

Online Safety Regulations Around The World:

The State of Play and The Way Forward - A Resource Guide

MARIANA OLAIZOLA ROSENBLAT, AYUSHI AGRAWAL & ISAAC YAP



 **NYU | STERN**

Center for Business
and Human Rights

April 2025

Contents

Executive Summary	1
1. Introduction	3
2. Lay of the Land	5
Content-based requirements	5
Design requirements.....	8
Transparency requirements.....	11
Procedural requirements.....	13
3. The Road Ahead	16
4. Appendix.....	19
Endnotes.....	24

Acknowledgments

We are grateful to the Open Society Foundations for their continued support of our work on technology and democracy.

Authors

Mariana Olaizola Rosenblat is a policy advisor on technology and law at the NYU Stern Center for Business and Human Rights.

Ayushi Agrawal and Isaac Yap, LL.M. students at the NYU School of Law, provided valuable research and analysis.

Paul M. Barrett, the NYU Stern Center's deputy director, provided assistance with the editing of this report.

Executive Summary

“
Drawing on a selection of 26 online platform regulations in 19 jurisdictions, this study offers a framework for understanding the main regulatory approaches adopted so far.

”

Governments worldwide are increasingly regulating online platforms with a view to addressing a variety of risks and harms associated with digital communications tools.

In the EU, the UK, Australia, Singapore, and many other places, there is considerable diversity in regulatory approaches. This study reviews and compares the main pieces of legislation enacted, proposed and, in a handful of cases, enforced.

Drawing on a selection of 26 online platform regulations in 19 jurisdictions, the study offers a framework for understanding the main regulatory approaches adopted so far, which include:

Content-based requirements under which online platforms are compelled to take action on certain classes of content;

Design requirements under which platforms must implement or refrain from implementing certain design features, such as push notifications and geolocation tracking;

Transparency requirements including a variety of obligations to disclose information or share data with external entities; and

Procedural requirements such as obligations to provide users with ways to report violations of terms of service.

In addition to providing a system for classifying current regulatory efforts, the study discusses the advantages and disadvantages of each approach. The third part of the study aims to chart a path forward for future regulation. On the following page is a summary of our recommendations to regulators.

Recommendations In Brief

- 1 Ensure that requirements for platforms to remove or otherwise limit certain classes of content pertain only to explicitly illegal content.** Governments should not require platforms to remove content that could be harmful but is not illegal, unless the harmful content is defined precisely enough to meet the “[legality](#)” standard under international human rights law.
- 2 Compel platforms to disclose information about their business operations which affects consumers and society, and subject those disclosures to external audit and analysis by vetted researchers.** These disclosures, audits, and data access regimes for researchers should be accompanied by robust safeguards to protect user privacy and legitimate trade secrets.
- 3 Ensure that platforms allow users to customize key design features affecting their online experience, such as safety and privacy settings and algorithmic recommendation systems.** Any highly prescriptive design-based mandates should be based on empirical research and proportional to the regulation’s aims.
- 4 Give teeth to requirements that platforms fulfill the promises they make to users.** In setting out any procedural requirements, regulators should establish clear standards to inform platforms’ compliance efforts.
- 5 Delegate enforcement to an independent agency with limits on its authority.** Regulators should ensure that this agency is appropriately funded and staffed with expert personnel.
- 6 Tailor some requirements to platforms of different types, sizes and risk profiles.** Regulators should establish a broad scope in the regulation’s coverage, but differentiate among platforms based on their service (e.g., live-streaming versus e-commerce), features, and reach.
- 7 Work together with privacy and antitrust counterparts to ensure requirements are compatible.** Regulators should be aware of certain tradeoffs and avoid establishing conflicting obligations for platforms.
- 8 Participate in multilateral initiatives to enhance global regulatory coherence.** Regulators should consider participating in initiatives like the [Global Online Safety Regulators Network](#) (GOSRN) to share best practices, tools, and experiences.

1. Introduction

“
This study provides
a framework for
understanding existing
regulations, evaluating
regulatory efforts under-
way, and informing future
regulatory initiatives.
”

At the beginning of 2025, global internet regulation stands at a critical juncture.

The past two years witnessed an unprecedented wave of legislative action across major economies, with the European Union’s [Digital Services Act](#) (DSA) and [Digital Markets Act](#) (DMA) now in full effect, promising to reshape how tech giants operate in Europe. The United Kingdom, Ireland, and several jurisdictions across Asia and Oceania also passed muscular platform legislation that they are now poised to enforce.

Meanwhile, the United States has failed for the past two decades to pass significant federal digital regulation.¹ But online safety legislation has been enacted at the state level in the US, and federal regulation on discrete issues such as children’s online safety remains a possibility.

Against this complicated backdrop, the present study provides a framework for understanding existing regulations, evaluating regulatory efforts underway, and informing future regulatory initiatives.


This study focuses on regulation that addresses online safety. This includes legally binding instruments (laws, rules, codes of practice, etc.) that impose direct obligations on online services to prevent and address harms ranging from cyber-harassment to compulsive usage. Although some of these instruments also touch on data privacy and cybersecurity, those areas are not the focus of this analysis.²

While the study adopts a global perspective, it is intentionally selective. Regulations were deemed in scope if they met a threshold of legitimacy by virtue of being enacted within a constitutional democracy, under the assumption that measures promulgated by authoritarian regimes are less likely to serve as suitable models. The final selection consists of 26 distinct online safety regulations across 19 jurisdictions as listed on the following page:

Selection of 26 distinct online safety regulations across 19 jurisdictions

European Union		Digital Services Act (2022)
		Regulation to address the dissemination of terrorist content online (2021)
Ireland		Online Safety and Media Regulation Act (2022) and Online Safety Code (2024)
United Kingdom		Online Safety Act (2023)
		Age Appropriate Design Code (Children's Code) (2020)

Australia		Online Safety Act (2021) , with Social Media Minimum Age Amendment (2024) , including the Basic Online Safety Expectations (2022) and Industry Codes and Standards (ongoing)
		Sharing of Abhorrent Violent Material Act (2019)
Fiji		Online Safety Act (2018) and Online Safety Regulations (2019)
India		Information Technology Rules (2021)
New Zealand		Harmful Digital Communications Act (2015)
Singapore		Online Safety (Miscellaneous Amendments) Act (2022)
South Korea		Act on Promotion of Information and Communications Network Utilization and Information Protection (2016)
		Telecommunications Business Act (2023)

South Africa		Film and Publications Amendment Regulations (2022)
--------------	---	--

North and South America			
Brazil		Marco Civil of the Internet (2014)	
United States of America		California	
		Age-Appropriate Design Code Act (2021-2022)	
		Social media companies: terms of service (2022)	
		Protecting Our Kids from Social Media Addiction Act (Addiction Act) (2024)	
		Colorado	House Bill 24-1136 (2024)
		Florida	SB 7072: Social Media Platforms (2021)
		Louisiana	Secure Online Child Interaction and Age Limitation (SOCIAL) Act (2023)
		Maryland	Age-Appropriate Design Code Act (2024)
		New York	Stop Addictive Feeds Exploitation (SAFE) for Kids Act (2023-2024)
		Texas	Securing Children Online Through Parental Empowerment (SCOPE) Act (2024)
	HB 20 (2021)		
Utah	Minor Protection in Social Media (MPSM) Act (2024)		

2. Lay of the Land

“
The content-based approach generally involves establishing classes of prohibited content which online services are required to remove.
”

The duties outlined in online platform regulations fall into one or more of the following broad categories: content-based, design-based, transparency, and procedural requirements.

Content-based requirements

The content-based approach generally involves establishing classes of prohibited content which online services are required to remove.

Prohibited content can be material that is illegal or defined as harmful or undesirable. The duties of online services with respect to such content may be reactive—that is, triggered by an official takedown order or user report—or proactive, requiring the services’ ongoing monitoring and removal of proscribed content.

Four Common Versions of the Content-Based Approach

Reactive obligations with respect to illegal content only

Proactive obligations with respect to illegal content only

Reactive obligations with respect to illegal + other content deemed harmful or undesirable

Proactive obligations with respect to illegal + other content deemed harmful or undesirable

The most conservative version of the content-based approach establishes **reactive** obligations with respect to **illegal content only**. This is the version adopted by the EU under the Terrorist Content Online Regulation (TCOR):

EU TCOR: Hosting services must remove or disable access to terrorist content within 1 hour if given an order by the competent authority of the respective Member State.³

Most regulations have adopted a more expansive version of the content-based approach by enlarging the category of proscribed content to include content **deemed harmful or otherwise undesirable**. Several jurisdictions have adopted this variant, including:

Singapore: Service providers must comply with orders by the Infocomm Media Development Authority (IMDA) to block access for Singapore users to “egregious content,” which is defined as content advocating for or instructing self-harm, suicide, violence, sexual violence, terrorism, depicting children for sexual purposes, or content advocating engaging in conduct that might endanger public health.⁴

Australia: Services must comply with removal or “remedial” notices issued by the eSafety Commissioner regarding adult cyber abuse, image-based abuse (non-consensual sharing of intimate images), child cyberbullying material, Class 1 material, and Class 2 material.⁵

New Zealand: Services must comply with court orders to take down or disable public access to material found by the courts to constitute a harmful digital communication.⁶

Texas: In addition to responding to known instances of illegal content, providers must reactively restrict harmful content once identified, applying filtering technologies and human moderation to remove or block harmful material from known minors’ accounts.⁷

Some jurisdictions go further, requiring services to take **proactive** action with respect to prohibited content. Such proactive action can range from notifying authorities if they become aware of illegal content, to requiring services to engage in continuous scanning of all communications to detect potentially unlawful content. The following three jurisdictions exemplify this range:

EU DSA: Where a hosting service becomes aware of any information giving rise to a suspicion that a criminal offense involving a threat to life or safety has taken place, is taking place, or is likely to take place, it must promptly inform law enforcement or judicial authorities of the Member State(s) concerned and provide all relevant information available.⁸

Australia OSA: Across the Industry Codes and Standards administered by the eSafety Commissioner under the Online Safety Act 2021, services that become aware of Class 1A or 1B material on their platform must remove it as soon as practicable. Services are also required to implement systems, processes and technologies to detect and remove child sexual abuse and pro-terror material where technically feasible, reasonably practicable, and where the measures would not undermine end-to-end encryption or introduce a systemic weakness.⁹

India IT Rules: Significant Social Media Intermediaries (SSMIs) and Online Gaming Intermediaries must deploy automated tools to proactively identify content involving rape, child sexual abuse, or content that was previously removed.¹⁰

The most expansive version of the content-based approach involves establishing **proactive** duties with respect to **illegal and other content deemed harmful or undesirable**. This version has been adopted by seven of the jurisdictions under analysis: the UK, Singapore, Australia, Ireland, South Korea, South Africa, and Texas. Most jurisdictions, with the exception of Texas, stop short of prescribing specific proactive measures.

UK OSA: In addition to requiring reactive and proactive measures with respect to illegal content, services likely to be accessed by children must use proportionate systems and processes¹¹ to prevent children from encountering “primary priority content” (i.e., pornographic content, suicide and self-harm content, eating disorder content) and protect children in age groups judged to be at risk of harm from “priority content” (e.g., abuse and hate content, bullying content, violent content, harmful substances content) as well as “non-designated content.”¹²

Texas SCOPE Act: Digital service providers must proactively use a combination of filtering technology, hash-sharing, and a regularly updated list of harmful keywords or identifiers to block illegal content before it reaches minors.¹³ Further, providers are required to perform human reviews to ensure that filtering technologies are effective in identifying illegal content.¹⁴

A more recent variant of the content-based approach consists of preventing online services from taking down content. Often called “must-carry” provisions, these types of regulations have been motivated by the perception that platforms unfairly suppress some political viewpoints. Such laws have been introduced in the US and Brazil but have met with strong criticism and constitutional hurdles.¹⁵

Finally, a limited number of content-based provisions focus on requiring platforms to communicate certain content. For example, under Singapore’s **OSA**, some services (“designated services”) must provide users, including children, with local information including Singapore-based safety resources.¹⁶

Human rights standards relevant to platform regulation

International human rights law contains a body of standards relevant to the governance of digital spaces. These standards—particularly regarding freedom of expression, privacy, children’s rights, and non-discrimination—should guide how states approach regulating the internet.¹

The International Covenant on Civil and Political Rights (ICCPR), a widely ratified human rights treaty, protects the right to freedom of expression (Article 19). The UN Human Rights Committee, which interprets the treaty, has explicitly stated that Article 19 covers online expression. Under the ICCPR, any government restriction on freedom of expression, whether online or offline, must be:

1. Provided by law (the *legality* principle): The restriction must be clearly established in law, so platforms and users can understand what is prohibited.
2. In pursuit of a legitimate aim (the *legitimacy* principle): Commonly recognized aims include respect for the rights or reputations of others, and protection of national security, public order, public health, or morals.
3. Necessary and proportionate: Restrictions cannot be overbroad; they must be narrowly tailored to achieve the legitimate aim. Blocking entire websites or platforms is usually considered a disproportionate measure.²

ICCPR Article 17 also protects against unlawful or arbitrary interference with privacy. In the online context, the right to privacy is often read to require restrictions on user data collection, surveillance, and law enforcement access to data.³

¹ Global Online Safety Regulators Network (GOSRN) Position Statement, [Human Rights and Online Safety Regulation](#), September 2023.

² UN Human Rights Committee, General Comment No. 34.

³ Human Rights Committee, General Comment No. 16; OHCHR A/HRC/27/37.

Commentary

Many jurisdictions have been drawn to the content-based approach as a surgical way to address downstream harms caused by content hosted on the platforms. But this approach is deficient in several respects.

First, it may involve government infringement of freedom of expression. International human rights law allows governments to limit expression, but only in a way that is set out clearly in legislation and is necessary and proportionate to achieve a legitimate governmental objective.¹⁷ Regulations that require platforms to remove vaguely defined categories of “harmful” or “egregious” content fail to meet the “[legality](#)” principle established by international human rights law and invite overbroad government enforcement. Those that require platforms to take unspecified proactive measures fail the “[necessity](#)” and “[proportionality](#)” principles because they incentivize platforms to remove more content than necessary to avoid potential liability.

In the US, the First Amendment to the Constitution prevents the government from restricting freedom of speech except in very [narrow](#) circumstances, which makes the content-based approach unconstitutional in most cases.¹⁸

Second, the adoption of content-based regulation is likely to contribute to global fragmentation in online platform regulation and a balkanization of internet communications. What is illegal in one jurisdiction may not be illegal in another. While the advent of regulatory fragmentation is not dispositive, it complicates companies’ compliance efforts and hinders global regulatory coherence.

Third, the content-based approach, while necessary to address specific instances of illegal material, is unsuitable on its own for dealing with the scale of online interaction. Most content moderation systems, whether reactive or proactive, rely on algorithmic systems to implement content rules to the billions of pieces of content uploaded each minute. But those algorithms are imperfect to different degrees and require human oversight, including manual review of certain content. While there are ways for platforms to streamline and speed up content moderation, those efforts can never keep pace with the sheer volume and speed of online discourse. For this reason, relying on content regulation and moderation systems *alone* is insufficient to address harms at scale.

Design requirements

The design-based approach mandates technical and interface-related changes to achieve certain outcomes, such as protecting users’ data privacy, empowering users to customize their experience, and reducing compulsive usage. This approach differs from the content-based approach in two important ways: It focuses on upstream harm prevention, rather than downstream (after-the-fact) mitigation, and it regulates platforms as products, targeting their architecture and features.

Some jurisdictions take a highly prescriptive approach toward safety-by-design regulation, setting forth requirements for specific design features. For example:

Louisiana’s SOCIAL Act prohibits platforms from enabling the direct messaging feature between adults and Louisiana minors unless the two are already connected.¹⁹

California’s Addiction Act prohibits services from sending notifications to minor users between the hours of 12 a.m. and 6 a.m., and during school hours (8 a.m. to 3 p.m., Monday to Friday, from September through May) unless they have obtained verifiable parental consent.²⁰

Other regulations focus on enhancing user agency and choice through feature customization options. For example, Singapore’s [OSA](#) requires regulated services to provide users with “tools that enable them to manage their own safety,” including “tools to restrict visibility of harmful content,” “to limit visibility of the end-user’s account,” and “to limit location sharing.”²¹

Examples of design features and how they can be regulated¹

Privacy settings, including default settings

- **Geolocation tracking can be set to on or off**

Example: Under the UK Children’s Code, services must set geolocation tracking off by default, and make any active location tracking visible to children (Standard 10).

- **Account visibility can be set to public, semi-public or private**

Example: Under Utah’s Minor Protection in Social Media Act, social media companies must set default privacy settings for minor users so as to restrict their visibility to only connected accounts (Section 13-71-202(1)).

Notifications

- **Push notifications can be enabled or disabled, or enabled only at specified times**

Example: Under the NY SAFE for Kids Act, platforms are prohibited from sending notifications regarding addictive feeds to minors between 12 AM and 6 AM without verifiable parental consent (Section 1502).

Algorithmic feeds²

- **Feeds can be curated by the platform’s chosen algorithm based on specific data, e.g., user engagement data, or they can be customized by the user (e.g., based on stated interests, friend lists, or simply reverse-chronological)**

Example: Under the EU’s DSA, very large online platforms and search engines must provide at least one option for each recommender system that is not based on profiling (Article 38).

Self-help tools

- **Lists of blocked contacts**

Example: Under Australian Industry Codes and Standards, many services that enable peer-to-peer messaging must allow Australian users to block messages from other users and hide their online status (Industry Standard on Class 1A and Class 1B material, RES Standard, Section 18(4)).

- **Reporting tools**

Example: Under California’s Age-Appropriate Design Code, services likely to be accessed by children must “provide prominent, accessible, and responsive tools to help children, or if applicable their parents or guardians, exercise their privacy rights and report concerns” (Section 1798.99.31(a)(10)).

¹ For a comprehensive taxonomy of platform design features and their relation to consumer harms, see USC Neely, KGI & Tech Law Justice Project, [Design Element Taxonomy](#) (work in progress).

² While algorithmic recommendation systems are design elements, their design can involve instructions regarding content. For example, an algorithm could be instructed to recommend user posts which have gathered the largest number of comments and shares (a content-neutral instruction) and/or it could be instructed to elevate posts that have to do with sports (a content-dependent instruction).

A hybrid version of these two approaches consists of requiring platforms to make certain settings the default option for users while still allowing users or their legal representatives to change those settings according to their preference. The latter is the case under [Utah's Minor Protection in Social Media Act](#), which states that social media companies must set default privacy settings for users under 16 to prioritize maximum privacy and allow changes to those settings only with verifiable parental consent.²²

Several jurisdictions have chosen to limit the applicability of certain design-based requirements to underage users.²³ In creating various requirements for users of different ages, these laws often trigger the need for platforms to determine who is a minor. Accordingly, some regulations prescribe a specific method of age assurance.²⁴ Many regulations simply require that platforms adopt a “commercially reasonable” method without further elaboration,²⁵ but a number of regulatory bodies have started issuing more concrete guidance and requirements regarding the implementation of age assurance technology.²⁶

Instead of, or in addition to, establishing specific design requirements, some regulations impose a general obligation on platforms to implement features with user safety in mind. These regulations place the onus on platforms and enforcers to determine the scope of this duty and the criteria for assessing noncompliance.

Conversely, some jurisdictions require that platforms refrain from implementing features—sometimes called “dark patterns”—that nudge users toward harmful behaviors such as compulsive usage. The EU, UK, Ireland, California, and Maryland contain explicit provisions banning manipulative designs or dark patterns, without necessarily specifying what those features are.

The following is a summary of emerging approaches to design-based regulation:

- Requiring that platforms implement specific design elements or adjust features in accordance with the regulation. These requirements can apply to all users or a subset (e.g., children).
- Requiring that platforms make specific design elements and features adjustable by users—or their legal representatives—according to their needs and preferences.
- Establishing a general duty to implement features with user safety in mind.
- Prohibiting the use of features that nudge users toward harmful behavior, such as compulsive usage or unwanted spending (“dark patterns”).

Commentary

The advantage of design-based regulation is that it focuses on features that shape users’ online behavior and experience instead of regulating content directly.²⁷ Aside from being largely²⁸ content-neutral, thereby reducing government intervention in speech, regulation of platform designs takes a systemic and preventative approach by addressing harms upstream.

Nevertheless, design-based regulation has its limitations. The approach rests on an assumption that specific features are linked to the harms that regulators seek to reduce. Yet the academic study of these links is at its infancy, and there is a need for more evidence substantiating the impact of specific platform features on users.²⁹ Platforms contain some of this evidence; they routinely track the impact of their product choices on users.³⁰ Requiring platforms to disclose some of this internal data should be a precursor to establishing prescriptive design requirements. (See the related discussion below on transparency).³¹

While evidence on the effects of specific designs accumulates, regulators could opt for a general duty to design features with user safety in mind, which shifts the burden onto platforms to test the impact of new product features *before* they get rolled out and track that impact continuously. Platforms constantly evolve, and general obligations with respect to “[safety-by-design](#)”³² provide sufficient flexibility for regulators and platforms to adjust compliance measures based on the available evidence. Because of the ambiguity inherent in general obligations, however, such provisions should be accompanied by enough regulatory guidance and constraints on enforcement powers to ensure their fair and foreseeable implementation.

Another promising approach is to require platforms to implement features that enhance user agency by allowing them to customize key aspects of their experience, such as their exposure to contact by strangers and subjection to personalized algorithmic recommendations based on sensitive data. Empowering users to exercise choice through specific design features is a promising mitigation measure for a variety of online harms.

Regulations that impose design requirements for a subset of users—usually minors or children—are subject to another complication: they imply a requirement for platforms to disaggregate their user base according to age, with significant consequences for data privacy and security.³³ In adopting any method of age assurance, platforms face a tradeoff between efficacy and invasiveness.³⁴ Methods that more accurately and reliably determine whether a particular user is underage need to collect more personal data from that user. Conversely, those that limit data collection are more likely to lead to false positives and false negatives.

As a way to sidestep this tradeoff and associated compliance challenges, regulators could simply extend design-based protections to all users. Louisiana has taken this approach, by requiring that social media platforms either make commercially reasonable efforts to verify the age of account holders or “apply the accommodations afforded to minors” under the law to all users.³⁵

In sum, regulators can help realize the potential of design-based regulation by heeding these recommendations:

- Regulators should wait to establish highly prescriptive design requirements until there is enough evidence of their relationship to relevant harms. Meanwhile, they should prioritize mandating design changes that allow users to customize aspects of their online experience that impact their rights or wellbeing.
- When regulating algorithmic recommendation systems, regulators should ensure that the regulation targets content-neutral design aspects rather than content-dependent determinations.
- When enshrining a general duty to design features with user safety in mind, regulators should subsequently provide sufficient standards and metrics to clarify compliance expectations.
- Instead of requiring platforms to implement age assurance when applying design-based protections, regulators should require that platforms apply those protections to all users.

Transparency requirements

The transparency-based approach to platform regulation sets forth requirements for online services to disclose information about their operations, revenue streams, algorithms, and moderation processes. Rather than prescribing specific rules for what content is allowed or how platforms should function, this approach aims to make platforms accountable by exposing their practices to scrutiny by regulators, researchers, and the public. It is a favored approach among those who consider content-based regulation problematic on constitutional grounds and design-based regulation too prescriptive or premature.

Online platforms can be compelled to make a number of disclosures. A small number of jurisdictions require platforms to release basic information about their userbase, such as the total number of users and number of under-age users.³⁶ But by far the most common requirement is for platforms to produce reports disclosing aggregate data and other information on their moderation of third-party content. Several platforms started producing such reports before being required to do so by law,³⁷ but the reports have lacked consistency and standardization.

The EU’s DSA contains the most extensive provisions on transparency reports.³⁸ The UK, Singapore, Australia, Ireland, India, and California also contain explicit requirements for platforms to produce one off and/or periodic reports on their content moderation and safety features, with varying levels of specificity on the metrics they need to include.³⁹

Another emerging trend in transparency regulation is to mandate the disclosure of information about the workings of algorithms which recommend content and target advertisements to users. The EU's DSA again has the most extensive provisions on this front, although other jurisdictions such as Texas are starting to require some degree of algorithmic transparency.⁴⁰

EU DSA: Platforms must include in their terms and conditions the main parameters used in recommender systems and any options users have to modify or influence those parameters. Platforms must explain why certain information is suggested to a user, including, at least: Criteria that are most significant in determining information suggested to users and reasons for relative importance of parameters.⁴¹ Platforms must also ensure users are able to identify relevant information about each advertisement presented to them on the platform, including the provenance/sponsor and reasons behind targeting. Platforms must not present ads to users based on profiling using special categories of personal data.⁴¹ Very large platforms must compile and make publicly available on their online interface a repository containing certain information about an ad, for the entire period they present the ad and until one year after.⁴³

Texas SCOPE Act: Providers must disclose their algorithmic practices in clear and accessible language. This involves detailing how algorithms rank, filter, and present content to minors, as well as information on the types of personal data used in the algorithms.⁴⁴

Algorithmic disclosures can also be embedded in risk and/or impact assessments which platforms may be required to conduct as a procedural safeguard against facilitating human rights harms (see section on procedural requirements for more details). Risk and impact assessments, a type of human rights due diligence, serve as a transparency mechanism when platforms are compelled to publicly disclose their assessments, mitigation measures, and impact reports. This is the case under a number of online platform regulations, including the EU's DSA and UK's OSA.

Finally, to ensure the accuracy and comprehensiveness of such disclosures, regulators may mandate that very large platforms submit themselves to independent audits and provide vetted researchers with access to platform data.⁴⁵ The EU is the only jurisdiction in the sample explicitly requiring certain platforms to do both.⁴⁶

Commentary

Making information accessible and public fosters accountability and helps prevent harmful, self-serving behavior. But transparency on its own does not guarantee specific company actions or policy outcomes. It is a crucial but insufficient aspect of online platform regulation aimed at reducing harm. A key target of platform disclosures should be their algorithmic recommendation systems—the software mechanisms that platforms use to suggest relevant content, including advertisements, to users. These systems are typically proprietary⁴⁷ so policy makers and the public have little insight into the factors that determine which pieces of content, among the billions of possibilities, get recommended to users in their feeds. Requiring platforms to disclose a descriptive account of these factors—as the EU's DSA and Texas' SCOPE Act do—allows regulators to then evaluate whether they should establish content-neutral design standards alongside user engagement.⁴⁸

Another priority for regulators should be to expand outside researchers' access to platform data as a way to enable independent assessment of platforms' operations and impact. The EU's DSA has the clearest provision mandating such access, and EU regulators are now fleshing out the details of this data access regime.⁴⁹ Other jurisdictions should explore doing the same.⁵⁰ In doing so, regulators should balance research needs against user privacy and data security concerns as well as valid technical and cost-related challenges.

Alongside demanding content-neutral disclosures and expanding researcher access to platform data, regulators need to be more specific about the categories of information they expect platforms to divulge in their content moderation reports and risk assessment and impact reports. Currently, regulatory requirements tend to leave too much room for platform discretion on which metrics to disclose and methodologies to employ, leading to company disclosures that are biased or uninformative.⁵¹ Moreover, resulting

disparities in companies' reports complicate regulators' task of evaluating and comparing performance across time. Regulatory bodies should develop robust standards and metrics to guide companies in meeting their transparency reporting requirements.

Procedural requirements

The last approach to online platform regulation focuses on platform processes aimed at ensuring basic fairness and accountability. These include requirements for platforms to:

1. Lay out their terms of service, including content and conduct policies, in clear and accessible language.
2. Live up to those terms of service, including any commitments made toward users, through the actual operation of their services.
3. Assign points of contact and legal representatives to answer pertinent user and other stakeholder queries, and make their contact information publicly accessible.
4. Conduct risk and/or impact assessments identifying how their platforms might lead to individual or societal harms and describing efforts to mitigate those harms.

Most online platform regulations explicitly or implicitly contain a requirement for services to publish Terms of Service (ToS) laying out key policies and practices with respect to safety, data privacy, and other consumer interests. Some require that platforms provide such information in accessible language, including child-friendly language.⁵²

Topics for Disclosure in Terms of Service

Content policies

- Example: Under the EU's DSA, platforms must include misuse policy in terms and conditions and give examples of facts/circumstances taken into account (Article 23(4)).

Measures to protect children

- Example: Under the UK's OSA, services must include provisions in their terms of service which specify how children will be prevented from encountering "primary priority content" and how those in age groups judged to be at risk will be protected from encountering "priority content" and "non-designated content" (Sections 12(9) and 12(10)).

Functioning of algorithms

- Example: Under the TX SCOPE Act, providers using algorithms to deliver or filter content must disclose how they use algorithms, including details on ranking, promotion, and filtering. This information must be accessible in the terms of service or privacy policy (Section 509.056).

Data handling practices

- Example: Under Utah's MPSM Act, social media companies must, for minor users, "provide an easily accessible and understandable notice that: (a) describes any information the social media company collects from a Utah minor account holder; and (b) explains how the information may be used or disclosed" (Section 13-71-202(3)).

Additionally, some regulations provide that services must live up to their ToS. These requirements can be general. For example:

UK Children’s Code: Services likely to be accessed by children must uphold their own “published terms, policies and community standards (including but not limited to privacy policies, age restriction, behaviour rules and content policies)” (Standard 6).

California Age-Appropriate Design Code: Services likely to be accessed by children must enforce published terms, policies, and community standards established by the business (Section 1798.99.31(a)(9)).

Other regulations prescribe specific procedural mechanisms to ensure that platforms abide by their ToS. If a platform purports to moderate content in their ToS—as some jurisdictions require by law that they do—they must also make sure their moderation systems follow fair and effective procedures. This means that platforms must review and address user reports in a timely fashion; notify users when their accounts or posts have been subject to moderation and provide an explanation of that action; and give affected users an opportunity to appeal the enforcement action. The EU’s DSA contains the most extensive provisions on procedurally adequate moderation.⁵³ But such provisions are common among online safety regulations.⁵⁴

Some regulations require platforms to designate points of contact and legal representatives to respond to user queries and be held accountable in case of legal noncompliance. The most extensive requirement in this category again comes from the EU’s DSA, which mandates that all intermediary services make easily accessible the contact information of a single point of contact for users to communicate with directly, rapidly and by electronic means.⁵⁵ Intermediary services must also designate a legal representative in one of the EU Member States where they offer services and make their information publicly available.⁵⁶

Finally, a growing number of jurisdictions require platforms to proactively assess, prevent, and mitigate human rights risks in their operations.⁵⁷ A type of human rights due diligence, such risk and/or impact assessments are increasingly a cornerstone of online platform regulations. In general, they establish platforms’ obligation to engage in a recurring process to (1) identify human rights risks related to content moderation, data processing, and algorithmic-driven decisions; (2) engage with stakeholders, including affected communities and civil society groups; and (3) implement mitigation strategies and remedial mechanisms.

Commentary

Procedural requirements are among the most common and least controversial in online safety regulations because they relate to fundamental expectations of fairness. In general, these requirements aim to hold platforms accountable for the promises they make toward users.

Nonetheless, some obligations that fall into this category can be far-reaching and potentially unbounded. For example, requirements that platforms operate well-functioning reporting mechanisms can be vague and need to be accompanied by sufficiently specific standards to inform compliance and ensure fair enforcement.

Similarly, regulators’ expectations related to human rights risk and/or impact assessments have yet to be fleshed out. An initial review of the first risk assessments published pursuant to the EU’s DSA revealed considerable variation across platforms in terms of the specificity, comprehensiveness, and overall rigor of their assessments.⁵⁸ Regulators should therefore aim to produce more specific instructions and guidance for platforms on the expected methodologies while ensuring enough flexibility to account for differences across types of online services.

Approaches to enforcement

Online safety regulations exhibit a broad range of enforcement approaches. They vary in terms of type of **entities** tasked with enforcement and the **powers** conferred to those entities.

The most common types of enforcement authorities are agencies or commissions with variable degrees of independence from the political branches. In the US, by contrast, most state-level platform regulations confer enforcement powers on state attorneys general, who are either directly elected by voters or appointed by other state officials or entities.⁵⁹ A few regulations, such as New Zealand’s HDCA and Fiji’s OSA, confer enforcement powers on courts instead.⁶⁰

In terms of enforcement powers, the most common involve fines and blocking or access restriction orders. A minority of regulations under review provide for criminal penalties, empowering courts to issue prison sentences for the employees of platforms who fail to comply with specific orders.⁶¹ Finally, a small number of jurisdictions establish private rights of action, allowing individuals to sue platforms directly for failing to comply with their duties.⁶²

Approaches to enforcement		Enforcement Entity	
		Less Independent	More Independent
Enforcement Powers	Less Punitive	Example: Singapore	Example: New Zealand
	More Punitive	Example: China	Example: Fiji

3. The Road Ahead

“
Our recommendations
are intended to guide
policymakers in
developing effective,
evidence-based, and
human rights-compliant
online safety regulations
going forward.”

Regulation cannot reduce, let alone eliminate, every potential online harm. However, for too long, the internet operated without meaningful guardrails, creating an environment where platforms wielded immense influence over individuals and society with little accountability.

The proliferation of online safety regulations is evidence of a growing international consensus that some kind of oversight is necessary. The challenge now lies in finding meaningful approaches that are consistent with international human rights standards.

Some regulations—including several under consideration—do not meet these standards. In the US, a pair of [state laws](#) seek to bar platforms from “censoring” users based on their “viewpoint.”⁶³ While cloaked as a defense of free speech, these politically motivated proposals undercut platforms’ own free expression right to choose the content they wish to host, within legal bounds. A similar [bill](#) seeking to limit companies’ content moderation powers has been proposed in Brazil.⁶⁴ While these legislative proposals include some worthy provisions, including procedural safeguards and disclosure requirements, they vest too much power in government actors to determine what is a “just” moderation policy. In the US at least, these laws are likely to be found unconstitutional.⁶⁵

Similarly, laws that compel platforms to remove vaguely defined categories of “harmful,” but not illegal, content are problematic from a human rights perspective as they allow governments to ban potentially legitimate speech indirectly. Several of the enactments discussed in this study are problematic in this regard. Further, at least two bills under consideration—Brazil’s “Fake News” Law⁶⁶ and Canada’s Online Harms Act⁶⁷—are also subject to this critique.

On the other hand, some stalled legislation in the US, such as the [Platform Accountability and Transparency Act](#) (PATA)—which would expand independent researcher access to platform data—and the [Kids Online Safety and Privacy Act](#) (KOSPA)—which would create a duty of care in the implementation of platforms’ design features, strengthen data privacy for minors, and expand avenues for research, among other measures—merit reconsideration and approval.

The recommendations on the next page are intended to guide policymakers in developing effective, evidence-based, and human rights-compliant online safety regulations going forward.

Recommendations for Regulators

1 Ensure that content-based requirements pertain only to content that is explicitly illegal, or content that meets the “legality” standard under human rights law.

Governments have a legitimate reason to crack down on illegal speech and conduct online, but they should do so consistent with human rights principles.⁶⁸ Regulations that require platforms to reactively or proactively remove vaguely defined categories of harmful but not illegal—or “awful but lawful”—content invite subjective and overbroad government enforcement. Requirements that platforms proactively scan for and remove unlawful content should be reasonable, consistent with data privacy rights, and accompanied by mechanisms which enable platforms to meet their obligations (e.g., the [CyberTipline](#) run by the National Center for Missing and Exploited Children (NCMEC) to report suspected child sexual abuse material).

2 Establish meaningful transparency requirements to enhance understanding of platforms’ systems and impacts.

Regulators should demand content-neutral disclosures, ranging from basic quantitative data, such as number of daily active users, to information about the main parameters used in algorithmic recommendation systems, and resources employed in any content moderation undertaken.⁶⁹ Regulations should also require platforms to disclose detailed information about how they handle user data, including how data is fed into recommender systems.⁷⁰

Additionally, regulators should mandate that platforms provide meaningful information about their content moderation systems and actions. Several major platforms already publish periodic “transparency reports” with rudimentary data about their policy enforcement, but these reports tend to leave out key information.⁷¹ While platforms should have a right to determine the substance of their content policies, regulators should require that companies disclose data showing what the experience inside the platform is *actually* like and demonstrating their efforts to change or improve that experience. These disclosures should be subject to external scrutiny, such as by independent auditors.

Finally, regulators should establish secure and privacy-compliant mechanisms for vetted researchers to request more granular data from platforms to enable independent study of platforms’ systems and impacts. For example, researchers should be able to request access to user communication data and platforms’ product experimentation data, with enough safeguards to protect user privacy and legitimate trade secrets.⁷²

3 Regulate design features to enhance user agency and foster development of evidence-based design standards.

Regulators should crack down on the use of “dark patterns” and incentivize platforms to create design features that allow users to customize aspects of their online experience that impact their rights and wellbeing. These protections should apply to all platform users, not just children. A key target for user customization should be platforms’ algorithmic recommendation systems which largely determine the content users consume. Any highly prescriptive design-based mandates, such as mandatory limits on the use of push notifications and direct messaging, should be evidence-based⁷³ and proportional to the regulation’s aims. Meanwhile, regulators should incentivize platforms to test the safety of their design features *before* rolling them out, consistent with the principle of [safety-by-design](#).

4 Ensure that procedural requirements are about more than just box-ticking.

For procedural safeguards to be meaningful, regulators need to issue binding codes of practice and concrete implementation guidance setting out clear expectations. For example, platforms need to know what an adequate risk assessment should contain, or what a functional user reporting mechanism looks like. Regulators should also put teeth in any requirements that platforms fulfill the promises they make towards users in their ToS. If a platform claims to prioritize user safety, as some do, regulators should require that companies demonstrate their investments in trust and safety, including in content moderator workforces and systems.

5 Assign enforcement to an independent agency with limits on its authority to safeguard individual freedoms.

Regulators should ensure that this agency is appropriately funded and staffed with expert personnel. Penalties for noncompliance, such as fines or content access restrictions, should be proportional to the harms and narrowly tailored to minimize infringing on user rights. Sweeping platform bans should be a measure of last resort. Criminal sanctions, such as prison sentences for platform employees due to noncompliance with a government orders, should only be used in cases of clear criminal conduct.

6 Adopt nuanced definitions that account for differences across platforms.

When setting out the regulation's scope, regulators should ensure that they do not overlook and inadvertently exclude important sectors, such as gaming platforms, from coverage. At the same time, regulators should understand the differences across types of platforms—such as social networking, video streaming, messaging, and gaming—and tailor requirements appropriately. Regulators should also consider differentiating between platforms based on their size (by revenue or user number), imposing more onerous requirements on large platforms which have an outsized impact and greater resources to comply.⁷⁴

7 Work together with regulatory counterparts in privacy and antitrust departments to ensure requirements are compatible and, ideally, mutually reinforcing.

In some instances, regulations establish requirements that, while reasonable in isolation, are impractical in combination. For example, a few regulations mandate that platforms maximize both user privacy and safety—while these two interests are generally complementary, there are some cases where tradeoffs are necessary.⁷⁵ Regulators should acknowledge that these interests are sometimes in tension and allow platforms a reasonable degree of flexibility in striking a balance.

8 Work towards international cooperation and coherence on platform governance.

While regulations are specific to each jurisdiction, internet platforms operate across national borders. Regulators should strive to achieve gradual convergence in online safety regulations as a way to fortify protections for users and promote compliance. They can do so by participating in multilateral initiatives, such as the [Global Online Safety Regulators Network \(GOSRN\)](#),⁷⁶ which provides a forum for independent regulators worldwide to share information, best practices, and tools to enhance global regulatory coherence and effectiveness. Regulators should also engage with a wide range of stakeholders, including civil society organizations, academic researchers, and people with relevant lived experience, who can offer perspectives on the real-world impact of regulatory measures.

Appendix

	Content-based requirements				
	Reactive duties regarding illegal content only	Reactive duties regarding illegal + harmful content	Proactive duties regarding illegal content only	Proactive duties regarding illegal + harmful content	Must-carry content requirements
Singapore	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> *	<input type="checkbox"/>
Australia	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
New Zealand	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
South Korea	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Fiji	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
EU	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/> ‡	<input type="checkbox"/>
Ireland	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
UK	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Canada	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
California	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
New York	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Maryland	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Texas	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Utah	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Brazil	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
India	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
South Africa	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Colorado	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Louisiana	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Florida	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

* † ‡ - See page 23 for legend

	Design-based requirements						
	Default settings	Feature restrictions	User customization + self-help tools	Parental controls	Prohibition of manipulative designs (“dark patterns”)	Algorithmic recommendation systems	General duty to implement features with safety in mind
Singapore	✓*	<input type="checkbox"/>	✓*	✓*	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Australia	✓	✓	✓	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
New Zealand	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
South Korea	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Fiji	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
EU	✓	<input type="checkbox"/>	<input type="checkbox"/>	✓*	✓	✓	<input type="checkbox"/>
Ireland	<input type="checkbox"/>	<input type="checkbox"/>	✓	✓	✓	✓	✓
UK	✓	<input type="checkbox"/>	✓	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>
Canada	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓
California	✓	✓	✓	✓	✓	<input type="checkbox"/>	<input type="checkbox"/>
New York	<input type="checkbox"/>	✓	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Maryland	✓	<input type="checkbox"/>	✓	✓	✓	<input type="checkbox"/>	<input type="checkbox"/>
Texas	✓	<input type="checkbox"/>	✓	✓	<input type="checkbox"/>	✓	<input type="checkbox"/>
Utah	✓	<input type="checkbox"/>	✓	✓	✓	<input type="checkbox"/>	<input type="checkbox"/>
Brazil	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
India	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>	✓
South Africa	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Colorado	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Louisiana	<input type="checkbox"/>	✓	<input type="checkbox"/>	✓	<input type="checkbox"/>	✓	<input type="checkbox"/>
Florida	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

* † ‡ - See page 23 for legend

	Transparency requirements					
	Content moderation reports	User numbers and demographics	Algorithmic recommendation systems	Publication of human rights risk and/or impact assessments	Independent Audits	Researcher access
Singapore	✓*	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓
Australia	✓	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>
New Zealand	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
South Korea	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Fiji	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
EU	✓	✓	✓	✓*	✓*	✓*
Ireland	✓	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>
UK	✓*	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/> †
Canada	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>
California	✓	✓	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>
New York	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Maryland	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Texas	✓	<input type="checkbox"/>	✓	✓	✓	<input type="checkbox"/>
Utah	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Brazil	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
India	✓	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
South Africa	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Colorado	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Louisiana	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Florida	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

* † ‡ - See page 23 for legend

	Procedural requirements						
	Notice and informed consent	Fulfilling ToS	Fair terms of use	Points of contact and legal reps	Reporting and appeal mechanisms	Record-keeping	Human rights risk and/or impact assessments
Singapore	✓*	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓*	<input type="checkbox"/>	<input type="checkbox"/>
Australia	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	✓
New Zealand	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
South Korea	✓	✓	✓	<input type="checkbox"/>	✓	✓	<input type="checkbox"/>
Fiji	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
EU	✓	<input type="checkbox"/>	<input type="checkbox"/>	✓	✓	✓	✓*
Ireland	✓	✓	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	✓
UK	✓	✓	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	✓
Canada	✓	<input type="checkbox"/>	<input type="checkbox"/>	✓	✓	✓	✓
California	✓	✓	✓	✓	✓	<input type="checkbox"/>	✓
New York	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Maryland	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	✓
Texas	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	✓
Utah	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Brazil	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>
India	✓	<input type="checkbox"/>	<input type="checkbox"/>	✓	✓	✓	<input type="checkbox"/>
South Africa	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	✓	<input type="checkbox"/>
Colorado	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Louisiana	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Florida	✓	<input type="checkbox"/>	✓	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

* † ‡ - See page 23 for legend

The analysis of online safety regulations proceeded in three steps.

- First, we conducted a comprehensive global survey of internet regulations, relying on existing databases or “trackers”⁷⁷ to identify relevant regulations for this study.
- Second, we distilled the provisions in each regulation, noting recurring requirements that, when grouped into categories, revealed distinct approaches to online safety regulation.
- Third, we examined the provisions of the 26 selected regulations in light of these categories and developed subcategories corresponding to specific types of provisions.

This process, and the resulting classification, highlighted considerable diversity yet also notable convergence in online safety regulations.

The legend for the tables on pages 19-22 is as follows:

* Applies to a subset of services. For the EU, those designated as “very large online platforms and search engines”; for Singapore, “designated services” or “regulated services,” and for the UK, those designated as “categorized services.”

† Ofcom is required under Section 162 of the OSA to prepare a report evaluating independent researchers’ access to platform data for studying online safety. The report must assess the current level of access, legal and practical constraints, and potential methods to enhance data-sharing. While the OSA does not mandate platforms to provide access, Ofcom must review and report on the issue.

‡ Under Article 35, the EU’s DSA requires the providers of very-large online platforms (VLOPs) and very large online search engines (VLOSEs) to “put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to Article 34.” The systemic risks in Article 34 include risks that flow from content that is harmful but not illegal. Adequate mitigation measures may include “adapting their terms and conditions and their enforcement” (Article 35(b)). Further, Article 36(1) of the DSA allows the European Commission to require providers of very-large online platforms (VLOPs) to change their policies with respect to what content is allowed on their platforms. Article 36(3) provides that these requirements, which can only be implemented in periods of crisis, be “strictly necessary, justified and proportionate.” These provisions have never been implemented so their interpretation is unclear. Thus, some argue that the EU DSA’s risk mitigation provisions could be a form of indirect content-based regulation of harmful but not illegal content.

Endnotes

- 1 Emphasis placed on “significant.” The US federal government has passed legislation addressing narrow aspects of online platforms. The most recent is the Protecting Americans from Foreign Adversary Controlled Applications Act, a law that requires the sale of the Chinese-owned short-video platform, TikTok, to a non-Chinese owner or face a ban in the US. At the time of writing, however, this law is not being enforced pursuant to an executive order from President Trump. The US federal government has passed other federal legislation that impacts platforms—such as the REPORT Act, which made amendments to the federal framework concerning the reporting of online child sexual exploitation, and FOSTA-SESTA Acts, which allow federal and state governments to enforce anti-trafficking laws on online platforms.
- 2 The reason for this scope delimitation is that obligations concerning data privacy and cybersecurity are often contained in pieces of legislation that are separate from the regulations addressing online safety as such.
- 3 EU TCOR, Article 3.
- 4 Singapore Broadcasting Act, Sections 45D and 45H.
- 5 Australia OSA, Parts 5-7 and Part 9. Class 1 material refers to content that is refused classification under the National Classification Code, such as child exploitation material, drug-related content, and pro-terrorism content, which cannot be provided under any circumstances. Class 2 material refers to content that is legally restricted to adults (i.e., classified as R18+ or X18+ under the National Classification Code), for example, explicit sexual content and scenes of intense violence.
- 6 New Zealand HDCA, Section 19(2). A communication can be considered harmful if it is posted online with the intention to cause harm to a victim, a reasonable person in the position of the victim would find the communication likely to cause harm, and the communication does, in fact, cause harm to the victim (Section 22 of HDCA).
- 7 Texas SCOPE Act, Section 509.053(b)(E). Harmful content includes material that could negatively impact minors, such as content promoting suicide, self-harm, eating disorders, bullying, harassment, and substance abuse (Section 509.053(a)(1)-(3)).
- 8 EU DSA, Article 18(1).
- 9 Australia Online Safety (Designated Internet Services—Class 1A and Class 1B Material) Industry Standard 2024, s 18; Online Safety (Relevant Electronic Services—Class 1A and Class 1B Material) Industry Standard 2024 s 16.
- 10 India IT Rules, Rule 4(4).
- 11 Specific proactive measures are not defined.
- 12 UK OSA, Sections 12(2), 12(3) and (8)(e).
- 13 Texas SCOPE Act, Section 509.053(b)(A)-(D).
- 14 Texas SCOPE Act, Section 509.053(b)(E).
- 15 *Moody v. NetChoice LLC*, 144 S. Ct. 2383 (2024); Joan Barata, [Regulating Online Platforms Beyond the Marco Civil in Brazil: The Controversial “Fake News Bill,”](#) May 2023.
- 16 Singapore Broadcasting Act, Code of Practice for Online Safety. Canada’s Bill C-63 also requires operators of online services to label content generated by automated programs if there are reasonable grounds to believe it involves harmful content or is overly prevalent due to automated generation (Section 60).
- 17 International Covenant on Civil and Political Rights, Article 19.
- 18 Some exceptions are laws requiring platforms to take down child pornography, violent incitement and other narrow categories of speech that are not protected under the US Constitution.
- 19 Louisiana SOCIAL Act, Section 51:1753(1).
- 20 California Addiction Act, Section 27002(a)(1). The NY SAFE Act contains a similar provision (Section 1502).
- 21 Singapore Broadcasting Act, [Code of Practice for Online Safety](#).
- 22 Utah MPSM Act, Sections 13-71-202(1), 13-71-204(1).
- 23 The age of a minor is defined differently across jurisdictions.
- 24 See Scott Brennen & Matt Perault, [Keeping Kids Safe Online: How Should Policymakers Approach Age Verification?](#) The Center for Growth and Opportunity at Utah State University, June 2023.
- 25 Some jurisdictions provide that such methods cannot be limited to collecting an official government ID given the implications for privacy and inclusion. See, e.g., NY SAFE Act Section 1501(2)(c) and Louisiana SOCIAL Act Section 51:1752(D).
- 26 See, e.g., eSafety Commissioner, [Age Assurance](#), Tech Trends Issue Paper, July 2024; Ofcom, [Statement: Age Assurance and Children’s Access](#), January 2025.
- 27 Uri Gal, [Want to combat online misinformation? Regulate the architecture of social media platforms, not their content](#), Australian Broadcasting Corporation, November 11, 2024.
- 28 Recommendation algorithms are properly conceived as a design feature since they are designed by platform software engineers. However, these designs can involve judgments and instructions about content. For example, an algorithm could be instructed to recommend user posts which have gathered the largest number of comments and shares (a content-neutral instruction) and/or it could be instructed to elevate posts that have to do with sports (a content-dependent instruction). This dual nature of recommendation algorithms makes them difficult to regulate, especially in jurisdictions like the US where most speech-related restrictions by government are unconstitutional. Nevertheless—and despite blanket [arguments](#) to the contrary—there are ways to regulate algorithms’ content-neutral elements without treading into content-based regulation. For example, the EU’s DSA stipulates that platforms must not present ads based on profiling using user personal data when providers are “aware with reasonable certainty” that the user is a minor (DSA, Article 28(2)). This provision regulates advertisement recommendation algorithms by restricting those system’s reliance on minors’ personal data. This is a content-neutral restriction. By contrast, Singapore’s OSA provides that “regulated services” must not target children with “harmful” or “inappropriate” content (Singapore Broadcasting Act, Code of Practice for Online Safety). The latter is a content-dependent requirement.
- 29 See, e.g., [research](#) produced by the Neely Center at the USC School of Business.
- 30 Nathaniel Lubin & Ravi Iyer, [How Tech Regulation Can Leverage Product Experimentation Results](#), Lawfare, July 2023.
- 31 Internal evidence, such as that showing the negative impact of Instagram usage on teenage girls, should be made available to inform design-based regulation. See, e.g., James Vincent, Instagram internal research: [“We make body image issues worse for one in three teen girls,”](#) The Verge, September 2021.
- 32 “Rather than retrofitting safeguards after an issue has occurred, Safety by Design focuses on the ways technology companies can minimise online threats by anticipating, detecting and eliminating online harms before they occur.” eSafety Commissioner, [Safety by Design](#), September 2024.
- 33 Marta Beltrán & Luis de Salvador, [Implications of Age Assurance on Privacy and Data Protection: A Systematic Threat Model](#), August 2024.
- 34 Sarah Forland, Nat Meysenburg & Erika Solis, [Age Verification: The Complicated Effort to Protect Youth Online](#), New America, April 2024.
- 35 Louisiana SOCIAL Act, 51:1752(A).
- 36 Under the EU’s DSA, online platforms must publish at least once every six months average monthly active users in EU for each online platform or online search engine (DSA, Article 24(2)). Under the California Addiction Act, operators must publicly disclose annually: The total number of minor users, the number of minors for whom verifiable parental consent has been obtained, and the number of minors with parental controls enabled (Section 27005).
- 37 E.g., Meta platforms, YouTube, TikTok, Discord, Twitch.
- 38 See EU DSA Articles 15, 24 and 42. In November 2024, the European Commission adopted an [Implementing Regulation](#) outlining the rules and templates for transparency reporting by providers of intermediary services under the Digital Services Act (DSA). The new law will ensure that all relevant providers give comparable information on their content moderation practices, from July 1, 2025.
- 39 Australia’s approach under the Basic Online Safety Expectations gives eSafety discretion to seek targeted questions, tailored to each service, about specific safety features services use across different surfaces of their services.
- 40 Under the Texas SCOPE Act, providers must disclose their algorithmic practices in clear and accessible language. This involves detailing how algorithms rank, filter, and present content to minors, as well as information on the types of personal data used in the algorithms (Section 509.056).
- 41 EU DSA, Article 27.
- 42 EU DSA, Article 26.
- 43 See EU DSA, Article 39(2) for list of minimum information requirements.
- 44 Texas SCOPE Act, Section 509.056.
- 45 Under Article 42(4), very large online platforms (VLOPs) and very large online search engines (VLOSEs) must submit themselves to independent audit and, after receipt of independent audit report, transmit to Digital Services Coordinator and European Commission and make publicly available the following: results of risk assessment; specific mitigation measures put in place; audit report; audit implementation report; information about consultations conducted, if applicable. Under Article 40, VLOPs and VLOSEs must, upon reasoned request from Digital Services Coordinator, provide access, within a reasonable period, to data to vetted researchers to study systemic risks in EU and assess risk mitigation measures.

- 46 Under the Texas SCOPE Act, providers are encouraged, but not required, to engage third-party auditors to review their content filtering technology, participate in industry partnerships, and conduct periodic independent audits to maintain compliance and efficiency in content moderation (Section 509.053(b)(2)(A)-(C)).
- 47 Some platforms, notably X (formerly Twitter) have made their algorithms open source.
- 48 See, Integrity Institute, [Ranking and Design Transparency](#) (draft), September 2021. See also, Uri Gal, [Want to combat online misinformation? Regulate the architecture of social media platforms, not their content](#), November 2024.
- 49 Singapore has a provision that could support such data access as well but it has not been operationalized.
- 50 Some scholars note that such provisions may face constitutional challenges in the US. But the question is as yet unsettled. See, e.g., [Platform Accountability and Transparency Act, S. 1876, 118th Cong. \(2023\)](#), 137 Harv. L. Rev. 2104, May 2024.
- 51 For a critique of current practices in transparency reporting, see ADL, [What's Wrong with Transparency Reporting \(and How to Fix It\)](#), October 2021.
- 52 E.g., Utah MPSM Act, Section 13-71-202(3); MD Kids Code Section 14-4605(4); EU DSA Article 14, UK Children's Code Standards 4 and 11; California Age-Appropriate Design Code Section 1798.99.31(a)(7).
- 53 EU DSA, Articles 16, 17, 20 and 22.
- 54 Under Australian Industry Codes and Standards, services are required to provide tools for users to make reports. The tools must be available "in", "on or through" the service, "easily accessible and easy to use" and "accompanied by clear instructions on how to use them"; services must also promptly acknowledge the report, notify the complainant of the outcome, and conduct a review if requested by the complainant (Industry Standards on class 1A and class 1B material, RES Standard, ss 28-30, and DIS Standard, ss.27-29). See also, UK OSA, Sections 20, 21, 31 and 32; Australia OSA RES Standard, s 16; Singapore OSA; Ireland OSC Section 16; Brazil Marco Civil Article 20; India IT Rules, Rule 3A.
- 55 EU DSA, Articles 11(1), 12(1), 11(2) and 12(2)
- 56 EU DSA, Articles 13(1) and 13(4)).
- 57 These regulations might require companies to follow specific processes, such as: (1) Identifying human rights risks related to content moderation, data processing, and AI-driven decisions, (2) engaging with stakeholders, including affected communities and civil society groups; and (3) implementing mitigation strategies and remedial mechanisms.
- 58 Tim Bernard, [Reading the Systemic Risk Assessments for Major Speech Platforms: Notes and Observations](#), Tech Policy Press, December 2024.
- 59 See, e.g., California Age-Appropriate Design Code Act, Section 1798.99.35(a).
- 60 Under New Zealand's HDCA, Netsafe is the administrative agency but is limited to mediating complaints and seeking voluntary takedowns (HDCA, Section 8(1)(c)). Enforcement via orders remains in the domain of the courts.
- 61 Fiji OSA, Section 23(b). Noncompliance with a court order is a criminal offense, and the corporation must be convicted before an employee is imprisoned. The relevant employee is one "in charge for the time being."
- 62 See, e.g., the Texas SCOPE Act, Section 509.152(b)-(c).
- 63 See, Texas HB 20 and Florida SB 7072.
- 64 [Projeto de Lei nº 592, de 2023](#).
- 65 *Moody v. Netchoice*, supra note 15.
- 66 [Projeto de Lei nº 2630, de 2020](#).
- 67 Also known as Canada Bill C-63.
- 68 ICCPR, Article 19. Three-part test.
- 69 See, e.g., EU DSA, Article 27.
- 70 Existing privacy policies tend to be vague and underinclusive. Regulators should enforce data privacy laws, such as the EU's General Data Protection Regulation (GDPR), more rigorously.
- 71 ADL, [What's Wrong with Transparency Reporting \(and How to Fix It\)](#), October 2021.
- 72 For an overview of available safeguards, see Laura Edelson, Inge Graef & Filippo Lancieri, [Access to data and algorithms: For an effective DMA and DSA implementation](#), Centre on regulation in Europe, March 2023.
- 73 See, e.g., the Neely Center [Design Code for Social Media](#) and Prosocial [Design Network](#).
- 74 Paul Barrett & Lily Warnke, [Enhancing the FTC's Consumer Protection Authority to Regulate Social Media Companies](#), NYU Stern Center for Business and Human Rights, February 2022.
- 75 For example, a default setting where zero data from users is collected would not allow the platform to track abuse.
- 76 eSafety Commissioner, [The Global Online Safety Regulators Network](#).
- 77 The Policy Press [tracker](#); Integrity Institute [legislative tracker](#); Digital Policy [Alert tracker](#).

NYU Stern Center for Business and Human Rights
Leonard N. Stern School of Business
44 West 4th Street, Suite 800
New York, NY 10012
+1 212-998-0261
bhr@stern.nyu.edu
bhr.stern.nyu.edu

© 2024 NYU Stern Center for Business and Human Rights
All rights reserved. This work is licensed under the
Creative Commons Attribution-NonCommercial 4.0
International License. To view a copy of the license,
visit <http://creativecommons.org/licenses/by-nc/4.0/>.



Center for Business
and Human Rights