STANDARD-SETTING PROJECT

Content Governance in the Internet Media & Services Industry

RECOMMENDED CHANGES TO THE SASB INTERNET MEDIA & SERVICES STANDARD

Publication Date: June 2, 2022

Prepared by the Sustainability Accounting Standards Board®
# Table of Contents

**Overview**............................................................................................................................................................................. 3

**Basis for Conclusions on Recommended Changes to the Internet Media & Services Standard**............................................................................................................................................................................. 4
  - Introduction.................................................................................................................................................................................. 5
  - Summary of recommended changes to the Standard.................................................................................................................. 5
  - How were the recommendations developed? ............................................................................................................................ 7
  - Why was the project added to the standard-setting agenda? ....................................................................................................... 8
  - What is the basis for the Board’s recommended changes to the Standard? ............................................................................. 9

**Recommended Changes to the Internet Media & Services Standard**............................................................................................................. 23
  - Summary of recommended changes to the Internet Media & Services Standard........................................................................ 24
  - Content Governance & Freedom of Expression topic................................................................................................................. 26
Overview

Following the planned consolidation of the Value Reporting Foundation into the International Financial Reporting Standards (IFRS) Foundation on June 30, 2022, stewardship of the SASB Standards will pass from the SASB Standards Board to the International Sustainability Standards Board (ISSB). These recommended changes represent changes to the Standard that the SASB Standards Board would have pursued through the publication of an exposure draft, including an invitation to comment, if there were sufficient time to advance the project through the next phases of its due process as outlined in the SASB Rules of Procedure.

The SASB Standards Board has published these recommended changes for consideration by the ISSB, and for market participants interested in the continued improvements to the SASB Standards. Any future updates to the SASB Standards will subjected to ISSB due process prior to being finalized.

Key changes recommended by the Board include:

- The addition of a new disclosure topic, Content Governance & Freedom of Expression, and five new corresponding metrics intended to capture risks and opportunities associated with companies’ management of user-generated content; and
- The relocation of two existing metrics, TC-IM-220a.5 and TC-IM-220a.6, from the current Data Privacy, Advertising Standards & Freedom of Expression disclosure topic into the new Content Governance & Freedom of Expression disclosure topic.
Basis for Conclusions on Recommended Changes to the Internet Media & Services Standard
Introduction

1 The basis for conclusions accompanies, but is not part of, the “Recommended Changes to the Internet Media & Services Standard” (recommended changes) as part of the Content Governance in the Internet Media & Services Industry standard-setting project. The basis for conclusions summarizes the considerations and rationale of the SASB Standards Board (the Board) in developing the recommended changes. Individual Board members gave greater weight to some factors than to others.

2 The basis for conclusions is organized as follows:
   a) Summary of recommended changes to the Standard
   b) How were the recommendations developed?
   c) Why was the project added to the standard-setting agenda?
   d) What is the basis for the Board’s recommended changes to the Standard?

Summary of recommended changes to the Standard

3 The Board recommends that a current disclosure topic in the Internet Media & Services (IM) Standard—Data Privacy, Advertising Standards & Freedom of Expression—be split into two separate topics, with one of them being significantly expanded in scope to capture risks and opportunities associated with content governance themes. As proposed, the existing metrics regarding user privacy and advertising practices would not change but would continue to exist in the Standard in a stand-alone topic with a revised scope. Elements of the current topic addressing how companies operate and comply with governments to control user-facing content, products, or features across jurisdictions are recommended to be combined with new metrics in a new disclosure topic: Content Governance & Freedom of Expression.

4 The primary goal of these recommended changes is to increase the completeness of the Standard, which, according to market feedback, doesn’t fully capture risks and opportunities related to content governance on internet platforms.
Table 1. Recommended changes to disclosure topics in the IM Standard

<table>
<thead>
<tr>
<th>Sustainability angle</th>
<th>Current Standard</th>
<th>Proposed changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy and handling of user data</td>
<td>Data Privacy, Advertising Standards &amp; Freedom of Expression</td>
<td>ADD content governance topic: REVISE scope of privacy topic</td>
</tr>
<tr>
<td>Law enforcement requests for user data</td>
<td></td>
<td>Revised topic: Data Privacy &amp; Advertising Standards</td>
</tr>
<tr>
<td>Government requests to remove content or services</td>
<td></td>
<td>New topic: Content Governance &amp; Freedom of Expression</td>
</tr>
<tr>
<td>Freedom of expression – platform content removal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harmful online content</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Recommended new disclosure topic: Content Governance & Freedom of Expression

5 The proposed new Content Governance & Freedom of Expression topic captures risks associated with user-generated content in the Internet & Media Services industry. Users of these platforms create or upload a variety of content that can carry negative societal impacts, including but not limited to inciting violence, glorifying terrorist acts, and depicting child sexual abuse.

6 The proposed establishment of the Content Governance & Freedom of Expression disclosure topic would include the addition of five new corresponding metrics:

   a) **Content moderation table:** (1) number of content items removed, (2) percentage of removed content discovered proactively, (3) percentage of removed content appealed, (4) percentage of appealed content restored, (5) average user impressions of removed content

   b) **Total amount spent on content governance**

   c) **Description of approach to identification and management of significant content- and conduct-related risks**

   d) **Description of approach to content display and recommendations, and how these systems account for content- and conduct-related risks**

   e) **Description of approach to content moderation**
Additionally, two metrics from the current disclosure topic Data Privacy, Advertising Standards & Freedom of Expression, TC-IM-220a.5 and TC-IM-220a.6, are recommended to be relocated to the proposed new disclosure topic.

Revised disclosure topic: Data Privacy & Advertising Standards

As proposed, the revised Data Privacy & Advertising Standards disclosure topic will continue to focus on the risks related to the use and storage of sensitive customer data, including law enforcement requests to access user information.

The recommended changes include moving the following metrics to the newly proposed disclosure topic, Content Governance & Freedom of Expression. There are no recommended changes to the text of these metrics or their technical protocols at this time:

a) TC-IM-220a.5. List of countries where core products or services are subject to government-required monitoring, blocking, content filtering, or censoring.

b) TC-IM-220a.6. Number of government requests to remove content, percentage compliance with requests.

How were the recommendations developed?

As a general practice, the Board and technical staff actively monitor topics identified in each industry standard by assessing corporate disclosures and the effectiveness of the relevant standards at capturing performance on the issue in a decision-useful manner. Technical staff also monitor industry developments and solicit input from market participants and subject matter experts on the need for standard setting.

Since this project was added to the agenda, the technical staff has conducted research and engaged in consultations with companies, investors, and subject matter experts, including the SASB Standards Advisory Group, to support development of the recommended changes. Throughout the course of this project, the technical staff contacted more than 60 market participants and held consultations with 41 of them: 13 investors, 9 companies, 6 industry bodies with corporate membership, and 13 subject matter experts. The companies included mega-cap companies, large-cap companies, and small-cap start-ups. Staff consulted with a variety of investors, including fundamental equity analysts, portfolio managers, and environmental, social, and governance (ESG) and stewardship team members.

The research and consultation conducted by the technical staff led to a series of deliberations (discussed below) by the Board on key issues, considerations, and challenges related to the development of the recommended changes. Additional information related to the standard-setting process that the Board follows to maintain and update the SASB Standards can be found on the SASB Standards website, in the Rules of Procedure, and/or in the Conceptual Framework. Additional project-specific information can be found on the Content Governance in the Internet Media & Services Industry project webpage on the SASB Standards website.
Why was the project added to the standard-setting agenda?

The Board and technical staff actively monitored the evolving issue of freedom of expression on internet platforms after the previous Standards were codified in October 2018. This included assessing corporate disclosures and the effectiveness of the relevant Standards at capturing performance on the issue, monitoring developments in the industry, and soliciting input from market participants.

The same technologies that have enabled huge growth in productivity and connected people across the world also enable bad actors and facilitate the creation, hosting, and sharing of harmful content. In recent years, there has been a steadily increasing focus from investors, regulators, the media, and society at large on the role that internet platforms play in disseminating harmful online content and/or facilitating harmful online behavior. This content ranges from the illegal and harmful (e.g., content depicting the sexual exploitation of children) to the legal but potentially harmful (e.g., certain types of misinformation).

Businesses provide services that can facilitate the dissemination of harmful content or harmful behavior in a variety of ways. The existence of harmful content and behavior on the internet leads to platforms deciding what types of content they are willing to host or amplify. The nature of the responsibility that internet platforms have over user-generated content, and the ways in which these platforms both craft and enforce their content policies, is highly contested; the evidence gathered throughout this project also indicates that they pose financially material risks and opportunities for many companies in the technology sector.

The current IM Standard addresses risks and opportunities associated with how companies interact with government requests for user information and government requests to remove, block, or censor content. However, it does not include disclosure requirements on the voluntary actions that platforms take to set and enforce content policies.

In December 2019, the chair of the Board announced a research project, titled Content Moderation on Internet Platforms, with the objective of evaluating whether standard setting should be pursued in any of the industries in the Technology & Communications sector to address this topic. During that research project, the technical staff compiled the Content Moderation Taxonomy, which outlines many of the social externalities related to harmful online content and the business activities across the Technology & Communications sector to which those externalities might apply. The taxonomy was released to the public in November 2020.

In September 2020, the Board voted in favor initiating a standard-setting project with the objective of identifying topics, metrics, and disclosure guidance to address themes associated with harmful online content and user freedom of expression to enhance the completeness of the Internet Media & Services Standard. The Board’s decision was based on research and consultations for the content moderation research project, which consisted of evidence of both financial impact for companies in the IM industry and investor interest.
What is the basis for the Board’s recommended changes to the Standard?

In developing these recommended changes to the Standard, the Board considered two key types of evidence. As described in greater detail below, these inputs are (1) evidence that content governance is a financially impactful issue in the IM industry, and (2) evidence of investor interest in content governance.

Evidence of financial impact

The management of content governance risks and opportunities has significant financial implications for companies in the IM industry. These include the expense of content moderation, revenue risks and opportunities related to the ability to attract and retain advertisers, and regulatory risk, such as the potential for new regulations that could make platforms liable for the user-generated content they host in many jurisdictions.

Cost of content governance

Companies in the IM industry of all sizes dedicate a significant amount of resources to crafting and enforcing their content policies. This includes personnel costs for content moderators and investments in automated systems. While most companies do not disclose these costs, in 2020, Meta (then Facebook) CEO Mark Zuckerberg claimed in testimony before the U.S. Congress that the company was spending upwards of $3 billion on its force of 35,000 content moderation employees, many of whom work as contractors. In the same hearing, Alphabet CEO Sundar Pichai testified that his company had spent more than $4 billion on a force of over 10,000 content moderators. In 2021, Meta revealed that its spending on “safety and security” amounted to $13 billion over the course of the previous five years, or an average of $2.6 billion a year over that time frame.

While evidence and disclosures for smaller companies is more limited, research and consultations indicate that content governance challenges are not confined to large, U.S.-headquartered multinationals. Recently, the death of a content moderator at Chinese video-sharing site Bilibili led the company to announce it would hire an additional 1,000 moderators. Notably, the 2,400 moderators employed by Bilibili prior to this incident already constituted about 30 percent of its entire workforce. The gaming-focused social network Discord in 2020

stated that it had recently doubled the size of its trust and safety team, which continued to be “one of the biggest” at its organization.4

Revenue impacts tied to brand-safety concerns of advertisers

Many IM companies generate revenue through the display of third-party advertisements next to user-generated content. These companies face brand-safety challenges, as advertisers are wary of ads for their products and services appearing next to harmful (or potentially objectionable) content. In 2017, a newspapers’ analysis showed that advertisements from many of the world’s largest brands were appearing next to YouTube videos with potentially problematic depictions of children—and that platform algorithms were recommending similar content to users viewing these videos.5 Following further investigations, some of the world’s largest advertisers, including Coca-Cola, PepsiCo, Walmart, Starbucks, and General Motors, announced that they were suspending all advertising on YouTube. Researchers determined that advertisements from these companies were regularly appearing alongside content that could be considered hateful, racist, and even violent.6 The stock of YouTube’s parent company, Google, dropped by 4 percent in the immediate aftermath of these revelations, erasing billions of dollars of market capitalization in a few days.7 YouTube and many peer companies subsequently made significant investments in retooling their policies and procedures in order to regain the trust and business of many advertisers.

In 2020, a movement known as Stop Hate for Profit urged major advertisers to boycott Meta because of its perceived failure to remove hate speech and other harmful content. More than 400 companies, including Cola-Cola, Adidas, and Lego, agreed to temporarily pull their advertising off the platform.8 While there was not significant evidence that Meta’s revenues were materially impacted, the boycott received widespread media attention; competitors such as Snap and Pinterest posted strong results driven by increased advertising revenue, which many attributed directly to the Meta boycott having created space for these competitors to build better relationships with advertisers.9

Regulatory risk

---

Companies in the IM industry face a significant and growing number of regulations regarding user-generated content across the globe. In Europe, the proposed Digital Services Act would introduce sweeping obligations on platforms to detect, identify, and remove illegal content. Most recently, Ofcom, the U.K.’s communications regulator, introduced an even more far-reaching proposal that would impose large fines and give regulators the ability to prosecute executives personally for failures to comply with laws around online harms.

Germany already has in place a law, the Network Enforcement Act (NetzDG), that requires social networks with two million users or more to set up user-friendly complaint mechanisms and remove “manifestly unlawful content” within 24 hours. In the United States, the proposed Eliminating Abusive and Rampant Neglect of Interactive Technologies (EARN IT) Act would make platforms liable for failing to remove content that is exploitative of children. And in China, regulators have broadened their focus in recent years, drafting new rules that attempt to broadly control what platforms show their users, forcing them to remove certain kinds of “socially unacceptable” content and to promote certain social values.

Evidence of investor interest

In weighing whether the IM Standard should be updated, the Board noted significant evidence of investor interest regarding content governance practices of companies, as demonstrated through (1) the consistent presence of the topic on earnings calls in recent years and (2) investor feedback provided to the technical staff during consultations.

In its review of earnings call transcripts of companies that operate user-generated platforms, the Board noted that content governance is a consistent area of questions from investment analysts. Topics brought up by analysts in recent years include the impacts that information-quality efforts might have on user growth, the protections in place for children who use platforms, and the financial impacts of potential new regulations that would increase companies’ liability for user-generated content on their platforms.

---


14 During Twitter’s Q3 2020 earnings calls, Goldman Sachs Group’s Heath Patrick Terry asked: “Were there any impacts this quarter from information quality efforts or similar initiatives that you’ve taken in the past that have impacted user growth?”

15 During Alphabet’s Q2 2019 earnings call, Morgan Stanley’s Brian Nowak asked: “There has been a decent amount in the press about children’s safety on YouTube. And I’d be curious to hear for advertisers and parents on the call, what steps have you taken to really make sure that YouTube is safe for kids?”

16 During Meta’s Q3 2020 earnings call, MoffettNathanson’s Michael Nathanson asked: “Mark [Zuckerberg], I want to take you back to Section 230 and [your congressional testimony] for a second. It seems like you’re the most comfortable CEO about making changes to the law. I wonder
The Board also considered the feedback provided by investors during consultations held in 2020 for the content moderation research project, as well as during consultations held in 2021 for the content governance standard-setting project. Throughout dozens of consultations, investors consistently expressed their belief that content governance activities present financially material risks and opportunities for companies in the IM industry. Investors expressed different opinions regarding which aspects of content governance were the most financially impactful and relevant, with some investors placing more importance on potential regulatory risk, others focused on the overall costliness of content review, and others noting concerns regarding the ability of companies to attract and retain users and advertisers.

Other considerations

The Board also considered the prevalence of content governance themes across the IM industry. At its September 2020 meeting (see Summary of Meeting Outcomes), the Board noted that there are a variety of internet platforms that fall into the IM industry, not all of which focus on the dissemination of user-generated content. Ultimately, however, the Board decided to proceed with the project on the basis of (1) the importance of the topic to many of the companies in the IM industry and (2) the fact that the companies that do focus on user-generated content make up a significant majority of the market capitalization of publicly listed IM companies.

Finally, the Board considered the absence of standardized disclosures regarding the topic of content governance. While a number of companies now produce transparency reports regarding their content moderation outcomes and also produce blog posts and other information regarding their approaches to challenging and dynamic issues such as misinformation, different companies take significantly different approaches to which qualitative and quantitative information they share publicly. Therefore, the Board considered the potential value that SASB Standards could add to the space through executing this project.

What is the basis for the Board’s recommended changes to the disclosure topics?

The Board deliberated the appropriate disclosure topic scope and structure to reflect the different nature of social impacts and management approaches used for issues around privacy, freedom of expression, and harmful online content. While recognizing that there are areas of overlap between user privacy and content governance themes, including freedom of expression, the Board recommends that the IM Standard address these issues through two separate disclosure topics: (1) Data Privacy & Advertising Standards and (2) Content Governance & Freedom of Expression.

---

what are the changes that you had proposed. Does that pact to have a proposal in June resemble where we need to go? And what do you think the cost of these changes would be for Facebook?
In making this determination, the Board considered the evidence gathered by staff during its research and consultations regarding the different sustainability risks, opportunities, and management approaches associated with the two topics.

The current IM Standard contains the disclosure topic Data Privacy, Advertising Standards & Freedom of Expression, which captures the following concepts:

- a) Companies’ use and collection of sensitive user data, including for advertising.
- b) Companies’ compliance with law enforcement requests for user information.
- c) Companies’ compliance with government requests for removal of content and services.

The key concepts the Board recommends to add in an updated Standard are the following:

- d) Companies’ approach to managing risks around harmful user-generated online content and behavior.
- e) The impacts of companies’ content policies and enforcement actions on user freedom of expression.

The Board recommends that this set of concepts be captured in two disclosure topics. The first, Data Privacy & Advertising Standards, would be a revision of the scope of the current Data Privacy, Advertising Standards & Freedom of Expression disclosure topic and would cover concepts a) and b), above. A newly proposed disclosure topic, Content Governance & Freedom of Expression, would cover issues c), d), and e), above. This proposed disclosure topic would be mapped to the Customer Welfare General Issue Category of the SASB Standards.

Based on the evidence gathered during research and consultations throughout the course of this project, the Board believes that each disclosure topic individually meets the criteria for topic selection in the SASB Conceptual Framework, namely that each disclosure topic is financially impactful, of interest to users of corporate disclosures for financial decision-making purposes, prevalent across the IM industry, and actionable by companies.

**Rationale for recommendation to split the current disclosure topic**

At its public meeting in July 2021, the Board deliberated whether these new concepts should be addressed through the expansion of the existing disclosure topic or through multiple disclosure topics.

The Board recommends that privacy and advertising practices be separated from content governance concepts primarily on the basis that the management approaches and risks associated with the collection and usage of sensitive user data are substantially different from the risks related to policing harmful content and its effect on freedom of expression. For example, management approaches to content governance generally involve crafting policies

---

17 The Summary of Meeting Outcomes for the July 2021 Board meeting can be viewed [here](#).
regarding what types of content are permissible on a platform and then developing an enforcement model that must react to the content users post on a continuous basis. Management approaches to user privacy typically involve a separate process of data governance and compliance with a different set of regulations, and generally involve different parts of organizations than the management approaches regarding harmful online content.

Additionally, while the topic of user privacy is nearly ubiquitous in the IM industry, many internet platforms do not focus on user-generated content or otherwise may have minimal risk exposure to content governance themes. Given this, the Board believes that a secondary benefit of separating privacy disclosures from content governance disclosures would be that companies that are not exposed to content governance risks and opportunities could omit the content governance topic and therefore more easily prepare disclosures aligned with the IM Standard.

Rationale for recommendation to separate law enforcement requests for user information from government requests for content removal

During its deliberations, the Board considered that the above concepts b), regarding law enforcement requests for user information, and c), regarding government requests for content removal, could fall under either a disclosure topic focused on privacy or one whose focus includes freedom of expression. Ultimately, the Board determined that the concept of requests for user information is most relevant to the topic of data privacy and that the concept of content removal is most relevant to a topic that includes freedom of expression.

Law enforcement requests for user information generally relate to the investigation of criminal activity, and companies in the IM industry implement policies and procedures for evaluating the legality of such requests prior to determining whether to divulge user information. This information can be highly sensitive, including location data and the contents of private communications such as emails.

While companies' handling of law enforcement requests for user information can therefore generally be viewed under the umbrella of how companies approach user privacy, there are instances of governments using such laws to commit human rights violations, such as suppressing dissidents. Therefore, the Board recognizes that a component of law enforcement requests for user information is related to user freedom of expression.

Government requests for platforms to remove content or disable services in their jurisdictions directly affect or limit content available to the public, and could be considered an act of censorship. So how companies approach government requests for content removal are best understood as challenges to freedom of expression.

The Board also considered the fact that the management approach to law enforcement requests for user information and government requests for content removal may be similar in process. Nonetheless, given the differences in subject matter and the ways in which users are impacted by the ultimate decisions that arise from these management decisions, the Board determined that it would be preferable to separate the concepts.
Rationale for recommendation to combine harmful content and freedom of expression into one disclosure topic

At its July 2021 meeting, the Board deliberated on whether freedom of expression and concepts related to harmful content should be addressed in different disclosure topics. The social issues tied to harmful online content—including but not limited to that disseminating child sexual abuse material, inciting violence, or promoting terrorism and violent extremism—certainly cover a greater range of issues than concerns regarding company impacts on user freedom of expression. Nonetheless, the Board recommends that these concepts be combined into a single disclosure topic because the management strategies regarding these two concepts are inextricably linked: all enforcement actions aimed at reducing harmful online content (removing content, downranking it, banning or limiting the permission of users, etc.) directly impact the ability of users to express themselves freely on a public platform.

Further, as outlined below in the discussion of metrics, the Board believes that quantitative information regarding how companies manage risks and opportunities related to freedom of expression is best interpreted in the context of content removal.

What is the basis for the Board’s recommended changes to the metrics?

The Board considered a set of metrics for each disclosure topic to help users understand and interpret performance on the risk exposure and risk likelihood associated with each disclosure topic. To assess the magnitude of potential impact (i.e., risk exposure), the Board considered and proposed metrics that aim to measure total potential harm. To assess the probability or likelihood of the risk, the Board recommended quantitative metrics that provide a general sense of company investment in, and exposure to, issue-related risks, as well as qualitative metrics that focus on company controls and governance structures.

In executing this project, the Board and technical staff did not explicitly seek to reevaluate the existing metrics in the Data Privacy, Advertising Standards & Freedom of Expression disclosure topic. Rather, the Board and technical staff focused research and consultations on developing new metrics intended to capture content governance risks and opportunities.

Table 2. Key concepts to measure: Content Governance & Freedom of Expression disclosure topic

<table>
<thead>
<tr>
<th>Sustainability angle</th>
<th>Concept to be measured</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmful online content and conduct</td>
<td>How do companies identify and manage the most significant impacts that their services may have on users and society?</td>
<td>Description of approach to identification and management of significant content- and conduct-related risks</td>
</tr>
<tr>
<td></td>
<td>How do companies design their content display systems, and how do they alter them to reduce the risk of harms?</td>
<td>Description of approach to content display and recommendations, and how these systems account for content- and conduct-</td>
</tr>
<tr>
<td>related risks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>How do companies create and enforce policies determining what is allowed on their platforms?</td>
<td>Description of approach to content moderation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of content items removed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentage of removed content discovered proactively</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average user impressions of removed content</td>
<td></td>
</tr>
<tr>
<td>Freedom of expression</td>
<td>Does the company allow users to appeal content moderation decisions? If so, how does it approach the appeals process?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentage of removed content appealed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentage of appealed content restored</td>
<td></td>
</tr>
<tr>
<td>Freedom of expression</td>
<td>How do companies respond to government requests to remove content?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TC-IM-220a.5. List of countries where core products or services are subject to government-required monitoring, blocking, content filtering, or censoring</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TC-IM-220a.6. Number of government requests to remove content, percentage compliance with requests</td>
<td></td>
</tr>
</tbody>
</table>

**Recommended revision of disclosure scope in the Data Privacy & Advertising Standards disclosure topic**

The Board recommends that four metrics currently in the Data Privacy, Advertising Standards & Freedom of Expression remain in the IM Standard in a disclosure topic with a revised scope. The proposed topic, titled Data Privacy & Advertising Standards, would retain the following metrics:

a) **TC-IM-220a.1.** Description of policies and practices relating to behavioral advertising and user privacy.

b) **TC-IM-220a.2.** Number of users whose information is used for secondary purposes.

c) **TC-IM-220a.3.** Total amount of monetary losses as a result of legal proceedings associated with user privacy.

d) **TC-IM-220a.4.** (1) Number of law enforcement requests for user information, (2) number of users whose information was requested, (3) percentage resulting in disclosure.

The Board recommends that the following metrics be removed from the current Data Privacy, Advertising Standards & Freedom of Expression disclosure topic and added to a new disclosure topic, Content Governance & Freedom of Expression:
a) **TC-IM-220a.5.** List of countries where core products or services are subject to government-required monitoring, blocking, content filtering, or censoring.

b) **TC-IM-220a.6.** Number of government requests to remove content, percentage compliance with requests.

The Board’s rationale for moving TC-IM-220a.5 and TC-IM-220a.6 to the Content Governance & Freedom of Expression topic is discussed in the above section starting on paragraph 39.

**Recommended new metrics for the Content Governance & Freedom of Expression disclosure topic**

The Board recommends the addition of the following new metrics to the IM Standard in a new disclosure topic titled Content Governance & Freedom of Expression:

a) **Content moderation table:** (1) content items removed, (2) percentage of removed content discovered proactively, (3) percentage of removed content appealed, (4) percentage of appealed content restored, (5) average user impressions of removed content.

b) **Total amount spent on content governance.**

c) **Description of approach to identification and management of significant content- and conduct-related risks.**

d) **Description of approach to content display and recommendations, and how these systems account for content- and conduct-related risks.**

e) **Description of approach to content moderation.**

The Board intends for these new metrics to provide a complete and verifiable set of disclosures regarding how companies in the IM industry approach the risks and opportunities associated with content governance. Metrics a) and b) are quantitative metrics intended to provide information regarding the breadth and magnitude of companies’ content governance practices. Metrics c), d), and e) are qualitative metrics intended to provide users with information regarding the policies, practices, and procedures used by companies to identify and mitigate risks related to harmful content and conduct on their platforms.

While research and consultation indicate that content governance and the broader emerging field of “trust and safety” in the technology sector are still relatively nascent, the Board sought ways to align with existing initiatives in developing the recommended updates to the Standard. Therefore, much of the terminology used in the proposed metrics is aligned with the preliminary framework recently issued by a key industry body, the Digital Trust and Safety Partnership (DTSP).¹⁸

¹⁸Members of the DTSP include parent companies of many of the largest IM platforms, including Alphabet, Meta, Twitter, Discord, Reddit, Pinterest, and Vimeo.
The Board recognizes that the addition of five new metrics to the IM Standard significantly expands the scope of disclosure for preparers using SASB Standards but believes that the recommended changes are justified by the significant evidence of the topic’s financial relevance and the amount of interest expressed by investors.

Rationale for recommended metric: *Content moderation table*

As noted above, investors during consultations stated their interest in both qualitative and quantitative disclosures regarding content moderation. The intent of this metric is to provide (1) backward-looking quantitative data on the content moderation actions taken by companies during the reporting period and (2) quantitative data that supplements qualitative disclosures, such as the extent to which platforms use automated systems to remove unwanted content, and whether platforms provide users with an avenue to appeal content moderation decisions.

At its July 2021 meeting, the Board deliberated the best format for quantitative disclosures regarding content moderation and determined that an approach that accounts for the differing categories of harmful content would be desirable. As noted elsewhere in this document, a wide variety of harmful content and conduct can appear on platforms, including material with child sexual abuse, violent extremism, hate speech, and self-harm. The Board determined that the most appropriate terminology for what has frequently been referred to as “harmful content” in this project is better defined as “content- and conduct-related” risks (CCR risks), as harmful online content generally includes both the harmful content itself and harmful behaviors such as harassment. Threats to freedom of expression, being so closely linked to any enforcement action, can also be considered CCR risks.

Each category of CCR risk can carry different challenges, and they often require markedly different management approaches. For example, child sexual abuse material is illegal in all jurisdictions, while hate speech often is not (and can be significantly more complex for companies to identify). Therefore, an aggregated figure for content moderation decisions is unlikely to yield decision-useful information for investors interested in understanding company performance on specific CCR risks.

As proposed, this metric therefore requests that companies disclose several content moderation actions (number of content items removed, percentage of removed content discovered proactively, percentage of removed content appealed, percentage of appealed content restored, average user impressions of removed content) broken out by content category.

The Board also deliberated whether the Standard should define specific categories of CCR risks, or instead direct companies to disclose content moderation decisions in accordance with their existing content policies. While recognizing that there exists a trade-off with comparability, the Board recommends that companies choose their own content categories. This determination is based on research and consultations suggesting that (1) there is no consensus regarding definitions for most forms of harmful content, and (2) asking companies to collect content moderation data according to the definitions developed in the Standard would not be cost-
effective for companies. The Board also considered that this approach can more easily accommodate changes in policies and priority areas for harmful content, which is desirable given how rapidly this practice area is developing.

Rationale for specific data points in the content moderation table

62 **Number of content items removed**: Provides a rough idea of the scale of the company's content moderation actions.

63 **Percentage of removed content discovered proactively**: Measures how much infringing content a company discovers using its own systems, as opposed to how much is flagged by users or other third parties.

64 **Percentage of removed content appealed**: Indicates whether companies have an appeals process, while also providing insight into the degree to which users are pushing back on the company’s content moderation actions.

65 **Percentage of appealed content restored**: When paired with the number of appeals, may indicate the extent to which the company’s initial removal decisions were inaccurate. A large percentage here could also indicate a robust appeals process.

66 **Average user impressions of removed content**: Measures how quickly platforms are able to remove offending content before it reaches users. This metric also helps put removal figures into context regarding the potential harm caused: without this metric, all removals are counted equally, regardless of whether a content item was viewed one time or 100 million times.

Research and consultations indicate that there are a number of limitations to quantitative data regarding content moderation. Many participants noted, for example, that absent additional context, the number of content items removed in a given category is not decision-useful because (1) a single number does not indicate whether the content removed truly infringed the company’s policies, and (2) even if a platform’s moderation systems were 100 percent accurate in the decisions made, such a number does not provide insight into how much infringing content was not detected by the platform at all. There are similar considerations for each quantitative data point in the content moderation table. Moreover, market feedback stressed that the values do not necessarily scale to higher or lower risk (or opportunity)—that is, a higher or lower number in any category does not necessarily indicate better or worse performance on related issues.

68 During consultations, one company and one subject matter expert indicated concerns with a quantitative approach to metric design. The company stated its belief that given the rapid pace of change in management practices, any quantitative metrics would quickly be obsolete. The subject matter expert stated that because platforms are so different in objectives and format, quantitative comparisons would not be possible.
Despite these limitations, the Board believes that, on balance, this quantitative metric helps meet investors' informational needs and will lend a different type of insight into content moderation decisions than qualitative information can alone.

The Board believes that specific data points in the content moderation table can provide decision-useful information in two ways. First, they provide a baseline that can be used to measure an individual company’s performance over time on a granular set of issues. For example, an investor could observe a year-over-year increase in the number of proactive content removals within the hate speech category, combined with the knowledge that a key market had recently passed stricter laws around platform governance of hate speech, to gain a greater understanding of how a company is responding to evolving regulatory risks.

Second, the data points can provide a general sense of risks and opportunities associated with a platform’s purpose and design that can be compared across platforms, even if they are unlikely to be used to benchmark platforms against each other. For example, an investor may determine that one platform’s design leads to frequent appeals and reinstatements of content, whereas another platform fields a more limited number of appeals based on the type of content it hosts.

Rationale for recommended metric: Total amount spent on content governance

During consultations, investors consistently indicated their desire to understand the overall size of company investments into content governance efforts. At its March 2022 meeting, the Board deliberated whether the Standard should request a total dollar amount or the amount as a percentage of company expenses. Ultimately, the Board determined that a simple amount of expenses expressed in reporting currency would provide the most decision-useful information, as it allows investors to normalize the figure to their financial metrics of choice; an absolute amount may also be more useful when analyzing diversified companies with business lines that don’t involve user-generated content.

The Board considered that the relatively nascent state of this practice area and the natural variance in platform format and size may impede the comparability of disclosures between companies. Nonetheless, the Board determined that on balance, investor interest in this metric was strong enough to warrant its inclusion in the recommended changes to the Standard.

The Board also considered different levels of prescriptiveness in defining the expenses that preparers should include in their calculations. While the inclusion of more prescriptive guidelines might improve the comparability of disclosures, the Board considered the trade-off with cost-effectiveness (requiring preparers to calculate expenses differently from their current practices) and its representational faithfulness (content governance expenses for different kinds of platforms may include substantially different activities). The Board therefore recommends a balanced approach that prescribes some specific activities to be included in the scope of disclosure while requesting that preparers also disclose which activities were included in its calculation.
Rationale for recommended metric: Description of approach to identification and management of significant content- and content-related risks

This metric is intended to provide users with high-level qualitative information regarding the most significant risks related to the use of a company’s platforms. During consultations, companies and subject matter experts indicated that a high-level approach to the identification and management of CCR risks would yield comparable, decision-useful information, as each internet platform or service has differing levels of risk exposure depending on a variety of factors, including but not limited to the format of content enabled (e.g., text, audio, recorded video, livestreamed video) and the categories of potentially harmful content that could be disseminated (e.g., child exploitation, the promotion of violence, and self-harm). Investors and subject matter experts also specifically expressed interest in understanding how the geographic breadth of a company’s operations impacts its CCR risks, as global platforms must design systems that account for dozens of languages and with the understanding of cultural dynamics in regions all over the world.

In consultations, investors consistently expressed an interest in better understanding how companies broadly approach content governance, as this is a topic that many viewed as opaque and difficult to understand. Investors specifically expressed interest in understanding how the board and executive levels oversee the identification and management of CCR risks.

Rationale for recommended metric: Description of approach to content display and recommendations, and how these systems account for content- and conduct-related risks

This metric is intended to provide investors with information regarding how companies determine what users see on their platforms, and how these systems account for CCR risks. Participants in consultations stated their belief that quantitative disclosures in this area would not be useful, given the complexity and relative nascent nature of the subject matter.

Most companies have broad control over what users see on their platforms, and may design ranking and recommendation algorithms that consider hundreds of factors in determining what content is visible. These systems for content visibility and recommendations have been blamed by the media and regulators for a variety of societal ills, including but not limited to creating “filter bubbles,” radicalizing users by recommending increasingly extreme content, and harming the mental health of teenagers.

During consultations, many investors and subject matter experts stated that a focus on how companies set content policy and enforce against that policy would be overly narrow and insufficient in helping users understand the topic of content governance if it failed to address product design. Specifically, these participants suggested that a platform will be unable to mitigate significant CCR risks if the systems used by that platform continually lead users to harmful or objectionable content. In this scenario, participants suggested that no amount of effort focused on identifying infringing content will be sufficient.
While research and consultation indicate that the evidence linking platforms to the aforementioned social impacts is not firmly established, the Board recommends to include this aspect of content governance in a revised Standard because (1) investors expressed strong interest in this topic throughout consultations; (2) the evidence gathered during this project suggests that even the perception of platforms generating such negative social impacts poses significant financial risks for companies, especially through being a key driver of increasing regulatory scrutiny across the globe; and (3) the nuanced, qualitative approach recommended here could facilitate improved communication between companies and investors on a complex topic in which management strategies will vary significantly based on the services offered.

Rationale for recommended metric: *Description of approach to content moderation*

This proposed metric is intended to provide qualitative information regarding content moderation—defined as the set of activities associated with developing and enforcing the content policies of a company’s platforms. The metric is designed to capture a variety of approaches to content moderation, including whether enforcement models (1) rely on humans and/or automated technology, (2) rely on in-house or outsourced labor, and (3) rely on paid and/or volunteer moderators. The metric also addresses whether and how companies allow users to appeal content moderation decisions.

While investors noted that the details of enforcement models are complex, they emphasized the usefulness of understanding the overall structure of the various parts of companies’ content moderation efforts and how these parts worked together.

During consultations, participants indicated that both qualitative and quantitative information regarding content moderation is necessary to provide a complete picture of company performance, with many investors specifically expressing an interest in how content policy has been determined. Research and consultation indicate that platforms can serve a broad variety of functions and that therefore content policies can vary significantly: some platforms take an expansive approach to valuing user freedom of expression, while others have a narrower focus (such as providing services to children) that requires a more restrictive approach to permitted content.
Recommended Changes to the Internet Media & Services Standard

About the Recommended Changes

These recommended changes are presented for consideration by the International Sustainability Standards Board (ISSB) and other market participants interested in the continued development of the SASB Standards. **This version is not intended for implementation.**

The below does not include the entirety of the Standard. The following Summary of Recommended Changes to the Internet Media & Services Standard reflects the complete set of disclosure topics and associated metrics for this industry, after incorporating the recommended changes. The table is followed by the recommended new Content Governance & Freedom of Expression disclosure topic and corresponding metrics (with the exception of current metrics TC-IM-220a.5 and TC-IM-220a.6, which are recommended to be relocated to this new disclosure topic).

Prepared by the Sustainability Accounting Standards Board

May 2022
Summary of Recommended Changes to the Internet Media & Services Standard

**Recommended Changes to Sustainability Disclosure Topics & Accounting Metrics**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Accounting Metric</th>
<th>Category</th>
<th>Unit of Measure</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environmental Footprint of Hardware Infrastructure</td>
<td>(1) Total energy consumed, (2) percentage grid electricity, (3) percentage renewable</td>
<td>Quantitative</td>
<td>Gigajoules (GJ), Percentage (%)</td>
<td>TC-IM-130a.1</td>
</tr>
<tr>
<td></td>
<td>(1) Total water withdrawn, (2) total water consumed, percentage of each in regions with High or Extremely High Baseline Water Stress</td>
<td>Quantitative</td>
<td>Thousand cubic meters (m³), Percentage (%)</td>
<td>TC-IM-130a.2</td>
</tr>
<tr>
<td></td>
<td>Discussion of the integration of environmental considerations into strategic planning for data center needs</td>
<td>Discussion and Analysis</td>
<td>n/a</td>
<td>TC-IM-130a.3</td>
</tr>
<tr>
<td>Data Privacy, Advertising Standards &amp; Freedom of Expression</td>
<td>Description of policies and practices relating to behavioral advertising and user privacy</td>
<td>Discussion and Analysis</td>
<td>n/a</td>
<td>TC-IM-220a.1</td>
</tr>
<tr>
<td></td>
<td>Number of users whose information is used for secondary purposes</td>
<td>Quantitative</td>
<td>Number</td>
<td>TC-IM-220a.2</td>
</tr>
<tr>
<td></td>
<td>Total amount of monetary losses as a result of legal proceedings associated with user privacy</td>
<td>Quantitative</td>
<td>Reporting currency</td>
<td>TC-IM-220a.3</td>
</tr>
<tr>
<td></td>
<td>(1) Number of law enforcement requests for user information, (2) number of users whose information was requested, (3) percentage resulting in disclosure</td>
<td>Quantitative</td>
<td>Number, Percentage (%)</td>
<td>TC-IM-220a.4</td>
</tr>
<tr>
<td></td>
<td>List of countries where core products or services are subject to government-required monitoring, blocking, content filtering, or censoring</td>
<td>Discussion and Analysis</td>
<td>n/a</td>
<td>TC-IM-220a.5</td>
</tr>
<tr>
<td></td>
<td>Number of government requests to remove content, percentage compliance with requests</td>
<td>Quantitative</td>
<td>Number, Percentage (%)</td>
<td>TC-IM-220a.6</td>
</tr>
<tr>
<td>Content Governance &amp; Freedom of Expression</td>
<td>Content moderation table: (1) content items removed, (2) percentage of removed content discovered proactively, (3) percentage of removed content appealed, (4) percentage of appealed content restored, (5) average user impressions of removed content</td>
<td>Quantitative</td>
<td>Number, Percentage (%)</td>
<td>TC-IM-260a.1</td>
</tr>
<tr>
<td></td>
<td>Total amount spent on content governance</td>
<td>Quantitative</td>
<td>Reporting currency</td>
<td>TC-IM-260a.2</td>
</tr>
<tr>
<td></td>
<td>Description of approach to identification and management of significant content- and conduct-related risks</td>
<td>Discussion and Analysis</td>
<td>n/a</td>
<td>TC-IM-260a.3</td>
</tr>
<tr>
<td></td>
<td>Description of approach to content display and recommendations, and how these systems account for content- and conduct-related risks</td>
<td>Discussion and Analysis</td>
<td>n/a</td>
<td>TC-IM-260a.4</td>
</tr>
</tbody>
</table>

Note: TC-IM-220a.3 – The entity shall briefly describe the nature, context, and any corrective actions taken as a result of the monetary losses.
<table>
<thead>
<tr>
<th>Topic</th>
<th>Accounting Metric</th>
<th>Category</th>
<th>Unit of Measure</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Description of approach to content moderation</td>
<td>Discussion and Analysis</td>
<td>n/a</td>
<td>TC-IM-260a.5</td>
</tr>
<tr>
<td></td>
<td>List of countries where core products or services are subject to government-required monitoring, blocking, content filtering, or censoring</td>
<td>Discussion and Analysis</td>
<td>n/a</td>
<td>TC-IM-260a.6</td>
</tr>
<tr>
<td></td>
<td>Number of government requests to remove content, percentage compliance with requests</td>
<td>Quantitative</td>
<td>Number, Percentage (%)</td>
<td>TC-IM-260a.7</td>
</tr>
<tr>
<td>Data Security</td>
<td>(1) Number of data breaches, (2) percentage involving personally identifiable information (PII), (3) number of users affected</td>
<td>Quantitative</td>
<td>Number, Percentage (%)</td>
<td>TC-IM-230a.1</td>
</tr>
<tr>
<td></td>
<td>Description of approach to identifying and addressing data security risks, including use of third-party cybersecurity standards</td>
<td>Discussion and Analysis</td>
<td>n/a</td>
<td>TC-IM-230a.2</td>
</tr>
<tr>
<td>Employee Recruitment, Inclusion &amp; Performance</td>
<td>Percentage of employees that are foreign nationals</td>
<td>Quantitative</td>
<td>Percentage (%)</td>
<td>TC-IM-330a.1</td>
</tr>
<tr>
<td></td>
<td>Employee engagement as a percentage</td>
<td>Quantitative</td>
<td>Percentage (%)</td>
<td>TC-IM-330a.2</td>
</tr>
<tr>
<td></td>
<td>Percentage of gender and racial/ethnic group representation for (1) management, (2) technical staff, and (3) all other employees</td>
<td>Quantitative</td>
<td>Percentage (%)</td>
<td>TC-IM-330a.3</td>
</tr>
<tr>
<td>Intellectual Property Protection &amp; Competitive Behavior</td>
<td>Total amount of monetary losses as a result of legal proceedings associated with anti-competitive behavior regulations</td>
<td>Quantitative</td>
<td>Reporting currency</td>
<td>TC-IM-520a.1</td>
</tr>
</tbody>
</table>

20 Note to TC-IM-220a.5 – Disclosure shall include a description of the extent of the impact in each case and, where relevant, a discussion of the entity’s policies and practices related to freedom of expression.
21 Note to TC-IM-230a.1 – Disclosure shall include a description of corrective actions implemented in response to data breaches.
22 Note to TC-IM-330a.1 – Disclosure shall include a description of potential risks of recruiting foreign nationals and management approach to addressing these risks.
23 Note to TC-IM-330a.2 – Disclosure shall include a description of methodology employed.
24 Note to TC-IM-330a.3 – The entity shall discuss its policies and programs for fostering equitable employee representation across its global operations.
25 Note to TC-IM-520a.1 – The entity shall briefly describe the nature, context, and any corrective actions taken as a result of the monetary losses.
Content Governance & Freedom of Expression

Topic Summary

Many companies in the Internet Media & Services industry offer products and services that rely on user-generated content. Users of these platforms can create content or engage in online conduct that creates negative societal impacts, including but not limited to the incitement of violence, the depiction or glorification of terrorist acts, the dissemination of child sexual abuse materials, or the harassment of other users. Most companies choose to suppress certain types of content and conduct to attract a large base of users and advertisers. However, when making content moderation decisions, companies risk alienating their user base or suppressing free expression in publicly accessible forums. Platforms therefore often establish policies for acceptable content and conduct, and design systems to enforce these policies. Some platforms also offer users a chance to appeal content they feel was wrongly removed. Additionally, companies may need to comply with local laws or government demands to remove sensitive content or services for cultural or political reasons. In order to retain users and advertisers and maintain their license to operate, companies must balance moderation activities with policies and procedures that appropriately account for concerns regarding user freedom of expression. This issue impacts company profitability through expenses such as personnel costs and investments in automated systems. Reputational damage related to controversial or ineffective moderation policy or enforcement may also lead to increased regulatory scrutiny or impact companies’ ability to attract users and advertisers, leading to reduced revenues. Conversely, companies that have a track record of maintaining a platform that is appealing to users while being “brand safe” for advertisers may be better positioned to attract more users and advertisers, and therefore improve financial performance.

Accounting Metrics (new metrics only)

TC-IM-260a.1 Content moderation table: (1) number of content items removed, (2) percentage of removed content discovered proactively, (3) percentage of removed content appealed, (4) percentage of appealed content restored, (5) average user impressions of removed content

1 The entity shall disclose metrics related to the removal of content, broken down by various categories of content that are prohibited by the entity’s content policies.

1.1 Content policies are defined as the policy or set of policies (such as terms of service or community guidelines) that clarify what kinds of content and behavior are not allowed on the entity’s platform or in various groups or subgroups on the platform.

1.2 Categories listed in the table should match the categories listed in the entity’s content policies.

1.2.1. The entity may omit categories of content that are prohibited by its platform but not directly associated with significant harm to individuals or society, such as certain types of spam, fake accounts, or consensual adult nudity.

1.3 The entity shall count each removed content item under only one category to avoid double counting. If a content item was removed for violating multiple policies, the entity shall use its discretion in determining which category is most appropriate.
For each category, the entity shall disclose: (1) number of content items removed, (2) percentage of removed content discovered proactively, (3) percentage of removed content appealed, (4) percentage of appealed content restored, and (5) average user impressions of removed content.

2.1 The entity shall disclose the total number of content items it removed from its platform deemed to be in violation of its content policies.

2.1.1 A content item is defined as a single piece of digital content, including but not limited to one post, comment, image, video, or audio file.

2.2 The percentage of removed content discovered proactively shall be calculated as the number of content items removed that are discovered proactively divided by the total number of content items removed.

2.2.1 Content items discovered proactively are defined as content that is identified by the entity’s systems as being in violation of its policies, as opposed to content reported to the entity by its users or external parties.

2.2.2 Content discovered proactively may include content identified by technology tools such as machine learning or hash recognition, or content identified by human moderators prior to being reported by users.

2.3 The percentage of removed content appealed shall be calculated as the number of content item removals that were appealed via the entity’s appeals process, divided by the number of content item removals, should such an appeals process exist.

2.3.1 The entity shall disclose “n/a” if no appeals process exists.

2.4 The percentage of appealed content restored shall be calculated as appealed content restored divided by the number of content removals appealed.

2.4.1 Appealed content restored is defined as content that was previously removed by the entity and then was subsequently restored to the platform after a successful user appeal via the entity’s appeals process, should such a process exist.

2.4.2 The entity shall disclose “n/a” if no appeals process exists.

2.5 Average user impressions of removed content shall be calculated as the total number of user impressions of content that was subsequently removed by the entity divided by the total number of content items removed.

2.5.1 A user impression is defined as an instance of an item of content appearing on the screen of a user of the entity’s platform.

3 The entity should disclose this information in the following table format:

<table>
<thead>
<tr>
<th>(a) Content category (entity-defined)</th>
<th>(b) Number of content items removed</th>
<th>(c) Percentage of removed content discovered proactively</th>
<th>(e) Percentage of removed content appealed</th>
<th>(f) Percentage of appealed content restored</th>
<th>(g) Average user impressions of removed content</th>
</tr>
</thead>
</table>

© 2022 VALUE REPORTING FOUNDATION
E.g., terrorist and violent extremist content

E.g., harassment

**TC-IM-260a.2 Total amount spent on content governance**

1. The entity shall disclose the total amount spent on business activities directly associated with the identification and management of content- and conduct-related risks on its platform(s).

1.1 Content- and conduct-related risks (CCR risks) are defined as reasonable threats to the well-being of people or society stemming from certain illegal, dangerous, or otherwise harmful content or behavior, including risks to human rights, on the entity’s platform, which may be prohibited by relevant policies and terms of service.

1.1.1 CCR risks may include, but are not limited to, the dissemination of child sexual abuse material or terrorist and violent extremist content on the entity’s platform, abuse and harassment of a platform’s users, and the incitement of violence on the entity’s platform.

1.2 Business activities may include but are not limited to the entity’s creation and enforcement of its content policies and activities associated with addressing CCR risk in the entity’s content display systems.

1.2.1 Content display systems are defined as the mechanisms that determine what pieces of content to display or recommend to individual users, and in what order.

1.3 The entity shall disclose a summary of the business activities and expenses included in its calculation.

1.3.1 The summary shall specify the major teams/departments, programs (including outsourced programs), and initiatives that support the identification and management of CCR risks.

**TC-IM-260a.3 Description of approach to identification and management of significant content- and conduct-related risks**

1. The entity shall disclose its most significant content- and conduct-related risks associated with its platform, as well as the most significant challenges it faces when responding to and managing these risks.

1.1 Content- and conduct-related risks (CCR risks) are defined as reasonable threats to the well-being of people or society stemming from certain illegal, dangerous, or otherwise harmful content or behavior, including risks to human rights, on the entity’s platform, which may be prohibited by relevant policies and terms of service.

1.1.1 CCR risks may include, but are not limited to, the dissemination of child sexual abuse material or terrorist and violent extremist content through the entity’s services, and the abuse and harassment of a platform’s users.
1.2 Significant risks are defined as those that could produce substantial harm to individuals or society, especially to vulnerable individuals or groups, that may be heightened by the use or misuse of entity’s platform.

1.3 Significant challenges are those that present notable management, investment, or compliance challenges from a business perspective, and may include challenges from operating in jurisdictions with complex or demanding regulatory landscapes.

2 The entity shall disclose its most significant CCR risks associated with (a) the format of its platform, (b) the effects of categories of harmful content on its platform, and (c) the diverse and/or global nature of its user base.

2.1 Platform format is defined as the type of digital content (e.g., text, image, audio, video, livestream) hosted by the platform and the ways that this content may be shared and presented to other users.

2.2 Categories of harmful content may include issues such as child exploitation, terrorism and violent extremism, the promotion of violence, hate speech, harassment, self-harm, misleading or false information, and other content that could lead to significant harm to individuals or society.

2.3 CCR risks associated with the diverse and/or global nature of an entity’s user base are defined as specific risks associated with regions, languages, demographic groups, cultures, geopolitical situations, and global events.

2.3.1 The entity shall disclose which regions, languages, demographic groups, cultures, geopolitical situations, and global events present heightened CCR risks associated with its platform.

3 The entity shall describe its approach to identifying and managing the most significant CCR risks associated with its platform, including the governance structure, criteria, and controls it uses to ensure successful identification and management of those risks.

3.1 Criteria are defined as targets, goals, or key performance indicators that the entity uses to measure performance in its identification and management of significant CCR risks associated with its platform.

3.2 Controls are defined as the internal processes, management, and organizational structures that ensure identification and management of significant CCR risks.

3.3 The scope of disclosure shall include risks from ongoing business practices, as well as those that arise from new business practices.

3.3.1 New business practices may include launching new products and services, expanding products and services to a new geographic region, a recent merger/acquisition or joint venture/partnership that impacts products or services, or other factors that might increase the likelihood, scope, or severity of CCR risks.

3.3.2 The entity shall describe its use of established tools and frameworks, such as human rights impact assessments, to evaluate new business practices.
3.4 The scope of disclosure shall include ongoing CCR risks, as well as those that change over time because of external impacts, such as risks associated with regional conflicts or other global events.

3.4.1 The entity shall discuss whether and how it shifts resources to contend with emerging or evolving CCR risks.

3.4.2 The entity may present an example of how it shifted resources to contend with a real or theoretical emerging or evolving CCR risk.

3.5 The scope of disclosure shall include a discussion of how the entity manages and oversees the general business practices used to address CCR risk, including (a) adjusting content display systems, (b) creating content policies, and (c) enforcing content policies.

3.5.1 Content display systems are defined as the mechanisms that determine which pieces of content to display or recommend to individual users, and in what order.

3.5.2 Content policies are defined as the policy or set of policies (such as terms of service or community guidelines) that clarify which kinds of content and behavior are not allowed on the entity’s platform or in various groups or subgroups on the platform.

3.6 The entity shall disclose the executive role(s) or position(s) with significant decision-making authority in its management of CCR risks, as well as whom that (those) executive(s) report to.

3.7 The entity shall disclose whether and how its efforts to address CCR risks are understood and/or overseen at the executive level and by the board of directors.

3.8 The entity shall describe how it assesses and improves the quality of its identification and management of CCR risks over time.

4 The entity shall discuss the availability and experience of its platform for users of different age ranges.

4.1 The entity shall disclose whether and how its products and/or experiences vary for users of different age ranges, particularly younger users.

4.2 If the entity prohibits users of certain ages on its platform, or offers a substantially different product or experience to younger users, it shall describe its mechanisms for verifying the ages of users and for enforcing its age-related policies and practices.

5 The entity shall disclose its engagement with (a) industry coalitions, (b) third-party experts, (c) governments and regulators, and (d) affected communities to identify and manage its most significant CCR risks.

5.1 Industry coalitions are defined as efforts by companies to share best practices and cooperate in response to certain types of CCR risks.

5.2 Third-party experts are defined as academics, researchers, non-governmental organizations, advocacy groups, and others with expertise in CCR risks.
5.3 Affected communities are defined as groups of people who by their nature, location, or demographic characteristics face elevated impacts from CCR risks on internet platforms.

TC-IM-260a.4 Description of approach to content display and recommendations, and how these systems account for content- and conduct-related risks

1 The entity shall describe the general principles it uses to design and weight its content display systems.

   1.1 Content display systems are defined as the mechanisms that determine which pieces of content to display or recommend to individual users, and in what order.

   1.2 The entity shall disclose what it considers to be desirable or “high quality” content.

   1.3 The entity may describe values or norms of behavior it aims to promote on its platform, and how it identifies and promotes content or behavior that reinforces those values.

2 The entity shall describe how it identifies and manages significant content- and conduct-related risks related to the design of its content display systems.

   2.1 Content- and conduct-related risks (CCR risks) are defined as reasonable threats to the well-being of people or society stemming from certain illegal, dangerous, or otherwise harmful content or behavior, including risks to human rights, on the entity’s platform, which may be prohibited by relevant policies and terms of service.

      2.1.1 CCR risks may include, but are not limited to, the dissemination of child sexual abuse material or terrorist and violent extremist content through the entity’s services, and the abuse and harassment of a platform’s users.

   2.2 Significant CCR risks may include situations in which users could be harmed by seeing or interacting with certain types of content, even if that content does not violate the entity’s policies.

   2.3 The entity shall describe how it assesses and improves outcomes related to CCR risks in its content display systems over time.

TC-IM-260a.5 Description of approach to content moderation

1 The entity shall describe its approach to content moderation.

   1.1 Content moderation is defined as the set of activities associated with creation and enforcement of content policies, including how the entity defines content or behavior that is prohibited on its platform, identifies this content across the platform, and determines whether to take action on it.

      1.1.1 Content policies are defined as the policy or set of policies (such as terms of service or community guidelines) that clarify which kinds of content and behavior are not allowed on the entity’s platform or in various groups or subgroups on the platform.

2 The entity shall describe its processes for developing, maintaining, and updating its content policies.
2.1 The entity shall describe how it ensures that its content policies cover and appropriately account for significant content- and conduct-related risks.

2.1.1 Content- and conduct-related risks (CCR risks) are defined as reasonable threats to the well-being of people or society stemming from certain illegal, dangerous, or otherwise harmful content or behavior, including risks to human rights, on the entity’s platform, which may be prohibited by relevant policies and terms of service.

2.1.2 CCR risks may include, but are not limited to, the dissemination of child sexual abuse material or terrorist and violent extremist content through the entity’s services, and the abuse and harassment of a platform’s users.

2.1.3 Significant risks are defined as those that could produce substantial harm to individuals or society, especially to vulnerable individuals or groups, that are linked to the use of entity’s platform.

2.2 The entity shall describe how it adjusts its content policies to account for new and emerging CCR risks.

3 The entity shall describe how it enforces its content policies.

3.1 The entity shall describe its content moderation enforcement model.

3.1.1 A content moderation enforcement model is defined as a structure for identifying and taking action on potential violations of content policies that includes some combination of the following mechanisms: (a) humans and/or automated technology; (b) in-house and/or outsourced labor; and (c) paid and/or volunteer moderators.

3.1.2 The entity shall describe the general role(s) of each enforcement mechanism and how the enforcement mechanisms interact and work together to perform content moderation.

3.1.3 The entity shall describe how the components of its content moderation model are overseen and managed within its enforcement structure.

3.2 The entity shall disclose the range of enforcement actions it may take on (a) content items and (b) users who violate its content policies, and the threshold of unacceptable (a) content and (b) behavior that would trigger each enforcement action.

3.3 The entity shall describe its process for addressing borderline content moderation decisions.

3.3.1 A borderline decision is defined as one in which people tasked with content review might disagree if the content or behavior violates the entity’s content policies.

3.3.2 The entity may disclose an example of how a real or theoretical borderline content moderation decision might be handled or escalated through its systems.

3.3.3 The entity shall disclose its executive or bodies with final decision-making authority on borderline decisions.
3.4 The entity shall disclose whether there are exceptional circumstances in which it might contravene its own content policies, and how those exceptional circumstances may be determined.

4 The entity shall describe its transparency and review efforts around its content moderation practices.

4.1 The entity shall describe the extent to which and the methods by which it provides notice to users whose content or conduct is at issue in an enforcement action.

4.2 The entity shall describe the processes by which it proactively identifies enforcement actions on (a) content items and (b) users that may have been made in error, should such processes exist.

4.3 The entity shall describe the processes by which it allows users or third parties to appeal its enforcement actions on (a) content items and (b) users, and how it determines whether to uphold an appeal and, for example, restore a content item or user permission, should such processes exist.