Content Moderation Taxonomy:

A Foundation for Standard Setting on the Issue of Content Moderation

Introduction

This document has been prepared by SASB's technical staff as part of the research project, *Content Moderation on Internet Platforms*.[1] It outlines a taxonomy of the various social externalities related to content moderation and the associated business activities in SASB's Technology & Communications sector to which these social externalities may apply. In other words, this document shifts the complex and often ambiguous area of content moderation into a structured set of issues and business activities that have been defined through SASB's classification of sustainability issues and industries.

This document was prepared using a combination of research (news reports, academic papers, corporate regulatory filings and press releases, etc.) and consultation with subject matter experts, company representatives, and investors. The technical staff at SASB welcomes feedback on the taxonomy and additions to the body of evidence cited.

This document does not alter the SASB Standards, nor does it establish staff or Standards Board views on whether standard-setting should be pursued. Rather, it establishes a foundation upon which such decisions can be made when combined with further research and market input.

This document was released on November 12, 2020. SASB may choose to update the document over time if new evidence emerges that would suggest changes to this taxonomy are necessary.

[1] Interested stakeholders can sign up for updates and see other information at the project page on SASB's website: https://www.sasb.org/standard-setting-process/active-projects/content-moderation-on-internet-platforms-research-project/

Table of Contents

Executive Summary

What is Content Moderation?

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. The existence of harmful content on the internet leads to platforms deciding which types of content they are willing to host or amplify; it also leads to some companies deciding which types of customers to do business with.

“Content moderation” generally refers to the processes and procedures used to detect and potentially action a range of illegal or unwanted content, especially at social media platforms. Using a broader lens that includes the systems used to determine what users see on a given platform (such as algorithms that prioritize or recommend content), a more encompassing label for these activities might be “content governance.” A variety of companies operating at deeper levels of internet infrastructure, such as cloud services, website builders, or internet service providers, make decisions that are akin to content moderation or content governance.

Issues Associated with Content Moderation

SASB documents the social issues surrounding content moderation as follows:

1. **Harmful Content**
   There are negative social externalities associated with a range of content created, stored, and disseminated on the internet. As discussed in further detail in this document, this category includes content related to child exploitation, terrorism and violent extremism, the promotion of violence and other hate speech, coordinated disinformation campaigns, and some forms of misinformation. Depending on the jurisdiction in question, harmful content may be illegal and therefore the subject of additional responsibility for platforms.
2. **Freedom of Expression**

   In order to mitigate harm (or the risk of harm) that comes from certain types of content, platforms react by removing content, ceasing to provide certain customers with services, or limiting content visibility through different strategies and approaches. These actions therefore have an impact on user freedom of expression.
3. **Privacy**

   Platform actions to monitor their customers or users for harmful or illegal activity may go against expectations of privacy. Some social media platforms also collect and use sensitive personal information to determine which content users see, and in what order.
4. **Worker Health & Safety**

   Another consequence of platforms deciding to monitor and remove harmful content is concentrated worker exposure to this content. Emerging evidence suggests that addressing the mental health consequences of this important work—including post-traumatic stress disorder—is a relevant human capital management theme for platforms, especially for those monitoring large volumes of user-generated content. Much of this work is contracted through third-party vendors, which may complicate platforms’ management of this issue.

SASB Taxonomy of Content Moderation Issues

SASB Industry		**Harmful Content**	**Freedom of Expression**	**User Privacy**	**Employee Health & Safety**
Internet Media & Services	**Social Media Platforms**	■	■	■	■
	Messaging Applications	■	■	■	
Software & IT Services	**Gaming Platforms & Publishers**	■	■	■	
	Internet Infrastructure & Cloud Services	■	■	■	
	Business Process Outsourcing				■
Telecommunication Services	**Internet Service Providers**	■	■	■	
E-Commerce	**Online Marketplaces**	■	■	■	

Gray shade represents areas where staff has identified evidence that a content moderation issue may be applicable to the business activity

Associated Business Activities

In reviewing the types of businesses to which the aforementioned issues apply, staff identified several applicable industries within SASB's Sustainable Industry Classification System® (SICS®). These industries were further segmented into core "business activities."

1. **Internet Media & Services**

    a. *Social Media Platforms*

    Social media platforms facilitate the sharing and dissemination of user-generated content through a variety of media, including text and video. These platforms are exposed to the harmful content, freedom of expression, user privacy, and worker health and safety issues. The extent to which each issue applies appears to be related to the platform's scale. Platforms with large user bases may significantly amplify harmful content and therefore be expected to have more wide-reaching and effective content governance policies; a high volume of harmful content may also expose content reviewers to heightened mental health hazards.

    b. *Messaging Applications*

    Messaging applications facilitate one-on-one or group communication and can be stand-alone services or embedded inside other products, such as social media or gaming platforms. These applications can enable the sharing of harmful content. Any content governance mechanisms in this area may stand in opposition to user expectations around security, privacy and freedom of expression.

    c. *Other Internet Platforms*

    As part of a broader review of questions related to internet platform safety, staff reviewed other platforms that facilitate physical human interaction, including ride-hailing, dating, and vacation rental applications. The practices and procedures used to manage safety at these platforms generally involve the verification and vetting of platform participants, as well as the veracity of the information and legality of services they provide, which are significantly different management strategies from those associated with content moderation themes.

2. **Software & IT Services**

    a. *Infrastructure & Cloud Services*

    These companies provide a range of services that facilitate the storage, hosting and delivery of content on the internet. Traditionally, these companies have been viewed as neutral intermediaries and therefore have not been held to the same expectation of content governance as social media platforms. However, there is evidence of a growing demand for these companies to do more to combat harmful content that is stored on or travels across their networks, such as child exploitation and hate speech.

b. Game Platforms & Publishers

Game publishers and platforms have also come into the public conversation surrounding harmful content as video games migrate online and increasingly facilitate user interaction. The nature of this interaction is highly dependent on the game in question and may occur on messaging platforms not controlled by the game publishers or platforms themselves.

c. Business Process Outsourcing

A number of companies provide business process outsourcing services to major social media platforms to help monitor and remove harmful user-generated content. There is a growing body of evidence suggesting there are negative mental health impacts on the workers that review and remove harmful content, suggesting that the provision of this service may expose these companies to human capital management challenges pertaining to employee health and safety.

3. Telecommunication Services

a. Internet Service Providers

Internet service providers (ISPs) sit a layer below cloud service providers in the overall infrastructure of internet delivery in that they transmit bits of data from point A to point B without storing or hosting data. While generally viewed as a neutral layer of the internet, ISPs are also facing growing calls to do more to help combat illegal and other harmful content that travels across their networks. As with messaging applications, actions that ISPs take to monitor users or their content, such as deep packet inspection, may stand in opposition to societal expectations around security and privacy.

4. E-Commerce

a. Online Marketplaces

Online marketplaces are “two-sided” platforms that link buyers with third-party sellers. The practices and procedures used to enforce Terms of Service in this environment and ensure product safety—which generally revolve around verifying and vetting platform participants such as third-party sellers—appear to be distinct from the type of content governance activities undertaken by social media or gaming platforms. Nonetheless, some online marketplaces may perform content moderation (or actions similar to it), such as the monitoring of self-published e-books and other analogs to user-generated content.

Harmful Content

There are negative social externalities associated with a range of content created, stored, and disseminated on the internet that can be classified as harmful content. This content ranges from the illegal (e.g., content depicting the sexual exploitation of children) to legal but debatable areas (e.g., misinformation that could potentially cause harm). Businesses provide services that can facilitate the existence or amplification of harmful content in a variety of ways.

Below is a description of several types of harmful online content that a variety of technology companies are grappling with. This section is not intended to be an exhaustive list of potential types of harmful content that can be accessed or shared online, but rather an illustration of some broader negative social externalities associated with the use of modern technology.

Child Sexual Abuse Material

Content containing or coercing the sexual exploitation of children—often called child pornography and increasingly referred to as child sexual abuse material (CSAM)—is harmful both to depicted victims and viewers. Thorn.org, a nonprofit that builds technology aimed at preventing the spread of CSAM online, notes that CSAM is shared, traded, and sold through a variety of forms of internet technology, including "websites, email, instant messaging/ICQ, Internet Relay Chat (IRC), newsgroups, bulletin boards, peer-to-peer networks, internet gaming sites, social networking sites, and anonymized networks."[2] CSAM is not just shared, but also solicited and coerced through online grooming or blackmailing on a variety of platforms.

Several studies have cited the challenges of quantifying the amount of this material that is shared/viewed across the internet as well as the number of individuals involved, due in part to different definitions of the material and the nature of tracing content on the internet, including the "dark web" or encrypted messaging services.[3] Although definitive numbers are hard to come by, CSAM is widespread on the internet and appears to be growing at an alarming rate. The Internet Watch Foundation, a nonprofit collaboration between industry and the European Commission, found over 130,000 instances of confirmed CSAM content across nearly 5,000 domains in in 2019.[4] The National Center for Missing & Exploited Children (NCMEC)—to which companies are required to report identified content in accordance with US law—states on its website that it has received over 65 million tips since its CyberTipline was established in 1998.[5] The world's largest social media platform reported 37.4 million items of content being actioned that fell under its "child nudity and sexual exploitation" policy in 2019 alone, with the vast majority of this material identified prior to being seen by any users.[6]

Given its inherently abhorrent nature, CSAM is one area of harmful content where progress has been made in terms of industry collaboration. One popular tool used across the internet is PhotoDNA, a forensic software that creates a digital fingerprint (often called a "hash") of an

---

[2] Thorn.org website, accessed May 20, 2020.

[3] Rapid Evidence Assessment: Quantifying the Extent of Online-Facilitated Child Sexual Abuse: Report for the Independent Inquiry into Child Sexual Abuse. University of Huddersfield, January 2018.

[4] Internet Watch Foundation Annual Report 2019. Internet Watch Foundation, March 2020.

[5] NCMEC Website. Accessed May 19, 2020.

[6] Facebook Community Standards Enforcement Report. Accessed June 1, 2020.

image and automatically scans it against a database of known child exploitation imagery.[7] While this technology is an important tool, there are some shortfalls; most importantly, hashing can only flag previously identified CSAM. According to a recent *New York Times* investigation, not all major cloud storage platforms scan for CSAM, and none of them does so at the point of upload: rather, images must be shared, meaning viewers of the content can evade detection by sharing login information.[8] One major US-based telecommunications provider states on its website that it uses PhotoDNA and other tools alongside human screening to detect CSAM.[9]

Platforms are also being exploited by would-be predators and viewers of CSAM in more complex ways. For example, several major advertisers left a major video-sharing platform after predators were found to be exploiting the comments section of seemingly innocent videos.[10] [11] In South Korea, a national scandal erupted at the end of 2019 when a ring of criminals was found to have been blackmailing and extorting children for explicit photos and live video chats on one of the country's most prominent messaging apps.[12]

CSAM is also an area of relatively developed legal regimes globally: various laws are in place that mandate reporting of CSAM imagery to law enforcement or child protection groups.[13] In the United States, there is also a new bipartisan legislative initiative in the Senate named the EARN IT Act, which would revoke online platforms' liability protections if they are deemed to not be sufficiently addressing concerns around CSAM.[14]

Recently, some investors have begun asking technology companies to do more specifically to prevent the online spread of CSAM. In 2019, a major US-based telecommunications provider was subject to a shareholder resolution requesting that the company release a report on the potential sexual exploitation of children through its products and services.[15] The resolution drew support from over 30 percent of votes cast at the meeting.

As discussed in further detail in the "User Privacy" section of this document, efforts to combat the sharing of CSAM are significantly complicated by privacy-centric product design.

Terrorist and Violent Extremist Content

This content—which is frequently referred to as TVEC—is another form of harmful content that is spread over the internet. This content includes footage of violence, the promotion and celebration of real-world acts of violence, and the planning of violent acts.

---

[7] "Child Abusers Run Rampant as Tech Companies Look the Other Way." *The New York Times.* November 9, 2019. [**Warning**: contains graphic descriptions of acts of child sexual abuse]
[8] Ibid.
[9] "Verizon's Efforts to Combat Online Child Exploitation." Company website, accessed June 1, 2020.
[10] "YouTube Still Can't Stop Child Predators in its Comments." *The Verge*. February 19, 2019.
[11] "Advertisers Boycott YouTube After Pedophiles Swarm Comments on Videos of Children." *The New York Times.* February 20, 2019.
[12] "South Korea's Latest Sex Crime Scandal is a Blackmail Ring Streaming Abuse on Telegram." *Quartz.* March 24, 2020.
[13] Laws include US §2258A from the 2008 PROTECT Our Children Act and the 2011 EU Directive on combating the sexual abuse and sexual exploitation and child pornography.
[14] "Graham, Blumenthal, Hawley, Feinstein Introduce EARN IT Act to Encourage Tech Industry to Take Online Child Sexual Exploitation Seriously". *U.S. Senate Press Release.* March 5, 2020.
[15] SEC Proxy Memorandum filed on form PX14A6G.

The violent extremist group ISIS was able to recruit more than 40,000 fighters to its cause following its 2014 declaration of a "caliphate."[16] Much of the group's success has been attributed to its sophisticated use of the internet, including social media platforms. Since that time, technology companies have become more active in collaborating to address the spread of TVEC, including through the formation of a cross-industry group, the Global Internet Forum to Counter Terrorism (GIFCT). This group states that because different companies have different definitions of what may constitute terrorist content, they have instead focused on a common "hash library" of known TVEC with over 200,000 pieces of content identified to date; 85 percent of this content is categorized as the "glorification of terrorist acts."[17]

Another terrorist event that has had a dramatic impact on the technology sector was the livestreaming of the Christchurch shooting—a massacre of 51 unarmed civilians in New Zealand—on a major social media platform in March of 2019. The shooter also posted a video manifesto on several other prominent sites prior to committing the atrocity. The video was watched 4,000 times before being removed, but perhaps more disturbing were the efforts of bad actors immediately in the wake: the platform reported removing 1.5 million re-uploads of the footage of the shooting in the 24 hours afterward, with uploaders taking various approaches in their attempts to evade detection by automated systems.[18] The footage was cross-posted to various other social media platforms as well.[19]

Governments have begun responding: following the shooting, the Prime Minister of New Zealand and President of France announced the Christchurch Call to Action, which it describes as a "commitment by Governments and tech companies to eliminate terrorist and violent extremist content online."[20] Legislation is beginning to follow as well. For example, Australia quickly passed the *Sharing of Abhorrent Violent Material Act*, which requires ISPs, social media platforms, websites, and cloud solutions providers (but does not apply to email, messaging, SMS, or online gaming providers) to expeditiously report TVEC to Australian authorities.[21] In 2018, the EU proposed new rules governing terrorist content online as well.[22] Concerns with this type of legislative approach to content moderation are widespread and covered further in the "Content Governance and Freedom of Expression" section of this document.

Hate Speech

While hate speech can be difficult to define at the margins, there is a compelling body of evidence that dehumanizing messages spread online can and do lead to real-world harm. More generally speaking, the technologies that have enabled communities to find like minds, share content, and organize are also available to organized hate groups.

---

[16] "IS Foreign Fighters: 5,600 Have Returned Home." *BBC News.* October 24, 2017.
[17] GIFCT Transparency Report, accessed May 20, 2020.
[18] "Facebook Says That it Removed 1.5 Million videos of the New Zealand Mass Shooting." *The Verge*. March 19, 2019.
[19] "The Tricky Task of Policing YouTube." *The Economist.* May 4, 2019.
[20] Christchurch Call website. Accessed May 20, 2020.
[21] "Unprecedented Penalties for Enabling the Sharing of Abhorrent Violent Material Online." *Baker McKenzie (online memo).* April 30, 2019.
[22] EU Proposal for Regulation on Preventing Dissemination of Terrorist Content Online. Accessed May 26, 2020.

In November 2018, a major social media platform stated in a blog post that, "prior to this year, we weren't doing enough to help prevent our platform from being used to foment division and incite offline violence."[23] This announcement was made following an independent human rights impact assessment conducted in accordance with the UN's guiding principles on business and human rights. The company was referring to alleged ethnic cleansing of a minority group in Myanmar; military leaders there were involved in a coordinated operation to spread falsehoods and dehumanizing language against Rohingya Muslims for years.[24] Reportedly, the platform hired its first two content moderators fluent in Burmese in 2015; this lack of qualified personnel made enforcement of the platform's policies around hate speech impossible given the millions of active users speaking the language.[25]

Another example of how technology can be used to exploit ugly social undercurrents can be seen in India, where a popular messaging app has been blamed in helping fan ethnic tensions through the spread of false rumors that routinely spill into mob violence.[26]

Online gaming is another area where harmful online behavior has emerged. While for decades some have blamed video games for encouraging violence, a different social issue has taken form as gamers increasingly migrate online and begin interacting with each other: hate speech and harassment. In a blog post last year, the head of a major gaming platform recognized "a growing toxic stew of hate speech, bigotry, and misogyny" in digital life and pledged to expand content moderation resources across the company's services.[27] In private consultations with game publishers, SASB staff was told that efforts to address harassment and bullying are complicated by the fact that many of the communications between gamers are happening on messaging or social media platforms outside of the game publishers' or platforms' control.

Extreme examples of hate speech and the incitement of violence have also led to more traditionally "neutral" layers of the internet taking action. For example, in August 2017, a white supremacist rally in Charlottesville, Virginia, turned violent when it was met by counter-protestors. During the ensuing political furor, attention was drawn to a network of websites and online groups espousing holocaust denial and other neo-Nazi ideology, including a publication, *The Daily Stormer*, which had mocked the death of a counter-protestor in the immediate aftermath of the Charlottesville violence.[28] In response to heightened public scrutiny, a number of platforms, including cloud service providers and domain registration providers—generally companies that tend not to action offensive content—ceased providing services to *The Daily Stormer,* effectively removing it from the open internet.[29] A popular chat application disabled a number of servers hosting white supremacists in response to the violence at the rally, while payment processing platforms also banned a number of far-right figures from using their services.[30]

---

[23] Company Blog Post. November 5, 2019.
[24] "A Genocide Incited on Facebook, With Posts from Myanmar's Military." *The New York Times.* October 15, 2018.
[25] "Why Facebook is Losing the War on Hate Speech in Myanmar." *Reuters.* August 15, 2018.
[26] "How WhatsApp Fuels Fake News and Violence in India." *Wired*. December 12, 2018.
[27] Company Blog Post. May 20, 2019.
[28] "The Neo-Nazis of the Daily Stormer Wander the Digital Wilderness." *The New Yorker.* May 21, 2018.
[29] Ibid.
[30] "Tech Companies Turn on Daily Stormer and the 'Alt-Right' after Charlottesville." *The Guardian.* August 14, 2017.

A similar process unfolded with the "de-platforming" of 8chan, a forum known for its users actively encouraging violence and racism following mass shootings in the US cities of El Paso, Texas, and Dayton, Ohio.[31] The cases of *The Daily Stormer* and 8chan are illustrative examples of the challenges that companies operating in different layers of the internet must confront when crafting policy around content that is not illegal in many jurisdictions but still deeply objectionable.

Disinformation

Disinformation—defined here as the *intentional* spread of false information or narratives to pursue a specific goal—is another form of harmful content that nearly all would agree should not be spread over the internet. Awareness of disinformation, especially that propagated by state actors, has risen in prominence largely since the 2016 US Presidential election and the UK's referendum on leaving the European Union. Staff's research and consultation indicate that concerns here are more acutely focused on social media platforms, as opposed to the broader internet ecosystem. This is likely due to the large user bases of prominent platforms and their associated power to amplify desired narratives.

Following an investigation, U.S. intelligence agencies stated in a report to the Select Committee on Intelligence of the US Senate that "Russian operatives associated with the St. Petersburg-based Internet Research Agency (IRA) used social media to conduct an information warfare campaign designed to spread disinformation and societal division in the United States."[32] While much has been made of foreign interference in democratic processes across the globe, there is also evidence that similar tactics are increasingly being used by domestic actors.[33] One recent study from Oxford University found that organized social media manipulation doubled from 2017 to 2019, with 45 democracies and 26 authoritarian states using "computational propaganda tools" to manipulate public opinion.[34]

Some social media platforms group this state-sponsored disinformation, along with other activities such as scams or spam campaigns involving networks of fake accounts, into something referred to as "coordinated inauthentic behavior."

Misinformation

While many platforms are reticent to police the opinions of users and some misinformation is innocuous, misinformation that could directly and demonstrably lead to harm is a form of harmful content. More broadly, the spread of "fake news" over the internet—in the form of both misinformation and disinformation—is a central concern for many policymakers. While misinformation is separated from disinformation by the intent of the person spreading it, the negative societal externality remains the same. As discussed further under "Freedom of Expression," this is also an area fraught with controversy and reputational risks for companies, many of which are struggling with decision-making in this area given the issue's overlap with political viewpoints.

---

[31] "Notorious 8chan Forum is an Internet Nomad." *The Wall Street Journal.* November 16, 2019.

[32] Report of the Select Committee on Intelligence of the U.S. Senate. Accessed May 20, 2020.

[33] "Tackling Domestic Disinformation: What Social Media Companies Need to Do." *NYU Center for Business and Human Rights.* April 3, 2019.

[34] "The Global Disinformation Order: 2019 Global Inventory of Organised Social Manipulation." *Oxford University.* September 26, 2019.

One online community that has grown significantly since the advent of the internet is the "antivax" community, which spreads falsehoods related to the dangers of medical vaccines. Social media platforms have been blamed for declining vaccination rates leading to significant Measles outbreaks in the United States, Philippines, Ukraine, Venezuela, Brazil, Italy, France, and Japan in 2019; the disease was virtually extinguished just a few years ago.[35] In March of 2019, the head of the American Medical Association sent a letter to the CEOs of several prominent technology companies asking them to do more to prevent the spread of bad information leading to "easily preventable" diseases, stating that declining vaccination coverage "threaten[s] to erase many years of progress" in nearly eliminating several diseases.[36] At the beginning of 2019, the World Health Organization listed "vaccine hesitancy" among its 10 threats to global health for the year.[37]

Conspiracy theories are another form of misinformation that can lead into dangerous situations and real-world violence. In one now-infamous incident known as "pizzagate," a man opened fire in a Washington, DC, pizza restaurant after reading online that it was the center of a pedophile ring being operated by the highest levels of a major political party.[38] Another insidious conspiracy theory, led in part by a prolific online provocateur, has motivated a small group of people to persistently harass the parents of victims of the Sandy Hook Elementary School mass-shooting, alleging that the attack was fabricated, that the children never existed, and that the parents were paid "crisis actors."[39]

More generally speaking, there is increasing concern that social media platforms' role in facilitating a "marketplace of ideas" is breaking down as increasing amounts of bad information (in the form of misinformation and disinformation alike) drowns out the good, thereby reducing a collective sensemaking process and harming society in the aggregate.[40] [41] [42] [43] Part of this concern is driven by the belief in some quarters that social media platforms in particular are not just a conduit for disinformation or misinformation, but an amplifier.[44]

The coronavirus outbreak has sparked the most recent example of potentially dangerous misinformation being rapidly spread across the globe. In March of 2020, a group of some of the largest technology companies in the US released a joint press release stating that they would be "working closely together" to combat fraud and misinformation about the virus.[45]

---

[35] "Anti-Vaccine Decision-Making and Measles Resurgence in the United States." *Global Pediatric Health, U.S. National Library of Medicine, National Institute of Health*. July 24, 2019.
[36] AMA Press Release. March 14, 2019. Accessed May 21, 2020.
[37] "Vaccine Hesitancy, Climate Change, Ebola Among Top 10 'Threats to Global Health' this year, WHO Says." *ABC News.* January 17, 2019.
[38] "Pizzagate: Gunman Fires in Restaurant at Centre of Conspiracy." *BBC News.* December 3, 2016.
[39] "Battling Hoaxers in Court, Sandy Hook Families Replay a Tragedy." *The New York Times.* December 12, 2019.
[40] "Answering Impossible Questions: Content Governance in an Age of Disinformation." John Bowers and Jonathan Zittrain. *Harvard Kennedy School (HKS) Misinformation Review.* January 14, 2020.
[41] "How to Cope With an Infodemic." *Brookings Institute.* April 27, 2020.
[42] "Tackling Domestic Disinformation: What Social Media Companies Need to Do." *NYU Center for Business and Human Rights.* April 3, 2019.
[43] "Remediating Social Media: A Layer-Conscious Approach." Annemarie Bridy. *Boston University Journal of Science and Technology Law.* April 9, 2018.
[44] "It's Not Just the Content, it's the Business Model: Democracy's Online Speech Challenge." Nathalie Marechal & Ellery Roberts Biddle. *Ranking Digital Rights.* March 17, 2020.
[45] "Major Tech Platforms Say They're 'Jointly Combating Fraud and Misinformation' About COVID-19." *The Verge*. March 16, 2020.

Social media platforms that have advertising-based business models arguably should also monitor advertisements for misinformation. One platform's firm stance that it will not fact-check political ads, made on the basis that private companies shouldn't be an arbiter of political figures' speech, has attracted significant criticism from some politicians;[46] others have argued that the revenues generated from such ads aren't worth the reputational damage incurred.[47] A competing platform announced it would ban all "political ads" in response to the ongoing controversy over paid misinformation, although defining the term and enforcing the policy comes with its own challenges.[48]

Harmful Content in SASB's General Issue Categories[49]

Staff believes that companies' facilitation of harmful content falls in SASB's Social Capital sustainability dimension, in either the *Product Quality & Safety* or *Customer Welfare* general issue category. Determining which is most appropriate depends on the specifics of a disclosure topic related to harmful content.

If the disclosure topic, including its supporting evidence, frames harmful content as an unintended consequence akin to a manufacturing defect, then product quality & safety would be the appropriate category.

If a disclosure topic frames harmful content as a greater societal implication of a product or service that largely is operating as intended, then customer welfare would be a more appropriate category.

Please refer to the Appendix for full definitions of SASB's sustainability dimensions and general issue categories.

---

[46] "Elizabeth Warren Dares Facebook With Intentionally False Political Ad." *The New York Times.* October 12, 2019.

[47] "Facebook's Political Ad Business is Lots of Pain and Little Gain." *Bloomberg.* October 16, 2019.

[48] "It's Harder to Ban Political Ads on Twitter Than it Sounds." *The Verge.* October 31, 2019.

[49] SASB organizes the universe of sustainability risks and opportunities that companies can face into five broad sustainability dimensions: Environment, Social Capital, Human Capital, Business Model and Innovation, and Leadership and Governance. The five sustainability dimensions are further defined through a set of general issue categories associated with each sustainability dimension. Together, these sustainability dimensions and general issue categories serve as the high-level organizing structure for the disclosure topics covered in the SASB Standards. More specifically, all disclosure topics in the SASB Standards are classified under the most relevant general issue category (and sustainability dimension).

Freedom of Expression

The existence of harmful content on the internet leads to platforms removing content or limiting its visibility through different strategies and approaches. At social media platforms, these decisions—and the processes and procedures used to inform these decisions—are generally referred to as content moderation. A term that could be used to broaden these decisions to a larger universe of platforms and business activities is "content governance." Content governance decisions themselves have an associated second-order social externality: once platforms begin monitoring and removing content, their policies and procedures begin to interact with user rights such as privacy and freedom of expression.

From both a user-centric perspective and a business one, content governance decisions are foundational to the type of products or services offered. While at one point platform providers may have simply decided whether they were willing to facilitate the dissemination of pornography or violent material, technology companies now must grapple with far more complex decisions about how to provide effective products and services free of harmful content (and more mundane problematic content such as spam and malware) while balancing concerns surrounding freedom of expression and customer privacy. Several prominent platforms have taken firm stances on allowing for a wide range of political expression, while others offer more specialized experiences that more aggressively eschew any type of political controversy.[50]

Users also have increasingly high expectations regarding the processes and procedures deployed by companies regarding the content and activities they facilitate. The nature of this perceived responsibility varies not just based on the scale of the platform in question and the role of the platform in the broader internet ecosystem, but also on personal and cultural preferences.

There is also the challenge of scale: for companies that must make decisions on millions of pieces of content every day, even a 0.1 percent error rate will result in thousands of mistakes, some of which could result in reputational harm to the platform or even actual harm to users.

The following section is not meant to be an exhaustive body of evidence detailing the challenges that companies face when determining which content they are comfortable disseminating and moderating. Rather, this section uses a number of examples to demonstrate the complexity, novelty, and depth of these challenges.

Voluntary Content Moderation Pursuant to Platform Policy

The monitoring and removing of user-generated content at scale—a "tricky task"[51] or an "impossible job,"[52] depending on whom you ask—is often subjective and therefore subject to disagreement and conflict. Examples abound in news articles about social media platforms being criticized for failing to do more to combat such things as misinformation regarding the coronavirus pandemic, while simultaneously being criticized for censorship of users when they do act. The related questions quickly become philosophical and relate to a person's fundamental beliefs about the role that technology platforms—or private enterprises at large—should play in society.

---

[50] "The Tech Giant Fighting Anti-Vaxxers isn't Twitter or Facebook. It's Pinterest." *Fast Company.* February 26, 2019.
[51] "The Tricky Task of Policing YouTube." *The Economist.* May 4, 2019.
[52] "The Impossible Job: Inside Facebook's Struggle to Moderate 2 Billion People." *Vice.* August 23, 2018.

Complexities aside, it is not unreasonable for platforms to have differing approaches to how they balance freedom of expression with other rights that may be violated by the spread of harmful content, based on the nature of services they provide.

Even the removal of content that would seem relatively easy to define and police, such as violent content, can be challenging and have implications on freedom of expression. For example, after pressure from several NGOs, one social media platform determined that it should allow gruesome videos captured by cellphones in Syria to stay live on the site in order to document human rights abuses occurring during the civil war.[53]

Despite industrywide collaboration in the form of GIFCT and other initiatives, there remains widespread concern from industry and some policy bodies that over-policing platforms in the name of counter-terrorism initiatives could have negative side effects. For example, after the European Union released its Proposal for Regulation on Preventing Dissemination of Terrorist Content Online[54] in September of 2018, human rights experts from the United Nations raised a number of concerns with the proposal, including that the definition used for terrorist content "could encompass legitimate forms of expression, such as reporting by journalists and human rights organizations on the activities of terrorist groups and on counter-terrorism measures taken by authorities."[55] There are also concerns that organizations like GIFCT are not sufficiently transparent in making determinations around terrorist content, given their potential aggregate impact on freedom of expression: a shared database of banned TVEC from GIFCT could effectively remove flagged content from the internet entirely.[56]

Similar concerns—largely based around the idea that overly strict regulations will prompt internet platforms to over-moderate and remove legitimate information from the public domain—have been raised by many organizations, including digital rights advocates at the Electronic Frontier Foundation[57] and the Committee to Protect Journalists.[58] Another challenging aspect of the governance of TVEC is that governments have differing views on the definition of a terrorist organization; this leads to regimes using counter-terrorism laws to force platforms to remove legitimate criticism and other speech.

Many human rights experts argue that content governance decisions around freedom of expression should be grounded in international human rights law.[59] Overall, the case of online TVEC illustrates the lack of agreement among policymakers on how to approach moderating even clearly harmful content from a regulatory perspective, let alone a business one.

---

[53] "Social Media's Silent Filter." *The Atlantic.* March 8, 2017.

[54] EU Proposal for Regulation on Preventing Dissemination of Terrorist Content Online. Accessed May 26, 2020.

[55] "UN Human Rights Experts Concerned About EU's Online Counter-Terrorism Proposal." *United Nations Office of the High Commissioner: Human Rights Council.* December 12, 2018.

[56] "The Rise of Content Cartels." Evelyn Douek. *Knight First Amendment Institute at Columbia University.* February 11, 2020.

[57] "Caught in the Net: The Impact of 'Extremist' Speech Regulations on Human Rights Content." *Electronic Frontier Foundation.* May 30, 2019.

[58] "EU Online Terrorist Content Legislation Risks Undermining Press Freedom." *Committee to Protect Journalists.* March 11, 2020.

[59] "Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression." *United Nations Human Rights Council.* July 6, 2018.

In one example highlighting fears of how states could abuse content regulation laws, Singapore's inaugural enforcement of its first-of-its-kind "fake news" law—which allows ministers to order deliberate falsehoods to be removed from the internet—involved removal of posts containing criticism of the incumbent government.[60]

Hate speech, like terrorist content, is subject to varying definitions and an area where online content governance collides with the right to freedom of expression. It also is an area of increasing interest to regulators globally. For example, a German satirical magazine's cartoon that mocked a far-right politician's anti-Muslim stances was blocked by a major social media platform in 2018 following the passage of a law that placed significant restrictions on online hate speech.[61] The controversy is one of countless examples of how nuance and cultural context make content moderation—and global regulatory compliance in a rapidly evolving environment—a major test for social media platforms in particular. As with TVEC, there are growing calls from human rights experts to align social media speech codes with existing international human rights principles.[62]

Given the nuance of many content decisions and the fact that companies are highly unlikely to perfectly conform with their own content policies, most social media platforms allow for some type of appeal. In a particularly striking example of the novelty of content governance mechanisms, one large social media platform has formed an independent oversight body to handle a small fraction of content appeals and enlisted a nonprofit organization to conduct a human rights review of the process.[63]

Social media platforms also have a unique set of content governance concerns related to their ability to amplify certain types of harmful content, especially disinformation and misinformation. Amplification can take different forms depending on the platform in question, and includes algorithmic sorting of user posts, ranking of search results, and recommended groups or pieces of content. There is only limited empirical research that has been conducted in this area to date; for example, two studies regarding whether a popular video sharing platform "radicalizes" users through its recommendation algorithms have yielded varying results.[64] [65] The complexity of measuring the specific harms in this area—things such as radicalization, polarization, or the deterioration of the information environment—make it a particularly challenging area for platforms, governments, academics, and users to gather evidence or make decisions.

Platforms with ad-based models also have a financial incentive to increase the level of user "engagement"—interactions with content, shares of content, and generally spending more time actively involved with their products. In a 2018 blog post, the CEO of a major social media platform noted that "one of the biggest issues social media networks face is that, when left unchecked, people will engage disproportionately with more sensationalist content …. At scale it can undermine the quality of public discourse and lead to polarization. In our case, it can also

---

[60] "Singapore Just Used its Fake News Law. Critics Say It's Just What They Feared." *CNN.* November 30, 2019.
[61] "German Hate Speech Law Tested as Twitter Blocks Satire Account." *Reuters.* January 3, 2018.
[62] "The Future of Freedom of Expression Online." Evelyn Mary Aswad. *Duke Law & Technology Review.* December 8, 2018.
[63] "A Human Rights Review of the Facebook Oversight Board." *BSR.* December 12, 2019.
[64] "A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos." Hany Farid. School of Information, UC Berkeley. March 6, 2020.
[65] "A Supply and Demand Framework for YouTube Politics." Kevin Munger and Joseph Phillips. Penn State Political Science. October 1, 2019.

degrade the quality of our services."[66] The CEO specifically called out misinformation and "clickbait" as problematic, "borderline" content in this context, and provided a visualization of this challenge:

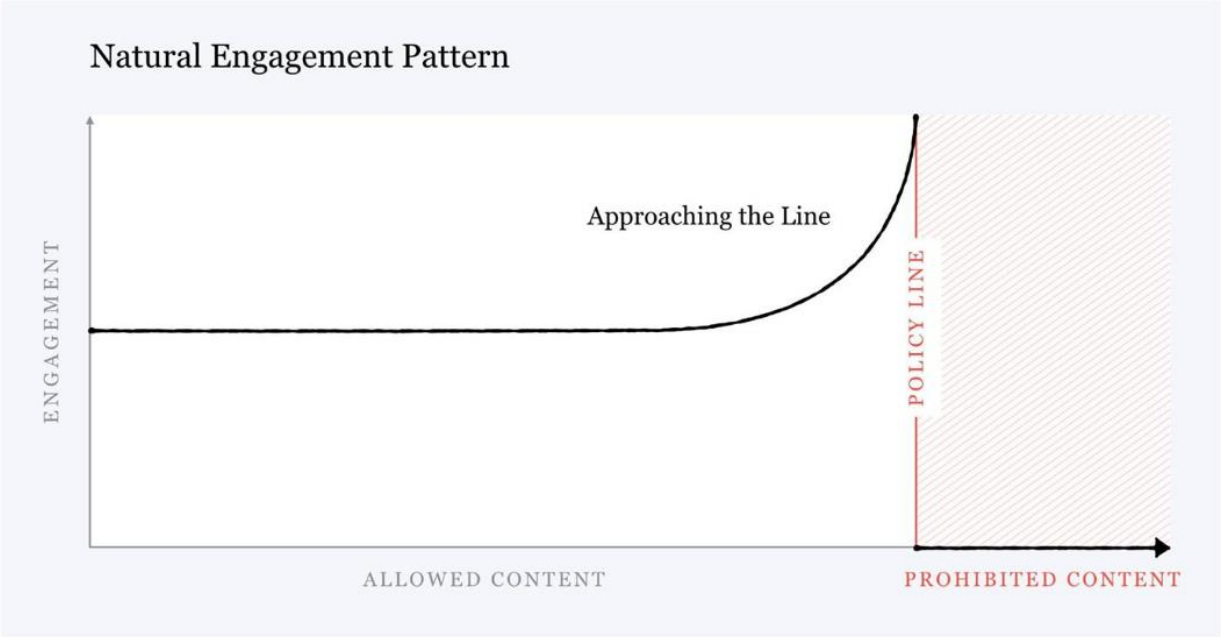

Figure 1.1: A social media platform's interpretation of the "Natural Engagement Pattern" on its network.[67]

While the topic of freedom of expression may be most prominent at social media platforms, there are also examples of other types of companies that are dealing with the same issues. For example, the controversy around the de-platforming of neo-Nazi websites like *The Daily Stormer* (discussed above in the "Harmful Content" section) is an example of cloud service providers, domain registration services, and payments processors all making a decision to stop providing services to customers based on their hosting of objectionable/harmful content.

One cloud service provider told SASB staff that it manages the issue of freedom of expression largely through vetting the customers it chooses to do business with. This company takes a relatively strict approach to the content it enables; for example, it does not partner with websites that host pornography. A different cloud service provider described itself as "incredibly uncomfortable" in deciding to halt service to the notorious forum 8chan given its philosophy that network providers shouldn't take on a content arbitration role, but did so after the users of the forum were found to be actively encouraging recent mass shootings in the US—specifically in the cities of El Paso, Texas, and Dayton, Ohio.[68]

---

[66] "A Blueprint for Content Governance and Enforcement." *Facebook Press Release.* November 15, 2018.
[67] Ibid.
[68] "Terminating Service for 8chan." *Cloudflare Press Release.* August 4, 2019.

Content Removal Requested by Governments

In addition to voluntary content moderation decisions, platforms also routinely receive requests from governments to remove content deemed to be illegal. This is an area with distinct challenges and risks for companies, and one with more developed disclosure and governance norms; for example, the current SASB standard for the Internet Media & Services industry includes a disclosure topic and relevant metrics pertaining to this issue.[69]

In countries with stricter regulations on permitted speech, nuances around freedom of expression are replaced with a far greater burden on companies to remove user-generated content deemed sensitive or inappropriate. In April of 2020, the Cyberspace Administration of China ordered a social media platform to clean up "low brow content" and criticized it for not having a strict enough content policy.[70] Another major platform based in the region notes in a regulatory filing that it may be found liable for content that is "alleged to be socially destabilizing, obscene, defamatory, libelous or otherwise unlawful" and that "it may be difficult to determine the type of content that may result in liability to us."[71] The challenge of avoiding liability for so much user-generated content has led platforms to take a variety of approaches; a recent research report from the University of Toronto found that a ubiquitous messaging app was monitoring the communications of foreign users in order to more quickly remove sensitive content behind China's "great firewall."[72]

Another recent incident involved the gaming industry, when a virtual card game player was banned by the game publisher after supporting ongoing protests in Hong Kong during a competition livestream.[73] The game publisher cited its policies prohibiting political speech, but this did little to quell criticism from customers, activists, politicians, and the media at large while setting off a broader international furor.

Freedom of Expression in SASB's General Issue Categories

Platforms' interactions with user freedom of expression as described in this section could be categorized in SASB's Social Capital or Leadership & Governance sustainability dimensions, likely in the *Customer Welfare* or *Business Ethics* general issue categories. Determining which is most appropriate depends on the specifics of a disclosure topic related to freedom of expression.

If a disclosure topic frames the relevant issues as platforms' restriction of users' freedom of expression, customer welfare would be an appropriate category.

A broader disclosure topic—which frames content removal and user freedom of expression as one piece of companies balancing their own economic interests with those of users, customers, and governments—could appropriately fit into the business ethics category.

---

[69] The disclosure topic is titled "Data Privacy, Advertising Standards & Freedom of Expression". It addresses and measures a variety of issues, including risks associated with providing access to user data to governments and government demands related to censorship of culturally or politically sensitive material.

[70] "China Orders Baidu to Clean Up Low-Brow Content." *CNBC.* April 8, 2020.

[71] Alibaba Group Holding Ltd. Form 20-F, filed June 5, 2019.

[72] "China's WeChat Monitors Foreign Users to Refine Censorship at Home." *The Wall Street Journal.* May 8, 2020.

[73] "Blizzard Bans Player for Supporting Hong Kong Protests." *The Verge.* October 8, 2019.

Please refer to the Appendix for full definitions of SASB's sustainability dimensions and general issue categories.

User Privacy

Privacy and Searching for Harmful Content

Another area where the policing of harmful content has its own social externalities is that of digital privacy. While examining the trade-offs between security and privacy is not a new subject of discourse in the public policy arena, the continued evolution of technology presents new challenges for users, companies, and policymakers.

For example, the ability for companies to counter the online spread of child sexual abuse material (CSAM) conflicts with a trend toward end-to-end encryption of messaging services. One investigation from *The New York Times* found that nearly 90 percent of reported CSAM to the NCMEC came from the world's largest social media/messaging provider; the natural concern of law enforcement agencies and advocates at the NCMEC is that a full-fledged pivot to encryption will prevent the use of digital fingerprinting tools and impede the reporting of illegal and harmful content.[74] The demand for encrypted messaging services from users is understandable, especially from those skeptical of their governments' surveillance efforts; as such, some argue that there may be an inherent trade-off between true privacy and improving online safety.

The concern that overly stringent regulations on CSAM, TVEC, and hate speech will have negative impacts on freedom of expression also apply to elements of data privacy. For example, a piece of legislation being considered by the US Senate dubbed the "EARN IT Act"—which would strip online platforms of their liability protections if they fail to do more to counter CSAM—has been criticized by some as a back-door attack on encryption by the Department of Justice.[75] [76]

SASB staff's review of sales materials indicates that companies that provide storage and computing services often make the strength of their data security processes a key selling point. The world's largest cloud provider cited user privacy in its response to reporters asking why it did not scan for CSAM on its network.[77] In 2015, a number of child pornography cases appeared to have been reported a major cloud storage provider; the company declined to comment when asked by a reporter about its practices.[78]

As noted above, public cloud providers' treatment of illegal content such as CSAM has increasingly come under scrutiny. However, there is a growing understanding that the

---

[74] "Facebook Encryption Eyed in Fight Against Online Child Abuse." *The New York Times.* October 2, 2019.
[75] "The EARN IT Act is a Sneak Attack on Encryption." *Wired.* March 5, 2020.
[76] "The EARN IT Act is a Disaster Amid the COVID-19 Crisis." *Brookings Institute.* May 4, 2020.
[77] "Child Abusers Run Rampant as Tech Companies Look the Other Way." *The New York Times.* November 9, 2019.
[78] "Dropbox Refuses to Explain its Mysterious Child Porn Detection Software." *Gizmodo.* August 12, 2015.

techniques cloud service providers and content delivery network providers use to prevent cyberattacks and other activities—such as pattern recognition and metadata analysis—could be applied to harmful content such as CSAM. For example, in May of 2020, a prominent social media platform and messaging application announced a series of new tools to alert users of messages that may be coming from suspicious accounts and bad actors; the tools use machine learning to analyze metadata and therefore do not require the breaking of encryption.[79]

Privacy and Content Governance

There are also privacy considerations for platforms that collect and use sensitive user data to personalize the content viewed by users. Privacy concerns with behavioral (often called "targeted") advertisements viewed on internet platforms are well documented, including in SASB's Internet Media & Services Industry Standard.

A new set of concerns has arisen around personalized content delivery encapsulated by the idea of the "filter bubble," a term coined by Eli Pariser in 2011. This term captures the concern that, in their quest to determine what users are most likely to engage with, platforms may deliver search results, recommendations, and sorted content that confirm or exacerbate existing biases.[80] While still an emerging area of research and inquiry, the potential social ramifications are diverse and could be significant.

User Privacy in SASB's General Issue Categories

User privacy can be directly categorized under SASB's Social Capital sustainability dimension, and the Customer Privacy general issue category. This is consistent with the relevant disclosure topic in SASB's existing Internet Media & Services Standard.

---

[79] "Facebook Messenger Adds Safety Alerts—Even in Encrypted Chats." *Wired.* May 21, 2020.

[80] "It's Not Just the Content, it's the Business Model: Democracy's Online Speech Challenge." Nathalie Marechal & Ellery Roberts Biddle. *Ranking Digital Rights.* March 17, 2020.

Worker Health & Safety

Another implication of harmful content being shared across the internet is that some of this content—which includes deeply disturbing content such as graphic violence and CSAM—must be reviewed by humans prior to removal. Advances in technology such as machine learning and hashing are helping to automatically remove harmful content, but a significant portion of this content must be processed by "human-in-the-loop" systems, where clips or posts are flagged by users or automated systems for review. The largest social media platforms have tens of thousands of workers (in-house, short-term contractors and contracted through vendor partners) performing content moderation work, although exact breakdowns or figures for these workers are not made public.

As discussed in greater detail below, there is a growing body of evidence that repeated exposure to harmful content may have significant and negative impacts on mental health, including the development of PTSD. As such, companies operating in this space may be required to dedicate additional resources toward protecting the employees performing this important work.

In recent years, there has been a growing focus in the media on what *The New Yorker* in 2017 called "the human toll of protecting the internet from the worst of humanity."[81] The earliest news report detailing "commercial content moderation" by workers that SASB staff identified was published by *The New York Times* in 2010.[82] In 2014, Dr. Sarah T. Roberts published a dissertation titled *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation* (subsequently adapted into a book) that significantly expanded public visibility into the commercial content moderation profession through research and a series of in-depth interviews with workers.[83]

In February 2019, reporting from *The Verge* detailed the grizzly images and videos that content moderators were reviewing for a living, the mental toll described by some of these workers, the challenge of meeting accuracy targets as measured against a complex and constantly-changing set of moderation guidelines, and unsanitary working conditions at a moderation facility.[84] The facility in question was located in Phoenix, Arizona, and run by a third-party vendor hired by a major social media platform. Other reporting has detailed the mental health struggles of moderators for various large social media platforms in the US,[85] Ireland,[86] the Philippines,[87] and India.[88]

Much of the recent media reporting on this "secondary trauma" is being driven by lawsuits brought by current and former moderators against their employers. These lawsuits have been brought in multiple jurisdictions, including the US and Ireland. In January of 2020, the *Financial*

---

[81] "The Human Toll of Protecting the Internet from the Worst of Humanity." *The New Yorker.* January 28, 2017.
[82] "Policing the Web's Lurid Precincts." *The New York Times.* July 18, 2010.
[83] "Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation." Sarah T. Roberts. Dissertation at the University of Illinois – Champaign. September 16, 2014.
[84] "The Trauma Floor: The Secret Lives of Facebook Moderators in America." *The Verge.* February 25, 2019.
[85] "The Terror Queue." *The Verge.* December 16, 2019.
[86] "Bestiality, Stabbings and Child Porn: Why Facebook Moderators Are Suing the Company for Trauma." *Vice.* December 3, 2019.
[87] "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014.
[88] "Inside the Secretive World of India's Social Media Content Moderators." *LiveMint.* March 18, 2020.

Times reported that moderators at a European site working through a third-party vendor were being required to sign a waiver explicitly acknowledging that their job could cause PTSD.[89] The article stated that the social media platform in question told reporters that it had not reviewed or approved the form distributed by the vendor and that it requires its partners "to offer extensive psychological support to its moderators on an ongoing basis."

There is not, to staff's knowledge, an established body of research regarding the mental health impacts of viewing a variety of harmful content, nor do there appear to be clear industry-wide best practices that have been developed. A cross-sector group geared toward combating CSAM online called the Technology Coalition released an Employee Resilience Handbook in January 2015, but it does not appear to have been updated since.[90] One potential parallel that could be explored is that of law enforcement officers that work in the area of CSAM: a 2013 article in a medical journal found that the use of coping mechanisms was inversely related with the development of secondary traumatic stress in personnel at the Internet Crimes Against Children task force.[91]

The relative novelty of this human capital management issue means companies are continuing to develop their own practices. For example, the CEO of one large social media platform stated in 2018 that the company would be limiting the exposure of part-time moderators to graphic content to no more than four hours per day, alongside other actions to improve employee wellness.[92] The articles cited above provide detail on some other ways in which platforms are attempting to protect moderators, including through making images and videos black and white, and disabling audio by default.

The continued development and use of machine learning and other technologies can also help identify and remove harmful content before it is viewed by humans, strengthening the case for companies to invest heavily in automated moderation. However, subject matter experts in this area told SASB staff that humans still review a vast majority of content and that reliable "tech-first" solutions are many years away given the nuance of many moderation decisions and the vast geographic scope of many social media platforms.

Much of the moderation work performed for the largest social media platforms is contracted through third-party business process outsourcing firms, generally referred to in the industry as vendors. While the exact breakdown of contracted vs. in-house workers is not disclosed by these companies, one large social media platform told SASB staff that a majority of the work on the "front lines" was done through vendor partners across the globe.

These vendors include some of the largest business process outsourcing firms, whose global footprint, experience running large-scale operations such as call centers, and ability to recruit workers with specialized language skills make them an obvious partner for social media platforms with hundreds of millions of daily active users. There are also a variety of services being made available for a broader set of use cases around brand safety concerns related to

---

[89] "Facebook Content Moderators Required to Sign PTSD Forms." *The Financial Times.* January 26, 2020.

[90] Technology Coalition Website; Handbook Link found via Google search (accessed May 29, 2020).

[91] "Secondary Traumatic Stress Among Internet Crimes Against Children Task Force Personnel: Impact, Risk Factors, and Coping Strategies." Michael L. Bourke and Sarah Craun. *Sexual Abuse Journal of Research and Treatment.* November 20, 2013.

[92] "YouTube Limits Moderators to Viewing Four Hours of Disturbing Content Per Day." *The Verge*. March 13, 2018.

user-generated content, such as boutique services that use AI and other techniques to scrub the comments sections of websites.

Given the amount of activities that continue to migrate online, content moderation services appear poised for continued growth. In March of 2020, *The Wall Street Journal* outlined services being provided by large China-based technology firms that can help smaller companies comply with the country's restrictions on permitted speech; one government official and board member at an online publication recently predicted that content moderation in China will soon grow into a $70 billion industry that will employ over 1 million people.[93]

This relationship with vendor partners means that social media platforms must manage this issue both from a human capital perspective and a vendor/supply chain management one. According to several of the articles cited above, the mental health impacts on workers were driven not just by the harmful content viewed, but by poor working conditions provided by vendors.

During staff's consultations on this topic, two subject matter experts stated that one challenge of content moderation broadly is the fact that the job function, which is often referred to as "Trust & Safety" at companies in the sector, is a relatively new profession lacking clearly defined career paths or best practices for maintaining mental health. Further, resource allocation to the enforcement of content rules is lacking in benchmarks. One expert told staff that they viewed the profession similarly to that of data security professionals 5 or 10 years ago in that this work is more likely to be viewed by executives—especially those at companies where harmful content is less of a direct threat to customer experience—as a cost center rather than an investment.

One social media platform provided staff with a different perspective, stating that many of the experts specialized in issues like TVEC and CSAM and had joined the company with a sense of purpose in hoping to prevent the spread of harmful content and identify bad actors. They also noted that executives, up to and including the CEO, have invested time in reviewing harmful content and understand the importance of the work being done by moderators.

Worker Health & Safety in SASB's General Issue Categories

Worker Health & Safety likely falls under the *Employee Health & Safety* General Issue Category in SASB's Human Capital sustainability dimension. This category captures "how companies ensure physical and mental health of workforce through technology, training, corporate culture, regulatory compliance, monitoring and testing, and personal protective equipment."

To the extent that a disclosure topic frames this issue as platforms managing their content moderation workforce through third parties, an appropriate GIC could also be *Supply Chain Management* under SASB's Business Model & Innovation sustainability dimension.

[93] "Made-in-China Censorship for Sale." *The Wall Street Journal.* March 6, 2020.

Review of Business Activities

This section describes the business activities that are most likely to be relevant to the content moderation issues identified in this document (harmful content, freedom of expression, user privacy, and worker health & safety). While SASB develops industry-specific standards using its Sustainable Industry Classification System® (SICS®), defining distinct sets of business activities within these industries is helpful for understanding the different ways in which these social externalities may manifest.

The SICS® industries of Internet Media & Services, Software & IT Services, Telecommunication Services, and E-Commerce have been further expanded to include business activities such as Social Media Platforms and Internet Infrastructure & Cloud Services.

SASB SICS® Industry: Internet Media & Services

Social Media Platforms

Social media platforms facilitate the creation, hosting, and dissemination of a variety of content, including user-generated content. These platforms are exposed to all four issues outlined in this Taxonomy. The extent to which each issue applies appears to be related to the platform's scale. Platforms with large user bases may significantly amplify harmful content and therefore be expected to have more effective content governance policies; a high volume of harmful content may also expose content reviewers to heightened mental health hazards.

Moderation can take a number of different forms, including removal, reductions in visibility (such as hiding objectionable posts from feeds, or hiding groups in search results), labeling of posts,[94] and suggestions to review information from reputable sources.[95]

Social platforms that rely on advertising revenue must also conduct content moderation in order to assuage their customers' brand safety concerns; this issue was brought into focus during a recent boycott of social media by hundreds of advertisers.[96]

Messaging Applications

Messaging applications facilitate one-on-one or small group communication and can be stand-alone services or embedded inside other products, such as social media or gaming platforms. These applications can enable sharing of harmful content. Any content governance mechanisms in this area may stand in opposition to user expectations around security and privacy, as discussed in greater detail in "User Privacy."

---

[94] "Twitter Official Explains Blue Exclamation Marks on Potentially Misleading Content." *NPR.* March 3, 2020.

[95] "Facebook Has Announced a Multistep Plan to Crack Down on Anti-Vax Information." *Buzzfeed News.* March 7, 2019.

[96] "Facebook Ad Boycott: Why Big Brands 'Hit Pause on Hate.'" *cnet.* July 30, 2020.

Other Internet Platforms

Several prominent technology platforms have been successful in building global brands in areas like ride-hailing, dating, and vacation rentals. Platforms that facilitate the physical interaction between humans may also be expected to have processes in place to protect the safety of users. These platforms, alongside social media platforms, are the subject of a debate over the extent to which platforms should be responsible (and legally liable) for the behavior of users. Dating services[97] and vacation rental services[98] both have faced scrutiny for their management of trust and safety concerns in recent times.

SASB SICS® Industry: Software & IT Services

Infrastructure and Cloud Services

Businesses that provide a range of services that facilitate the hosting and delivery of content on the internet are defined here as infrastructure and cloud service providers. These services include "public cloud" providers of online storage, content delivery networks, and domain registration service providers. Traditionally, these companies have been viewed as neutral intermediaries and therefore not held to the same expectation of content governance as social media platforms. However, there are growing demands that these companies do more to combat harmful content, such as child exploitation and hate speech (as discussed briefly with regard to neo-Nazi sites such as *The Daily Stormer* under "Harmful Content").

While the Terms of Service of these platforms clearly prohibit use of their services for illegal activities such as the sharing of CSAM, the enforcement of these terms is not an area where there is widespread disclosure or discussion. A review of regulatory filings indicates that the companies providing domain registration and other related services do not actively monitor their customers.

Game Publishers and Platforms

Game publishers and platforms develop and distribute interactive video games that can be accessed either through hardware, software, or a combination of both. These service providers have come into the public conversation surrounding harmful content as video games migrate online and increasingly facilitate user interaction. Two companies with businesses in this space told SASB staff that the nature of user interaction is highly dependent on the game in question, and that user interaction may occur on messaging platforms not controlled by the game publishers or platforms themselves. This complicates efforts to enforce Terms of Service, such as those that prohibit abuse and harassment.

The controversy surrounding a live eSports event discussed under "Freedom of Expression" also details the ways in which gaming companies may begin to have business lines that

---

[97] "Tinder Lets Known Sex Offenders Use the App. It's Not Alone." *ProPublica*. December 2, 2019.

[98] "Airbnb to Verify All Listing, CEO Chesky Says." *The New York Times.* November 6, 2019.

resemble traditional media, which may require additional controls and procedures in order to mitigate related risks.

Business Process Outsourcing

A number of firms perform contracted content moderation services to large social media platforms. As detailed in "Worker Health and Safety," content moderation services may be a significant source of revenue for some companies; however the providers of these services may also be exposed to human capital management themes pertaining to employee health and safety to the extent their employees are viewing large volumes of harmful content.

SASB SICS® Industry: E-Commerce

Online Marketplaces

Online marketplaces are "two-sided" platforms that link buyers with third-party sellers. The practices and procedures used to enforce Terms of Service in this environment and ensure product safety—which generally revolve around verifying and vetting platform participants such as third-party sellers—appear to be distinct from the type of content governance activities undertaken by social media or gaming platforms.

However, depending on the nature of a given marketplace, the goods being sold may be forms of user-generated content that pose risks to platform providers. For example, the world's largest e-book marketplace allows individual authors to self-publish their work; a recent investigation found that this marketplace contained neo-Nazi publications and other objectionable hate speech.[99]

SASB SICS® Industry: Telecommunication Services

Internet Service Providers

Internet Service Providers sit a layer below infrastructure and cloud service providers in the overall schematic of internet delivery in that they transmit bits of data from point A to point B without storing or hosting this data. As discussed under "Harmful Content," ISPs already are required to report illegal activities to law enforcement as they become aware of them; they are also facing growing calls to more actively monitor and remove harmful content that travels across their networks.

[99] "The Hate Store: Amazon's Self-Publishing Arm is a Haven for White Supremacists." *ProPublica.* April 7, 2020.

Conclusion

This document has been prepared with the objective of defining four issues central to content moderation—harmful content, freedom of expression, privacy, and worker health and safety—and plotting them out across seven business activities in the technology sector, in order to have a clear framework for evaluating whether and how to approach standard-setting.

Overall, the most significant volume of evidence reviewed during this research project relates to how social media platforms are addressing the challenges (and opportunities) posed by each of the four issues outlined in this document. This is likely because of the number of social media platforms whose businesses rely on the monetization of user-generated content. Pressure from users, advertisers, and regulators around the world to implement coherent, fair, and scalable content policies is only likely to increase over time.

The evidence cited here also suggests that a variety of businesses beyond the social media platforms conduct forms of content moderation, although their policies and decisions on content may not be as central to their management strategies and future value creation. Nonetheless, as awareness of unintended consequences or misuse of technology products and services spreads, all companies in the sector would likely be well-served to consider the broader ramifications of their products, alongside the attendant business risks and opportunities.

SASB encourages interested stakeholders to engage with the technical staff to help build our understanding of these social issues, the associated business activities, and the cited body of evidence. Those interested in following our continued research and standard-setting in this area are encouraged to do so through the project pages[100] found on our website and through listening in to future Standards Board meetings.

[100] SASB project pages

Appendix I: Issues Not Included in Taxonomy

As part of the research performed in preparing this document, staff reviewed a range of content moderation activities and policies. Staff believes that two important content moderation areas are more straightforward business issues that do not have a clear associated social externality and therefore are not considered "harmful content" under this Taxonomy. These two types of content are:

- Spam and other potentially fraudulent content, including fake accounts (excluding those set up as part of a disinformation campaign)
- Content that infringes the intellectual property of copyright owners

Appendix II: SASB Sustainability Dimensions and General Issue Categories

Social Capital

This dimension addresses a company's impact on external stakeholders and the management of those stakeholder relationships, including a company's license to operate. External stakeholders include customers, local communities, regulators, and the public. Impacts on these stakeholders may relate to issues such as human rights, protection of vulnerable groups, local economic development, access to and quality of products and services, affordability, responsible business practices in marketing, and customer privacy. Stakeholders that are directly or indirectly employed by the company are excluded from this sustainability dimension and are instead captured under the "Human Capital" and/or "Business Model & Innovation" dimensions.

Customer Privacy

The category addresses management of risks related to the use of personally identifiable information (PII) and other customer or user data for secondary purposes including but not limited to marketing through affiliates and non-affiliates. The scope of the category includes social issues that may arise from a company's approach to collecting data, obtaining consent (e.g., opt-in policies), managing user and customer expectations regarding how their data is used, and managing evolving regulation. It excludes social issues arising from cybersecurity risks, which are covered in a separate category.

Product Quality & Safety

The category addresses issues involving unintended characteristics of products sold or services provided that may create health or safety risks to end-users. It addresses a company's ability to offer manufactured products and/or services that meet customer expectations with respect to their health and safety characteristics. It includes, but is not limited to, issues involving liability, management of recalls and market withdrawals, product testing, and chemicals/content/ingredient management in products.

Customer Welfare

The category addresses customer welfare concerns over issues including, but not limited to, health and nutrition of foods and beverages, antibiotic use in animal production, and management of controlled substances. The category addresses the company's ability to provide consumers with manufactured products and services that are aligned with societal expectations. It does not include issues directly related to quality and safety malfunctions of manufactured products and services, but instead addresses qualities inherent to the design and delivery of products and services where customer welfare may be in question. The scope of the category also captures companies' ability to prevent counterfeit product.

Leadership & Governance

This dimension involves the governance and management of key industry issues that may create conflicts with the interests of broader stakeholder groups, and therefore may lead to liabilities or impacts on a license to operate. The dimension includes conducting business activities in compliance with industry laws and regulations, and in accordance with the industry's leading standards of professional integrity. Issues captured in this dimension include, anticompetitive practices, ethical conduct of business, and engagement with regulators on environmental, social, and human impacts. This dimension also addresses the management of risks related to low-probability, high-impact accidents and emergencies that generate a multitude of sustainability impacts.

Business Ethics

The category addresses the company's approach to managing risks and opportunities surrounding ethical conduct of business, including fraud, corruption, bribery and facilitation payments, fiduciary responsibilities, and other behavior that may have an ethical component. This includes sensitivity to business norms and standards as they shift over time, jurisdiction, and culture. It addresses the company's ability to provide services that satisfy the highest professional and ethical standards of the industry, which means to avoid conflicts of interest, misrepresentation, bias, and negligence through training employees adequately and implementing policies and procedures to ensure employees provide services free from bias and error