responder
Roles & Responsibilities					
PD.2 PD.7 PD.8	**PD.2:** Include Trust and Safety team or equivalent stakeholder in the product development process at an early stage, including through communication and meetings, soliciting and incorporating feedback as appropriate **PD.7:** Provide for ongoing pre-launch feedback related to Trust and Safety considerations **PD.8:** Provide for post-launch evaluation by the team accountable for managing risks and those responsible for managing the product or in response to specific incidents	**A1.1**	Do you have a trust and safety team or individual involved in the product development process? How are trust and safety risks evaluated pre- and post-product launch? Is there a team accountable for this?		
Policies					
PG.1 PG.6 PE.1a PI.2	**PG.1:** Establish a team or function that develops, maintains, and updates the company's corpus of content, conduct, and/or acceptable use policies **PG.6:** Document for internal use the interpretation of policy rules and their application based on precedent or other forms of investigation, research, and analysis **PE.1a:** Constitute roles and/ or teams within the company accountable for policy creation, evaluation, implementation, and operations **PI.4:** Establish processes to ensure policies and operations align with these Commitments	**A2.1**	Which team(s) is/are involved in updating or writing the product's content, conduct, or acceptable use policies? How do you document the interpretation and practical application of policy rules based on precedent, or other forms of research and analysis? Please describe current assessment methods for evaluating accuracy and effectiveness of content-related policies and operations.		

DTSP ID	Commitment	ID	Question	Answer	POC/ Responder
PG.3	**PG.3:** Develop user-facing policy descriptions and explanations in easy-to-understand language	**A2.2**	Are terms of service, policies, or applicable community guidelines made easily accessible to users? What is the frequency at which terms of service, policy updates, and community guidelines are communicated to users? By what means are these communicated to users?		
Training, Wellness and Awareness					
PE.2 PE.3	**PE.2:** Formalize training and awareness programs to keep pace with dynamic online content and related issues, to inform the design of associated solutions **PE.3:** Invest in wellness and resilience of teams dealing with sensitive materials, such as tools and processes to reduce exposure, employee training, rotations on/off content review, and benefits like counselling	**A3.1**	How do you invest in reviewer wellness and awareness? What types of training programs and benefits are available to team members?		

Annex E (informative)

Question Bank

	People, Process, Technology	Topic Area	Question
Commitment 1: Product Development	Process	Risk Identification and Assessment	How does your team evaluate and consider trust and safety risks during the product development lifecycle?
	Process	User Experience	How do you balance product useability with security when considering the design of product features?
	People	Roles and Responsibilities	Do you have a trust and safety team or individual involved in the product development process?
	People, Process	Risk Identification and Assessment	How are trust and safety risks evaluated pre- and post-product launch? Is there a team accountable for this?
	Technology	User Experience	How does your product allow users to control their own product experience as it relates to content? What sorts of technical measures (e.g., blocking or muting) are in place?
	Process	Risk Identification and Assessment	How does your team perform or participate in risk assessments to better understand potential risks?
	Process, Technology	Risk Identification and Assessment	What capabilities do you leverage to understand patterns of abuse prevalent on the product, or service?
	Process	User Feedback	How do you seek and incorporate user feedback related to trust and safety in the product design process?

	People, Process, Technology	Topic Area	Question
Commitment 2: Product Governance	Process	Policies and Terms of Service and Guidelines	Are terms of service, policies, or applicable community guidelines made easily accessible to users?
	Process	Policies and Terms of Service and Guidelines	What is the frequency at which terms of service, policy updates, and community guidelines are communicated to users? By what means are these communicated to users?
	Process	User Feedback	Do you have processes for taking user considerations into account when drafting and updating relevant Product Governance, such as policies, terms of service, or community guidelines?
	Process	Policies and Terms of Service and Guidelines	How do you document the interpretation and practical application of policy rules based on precedent, or other forms of research and analysis?
	People, Process	User Feedback	Do you have any forms of community-led self-regulation (e.g. forums for governance or tools for community moderation)?
	People	Policies and Terms of Service and Guidelines	Which team(s) is/are involved in updating or writing the product's content, conduct, or acceptable use policies?
	Process	Feedback and External Collaboration	Do you work with industry groups, third-party civil society groups, or external experts to solicit input on product policies?

	People, Process, Technology	Topic Area	Question
Commitment 3: Product Enforcement	Process, Technology	Detection by Users	Are users able to report/flag content, conduct, or a user account as potentially violating policy? If so, please describe the process.
	Process	Review Processes and Procedures	What is the process for reviewing content that has been identified or flagged as potentially violating policy?
	Process	Review Processes and Procedures	How are content reviews prioritized, and what factors are taken into consideration?
	Technology	Review Processes and Procedures	What types of tools/systems are used to review content or manage the review process?
	Process	Enforcement Actions	What types of actions may be taken against a piece of content or user in relation to policy violations?
	Technology	Enforcement Actions	What tools/systems are used to enforce content policies or manage the enforcement process?
	Process, Technology	User Notifications	How are users notified of enforcement actions taken relating to their content or activity on the product/service (e.g., broad public notices, icons)?
	Process, Technology	Detection Mechanisms and Processes	What types of processes or mechanisms are in place to proactively detect potentially violating content or conduct (automated or manual)?
	Technology	Detection Mechanisms and Processes	How do you make decisions about provisioning technology to conduct enforcement operations? How do you determine whether to build, buy, adapt, or collaborate when assessing available tools or technologies?
	Process	User Recourse	Is there a mechanism available for users to appeal decisions or actions taken on the product or service? If so, please describe the process.
	Process, Technology	Data Management and Retention	How are data related to enforcement actions (such as data relevant for investigations or key contextual data) retained and managed?

	People, Process, Technology	Topic Area	Question
Commitment 3: Product Enforcement	People	Training and Awareness	How do you invest in reviewer wellness and awareness? What types of training programs and benefits are available to team members?
	Process	Detection by Third Party Partners	If applicable, how do you collaborate or partner with third parties to identify and flag potentially violating content or conduct?
	Process, Technology	Detection Mechanisms and Processes	How does your team protect against coordinated dissemination of illegal or violating content (e.g. public health misinformation, content harmful to minors, electoral processes) through automated or manual means?
	Process, Technology	Detection Mechanisms and Processes	How does your team protect against the amplification of harmful content or conduct? What processes and systems are in place to deter bad actors and behaviours that violate product policies?
	Process	Feedback and External Collaboration	How and when are notifications or appropriate reporting sent outside the organization, such as to law enforcement, in cases of credible and imminent threat to life?

	People, Process, Technology	Topic Area	Question
Commitment 4: Product Improvement	Process, Technology	Process Quality and Continuous Improvement	Please describe current assessment methods for evaluating accuracy and effectiveness of content-related policies and operations.
	Process	Risk Identification and Assessment	How often do you conduct risk assessments and how are emerging threats or risks taken into account?
	Process	Risk Identification and Assessment	What are some of the key risk areas or focus areas that are top-of-mind as it relates to user trust and safety?
	Process, People	Risk Identification and Assessment	Please describe if and how you use risk assessments to determine allocation of resources for emerging content- and conduct-related risks.
	Process, People	Process Quality and Continuous Improvement	Please describe any existing methods for internal product feedback and evaluation, as it relates to mitigating content- and conduct-related risks.
	Process	User Feedback	How do you seek and incorporate user feedback in the organization's approach and processes to protect users?
	Process	Feedback and External Collaboration	Please describe how you work with recognized third party civil society groups and experts (e.g. qualified fact checkers or human rights groups) to help evolve efforts to mitigate content- and conduct-related risks.
	Process	User Recourse	Please describe any remedy mechanisms in place for users that have been directly affected by moderation decisions. (i.e. content removal, account suspension or termination).

	People, Process, Technology	Topic Area	Question
Commitment 5: Product Transparency	People	Transparency Reporting	If applicable, how is your team involved in developing or providing input into organization transparency reporting or content risk reporting?
	Technology	Transparency Reporting	How frequently, and via what means (e.g., publicly available website), are transparency reports made available to the public and other external stakeholders?
	Process	Transparency Reporting	Please describe at a high level metrics or data retained for the purposes of regular transparency reporting (e.g. abuses reported, processed, data requests processed and fulfilled).
	Process, Technology	Data Management and Retention	Do you have a process in place to log user complaints, decisions, and enforcement actions according to relevant data policies?
	Process, Technology	User Notifications	How and when are notices provided to users whose content or conduct is at issue in an enforcement action (with relevant exceptions, such as legal prohibition or prevention of further harm)?
	Process	Feedback and External Collaboration	How do you collaborate with academic and other researchers working on relevant trust and safety subject matter (to the extent permitted by law, security and privacy standards, and other business considerations)? Do you share data or insights on a regular basis?

Annex F (informative)

Illustrative example: product area report template

F.1 Overview

[brief introduction of the purpose of this report and methods to collect results (e.g. tailoring framework, assessment steps, etc.]

F.1.1 Introduction

[text]

F.1.2 Background and purpose

[text]

F.1.3 Approach

[text]

F.2 Scope of Review

[describe the level of the assessment (i.e. L1, L2, L3)]

F.3 Assessment Results

F.3.1 Commitment 1: Identify, evaluate, and adjust for content- and conduct-related risks in product development.

F.3.1.1 Findings

- [finding 1]
- [finding 2]
- [finding 3]

F.3.1.2 Opportunities for Improvement

- [improvement 1]
- [improvement 2]
- [improvement 3]

F.3.1.3 Conclusion

- Maturity: [1 ad hoc, 2 repeatable, 3, defined, 4 managed, 5 optimized]

Bibliography

[1] ISO 31071:2022, *Risk Management — Vocabulary*

[2] ISO/IEC TS 5928:2023, *Information Technology — Cloud Computing and distributed platforms —Taxonomy for digital platforms*

[3] ISO/IEC 27000:2108, *Information Technology — Security Techniques — Information Security Management Systems —Overview and vocabulary*

[4] ISO/IEC Guide 51:2014, *Safety Aspects — Guidelines for their inclusion in standards*

[5] ISO 32110:2023, *Transaction assurance in E-commerce — Vocabulary*

[6] ISO/IEC 29100:2024, *Information Technology — Security Techniques — Privacy framework*

[7] ISO/IEC 22989:2022, *Information Technology — Artificial intelligence — Artificial intelligence concepts and terminology*

[8] ISO/IEC TS 5723:2022, *Trustworthiness — Vocabulary*