



Reimagining Digital Regulation in India

A New Framework for
Online Safety



Contents

Executive Summary

- 01** Overview
- 02** Understanding Online Safety
- 03** Axes of Classification
- 04** Global Trends in Digital Service Regulation
- 05** Considerations in Designing an Online Safety Framework
- 06** Our Recommendations
- 07** Open Questions

Appendices

A Note to the Readers

Each chapter explores a distinct theme in online safety and can be read independently. Feel free to read the report from cover to cover or jump straight to the topics that interest you.

Acknowledgements

This research would not have been possible without **Google's** funding support. We extend our sincere gratitude for their commitment to fostering informed dialogue and advancing critical discussions in the evolving landscape of digital services regulation.

We are also deeply grateful to the nearly 50 experts and stakeholders from industry, academia, regulatory and policy institutions, civil society organisations, and other relevant communities across the world who generously shared their time, insights, and perspectives through interviews and focus group discussions. Their contributions significantly enriched this research and informed our understanding of key issues relating to content regulation, platform design, intermediary liability, and online safety.

Authors & Contributors

Authors

Ujval Mohan

Paavi Kulshreshth

Vishakha Singh Deshwal

Manan Katyal

Research Directed and Edited by

Rohit Kumar

Deepro Guha

Other Contributors

Tithi Neogi

Jeremy McKey

Kaushik Thanugonda

Saswat Mishra

Yashna Arora

S Muqieth Hussain

Design Visuals

Zakia Rafiqi

Cover artwork created using generative AI

Disclosure

At The Quantum Hub (TQH), we are committed to transparency in all our work. We consistently disclose the funding sources of all commissioned research. When a project is supported by a funder, we clearly identify them and acknowledge their role in shaping the research scope, providing methodological inputs, and sharing feedback. TQH retains full editorial independence and takes sole responsibility for the final research output, including its analysis and recommendations.

All insights and analysis presented in this publication are solely attributable to TQH and the named authors.

This report was prepared between March 2024 and February 2026. Given the pace at which digital technologies and their regulation are evolving, some laws, policies or regulatory positions discussed in this report may have changed since the research was undertaken. The analysis and recommendations should therefore be read in the context of developments up to February 2026.

Note on AI use: AI may have been used to refine the writing in some sections of this report; however, all such content was subject to significant human review and oversight. Any errors or omissions are exclusively our own.

Suggested Citation: TQH Consulting. (2026, October). Reimagining Digital Regulation in India: A New Framework for Online Safety.

Copyright License: This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). You are free to share, copy, and redistribute the material in any medium or format, and to adapt, remix, transform, and build upon the material for any purpose, even commercially, provided you give appropriate credit to TQH and the authors, provide a link to the license, and indicate if changes were made. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Executive Summary

The substantive drafting of this report was completed by February 2026, reflecting regulatory developments considered up to that point. A few references to subsequent developments are included selectively to contextualize the analysis and illustrate the continued evolution of online safety regulation.

The emergence of online safety laws can be contextualized by tracing three distinct waves of technology legislation around the world, beginning with a safe harbour wave (focused on enabling e-commerce); followed by data protection (focused on personal data rights); and the current, more fragmented third wave which seeks to advance online safety while also addressing adjacent priorities such as digital market competition and cybersecurity.

While there is no universally agreed definition of “online safety”, it is broadly understood as protection of end-users from online harms, typically emanating from user-to-user interaction on digital services. Unlike earlier waves of tech regulation, this third phase lacks convergence in legislative design and approach, with jurisdictions diverging on which harms are prioritized and what obligations digital service providers (DSPs) are required to meet. A recurring feature across many new laws is the articulation of a **duty of care** requiring DSPs to take steps to mitigate service-related harms. This represents a shift from earlier regulatory frameworks that focused primarily on liability shields for third-party content, towards models that impose affirmative risk-mitigation responsibilities on service providers.

This does not, however, mean that safe harbour protections have been abandoned. **Rather, it reflects a move towards a layered regulatory design in which service-level obligations coexist with intermediary liability frameworks.** For India, this also calls for a broader conception of the entities subject to online safety regulation. The existing framework is rooted in the concept of an “intermediary”, reflecting a regulatory architecture developed when many digital services primarily acted as passive conduits to facilitate the transmission or hosting of third-party content. Today, however, services increasingly combine multiple functions and actively shape user experiences. **This report therefore uses the broader terms “digital service” and “digital service provider” (DSP)** to reflect the range of services that may warrant online safety obligations, rather than limiting its scope to entities classified as intermediaries under existing law.

To ensure that online safety laws are targeted to the correct set of DSPs (without under- or over-inclusion of DSPs), a **classification framework becomes critical**. In Chapter 3, we identify three primary ways of classifying DSPs. In practice, most online safety regimes draw on a combination of these approaches, reflecting the need to balance precision, proportionality, and enforceability in regulatory design.

- **Functionality-based:** Classifies DSPs based on their *function or end-use*, using this as a proxy to assess their impact and potential risk.

- **Size-based:** Uses DSPs' scale (e.g. user base, revenue) as a proxy for impact, and thus distinguishes “large” (or, significant) DSPs from others. In practice, this often takes the form of defined thresholds of user base, revenue, or similar metrics, with only those meeting the threshold subject to specific safety obligations.
- **Risk-based:** Classifies DSPs based on their exposure to, or the likelihood of facilitating, particular harms, irrespective of function or scale. *For example, applying child safety focused interventions to all DSPs that children are likely to access.* While this approach most directly aligns regulatory design with the underlying policy objective, it is also the most complex to operationalize in practice.

Jurisdictions sometimes also adopt **ad hoc approaches**, where classification is implicit in the application thresholds rather than a central feature of the law. Whether explicit or implicit, all regulatory regimes must eventually define which DSPs are within its scope and why. Over-simplified or inconsistent approaches risk legal challenges, overburdening low-risk players, and under-regulating high-risk ones. In *Chapter 4* of this report, we examine online safety laws across diverse jurisdictions - **Australia, Canada, European Union, Singapore, Thailand, United Kingdom, and the United States** - to provide an insight into different classification philosophies and the nature of related obligations.

An ideal framework must distinguish between DSPs based on risk and function, and ensure that safety mandates are targeted, proportionate, and legally sound. This forms the backbone of our approach to building a future-ready online safety framework for India.

Key Considerations and Recommendations

The considerations and recommendations set out in *Chapters 5 and 6* of this report are based on a comprehensive review of trust and safety literature, an in-depth analysis of global approaches to online safety regulation, and consultations with a wide range of stakeholders, ranging from regulatory and legal experts to civil society organizations and trusted flaggers from India and around the world.

Together, these chapters are structured around **four foundational questions** for designing an effective online safety framework: *Which risks warrant prioritization? Which entities should regulatory action be directed toward? What form should safety interventions take? And how should these be implemented to ensure proportionality and effectiveness?*

A. Which problems to prioritize?

We understand online safety as the protection of end-users from harm arising from digital environments, particularly through user interaction on online services, including interactions with AI-driven systems. Such harms are usually a product of or a combination of (online) **content**, (user) **conduct**, or (user-to-user) **contact**. While these harms may originate in user-generated content, conduct or contact, they can be enabled, exacerbated or mitigated by the design and architecture

of digital services. DSPs therefore have a duty to mitigate harms that are reasonably foreseeable within their environments.

Online harms can also vary significantly in severity, persistence, and scale. Some are episodic and can be addressed through individual remedies, such as content removal or account-level enforcement. Others manifest as recurring and/or structurally embedded phenomena that cannot be meaningfully addressed through case-by-case responses alone. These broader patterns require systemic interventions and, in some cases, coordinated multi-stakeholder action.

In this report, we describe such phenomenon-level challenges as “**Systemic Risks**”. These represent priority concerns for an online safety framework because they arise from structural features of digital ecosystems and can produce widespread or sustained harm. Yet few existing online safety laws articulate a principled method for identifying which risks warrant systemic treatment; in fact, priority online harms covered by many such laws are a result of ad hoc and politically expedient choices. To provide a litmus check to regulators and limit unpredictability for DSPs, we develop a **three-part test** to assess which risks qualify as systemic risks.

A systemic risk is one that satisfies **ALL** of the following tests:

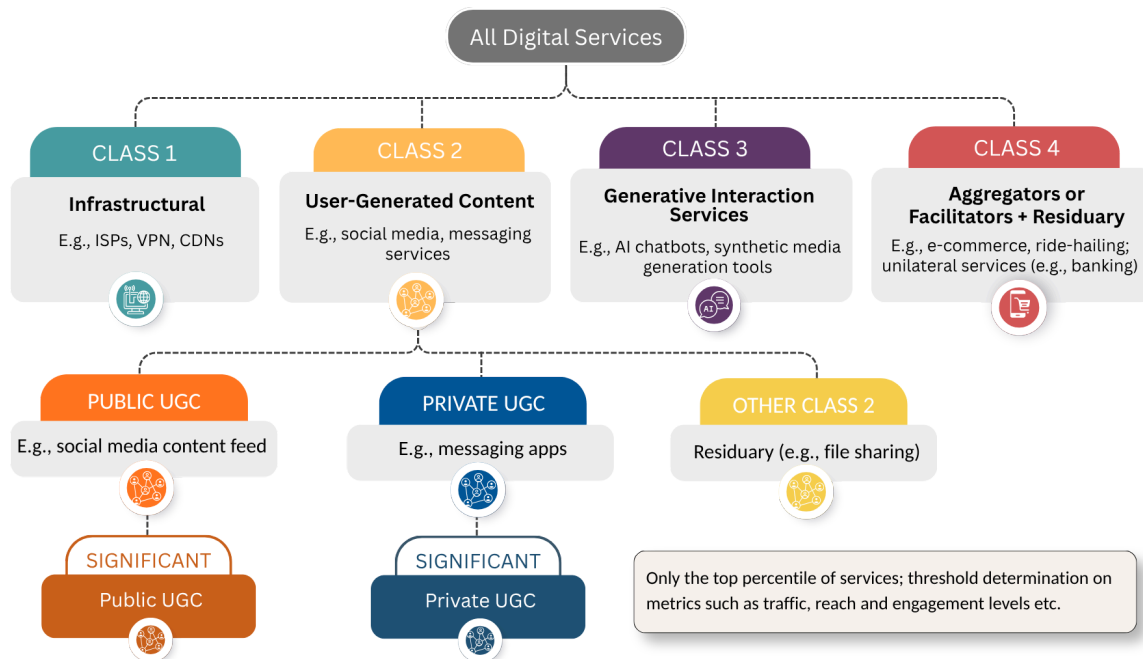
1. **Risk is widespread or systematic in its manifestation** - Filters out sporadic and isolated harms, thereby justifying the need for regulatory intervention
2. **Risk is caused or primarily exacerbated by digital services** - Establishes that risk is rooted in platform design or digital governance, thereby justifying the need for intervention under an online safety framework
3. **Vertical regulation is not adequate or effective** - Justifies the role of a general digital services regulator rather than a more narrow approach driven by sectoral regulators¹

B. Where should safety interventions be targeted?

Given our understanding of online safety harms as those emanating from user interaction, we argue that “**functionality**” of a digital service is the most reliable proxy for assessing the kinds of online safety risks the service is susceptible to. Functionality defines what users can and cannot do on the service. For instance, a private messaging service enables user-to-user communication, but typically restricts content discoverability unlike public feeds. Therefore, in our view, functionality should be the primary basis of differentiation between DSPs.

Building on this, we propose a three-tier classification of digital services wherein DSPs would self-classify based on the functionalities they offer, subject to appropriate regulatory oversight.

¹ We recognise that from a legal and policy hygiene perspective, we must lean towards narrower and *least-disruptive* options in policymaking, and therefore this third test creates an avenue for deliberate engagement on whether less-intrusive options from a sector-specific perspective are available that would achieve the same result.



Note: This is a technology-neutral, non-exclusive classification of **services** and **NOT service providers**. A single service provider may offer services across multiple categories.

At Tier I we cover all DSPs, but distinguish between four fundamentally different types of digital services, based on the functionalities they offer and the degree of control they exercise over content and user interaction. The purpose is not to subject all DSPs to uniform regulation, but to identify which categories are structurally relevant when addressing systemic risks. In particular, this tier isolates services that facilitate user interaction, while clearly *separating* other DSPs to ensure proportionate targeting of obligations.

- **Class 1 (Infrastructural)** services represent infrastructure and back-end enablers of the digital economy. They typically have negligible-to-no control over user-generated content and how users interact on digital services.
- **Class 2 (User-Generated Content or UGC)** represents services that actively facilitate user interaction, particularly through the creation, sharing, and dissemination of content. Examples include social media platforms, dating services, video and audio sharing platforms, messaging applications, and video-calling services, among others.
- **Class 3 (Generative Interaction Services)** represents services whose primary function is to generate novel, first-party output in direct interaction with a user, rather than facilitating an exchange between users. Common examples include AI chatbots such as ChatGPT, Gemini and Claude, companion applications, and services generating synthetic images, audio or video at a user's prompting. Although no second user is involved, they enable *quasi-user* interactions that can give rise to comparable online safety harms.

- **Class 4 (Aggregators or Facilitators)** represents services that typically help connect users to other, non-user-interaction focused services such as ride hailing apps, e-commerce platforms etc. While users may transact or engage with service providers through these platforms, they do not typically facilitate user-to-user interaction. Services in this category are often regulated vertically by respective sectoral laws or regulators, and are therefore generally excluded from additional online safety obligations. Importantly, Class 4 also houses the **residuary category** - containing services that do not neatly fit within the definitions of the pre-defined classes.

In our proposed framework, Class 2 and Class 3 together serve as the primary gateway for identifying services that warrant targeted regulatory attention under an online safety framework.

At Tier II, only Class 2 services are split into further sub-categories based on the varying nature of user interaction enabled by these services:

- **Public UGC Services** enable generation and dissemination of UGC to large, including unrestricted, audiences directly by users; these are the proverbial public squares of the internet. Common examples include *social media feeds, blogs, etc.*
- **Private UGC Services** enable private interaction within a limited number of users. Common examples include messaging platforms, email services, video calling services etc. Importantly, these are not services that allow public discoverability of UGC.
- **Other Class 2 Services** enable user-created content, conduct, or contact, but do not fall within the intended scope of either Public or Private UGC Services. These services may facilitate content creation or limited collaboration, yet are not designed primarily as social interaction environments. For example, document creation and editing tools that allow users to draft, store, and optionally collaborate on files would fall within this category. This residuary category is important to avoid force-fitting services into the Public or Private UGC classes, which could lead to disproportionate or unnecessary regulatory burdens. Where patterns of risk justify it, specific sub-classes within this category may be delineated in the future.

At Tier III, Public and Private UGC services are further classified based on size thresholds – into “Significant Public UGC” and “Significant Private UGC” services. Although we do not prescribe a specific metric, the intention is to create a class of the top percentile of DSPs in the market that, when taken together, represent most user experiences.

This structure creates sufficiently coherent categories of digital services, such that different regulatory expectations can be set for different “classes” without having to restrict legal requirements to the lowest common denominator, thus enabling regulators to introduce **proportionate** obligations that are well matched to the DSPs **risk profile** and **technology architecture**.

C. What should the interventions look like?

A functionality-based classification of digital services provides the foundation for translating the principle of duty of care into actionable, proportionate obligations. Building on this, we propose a three-level structure that progressively raises requirements as risks grow, incentivizing services to adopt appropriate countermeasures.

This approach offers an alternative to relying on tools such as blocking and takedowns, which were designed primarily to address liability for third-party content and can be blunt instruments for enforcing a duty of care. *In June 2026, for instance, India imposed a temporary nationwide ban on Telegram, acting on a National Testing Agency request alleging the app was being used to sell fake or leaked papers for a national medical entrance exam. While concerns regarding platform accountability are justified, this move relied on a disproportionate access-blocking mechanism, rather than uncovering the potential design failure that allowed the cheating rackets to operate in the first place. Digital platforms today are deeply embedded into how individuals and businesses function, and such shutdowns can create significant disruption and uncertainty. They also set a difficult precedent and risk undermining the digital ecosystem, without actually incentivizing platforms to investigate harms and deploy effective countermeasures.*

To address this, we propose three levels of obligations under our online safety framework: **Baseline, Proactive and Advanced.**

- **Baseline obligations establish the minimum required accountability owed by service providers.** These include core requirements such as informing users of terms of service; responsiveness to user grievances; and lawful content blocking. Baseline obligations apply to all classes of digital services.
- **At the next level, DSPs are required to take reasonable steps to prevent user harm through proactive safeguards.** These may include implementing accessible reporting mechanisms, limiting contact between users when necessary, adopting appropriate moderation practices, and for Generative Interaction Services, output safeguards to mitigate risks arising from the content generated by the service. Since Proactive obligations focus on safeguarding against user harm, these apply to Class 2 and Class 3 services alone. This helps insulate functionalities that involve neither interaction between users nor direct interaction with a content-generating service from legal obligations that are not relevant to them.
- **Advanced obligations** are designed to address identified systemic risks. Such risks often manifest as complex, structural phenomena that require calibrated and context-specific responses. Accordingly, obligations at this tier must be tailored both to the nature of the risk and the class of the digital service concerned. Different systemic risks may also warrant different combinations of measures.

For this reason, advanced obligations represent the highest tier of regulatory expectations

under the framework. We argue that by default, they should apply only to Significant service classes, namely Significant Public UGC and Significant Private UGC services, where the scale and reach of impact justify heightened intervention. However, their scope may be extended to other classes where risk-based justification exists.

To guide policymakers and enhance predictability for DSPs, we recommend specifying pre-defined “families” of advanced obligations that function as a structured toolkit for regulatory design. **We propose three such families: safety-by-design, transparency, and user empowerment.** This approach enables the regulator to respond to emerging systemic risks without reinventing the framework each time, while signalling to DSPs the broad nature and direction of potential compliance expectations.

The Indian government's response to WhatsApp's planned username feature illustrates the need for such families of obligations. In July 2026, WhatsApp announced plans to introduce usernames, allowing users to connect with each other without sharing their phone numbers. The government raised concerns about potential misuse, including unsolicited contact and impersonation scams, and directed WhatsApp to pause the rollout pending consultations. While these concerns are legitimate, a reactive approach of this kind risks setting a precedent where the government must weigh in on substantial product updates. For DSPs, this can mean delayed launches or rolling back features developed over months or years, at significant cost. Our proposed framework offers an alternative: DSPs can assess whether new features pose systemic risks and draw on predefined families of advanced obligations to identify and implement appropriate safeguards before launch. For the government, these families provide a consistent basis for assessing the adequacy of those safeguards.

D. How should this be implemented?

Given the rapidly evolving nature of the online harms landscape, an online safety law must incorporate predictable and periodic review mechanisms to remain responsive to changes in the digital economy and shifting risk profiles. Regular reviews embed iterativeness into policymaking, lowering the stakes of individual regulatory decisions and creating space for experimentation, course correction, and sustained dialogue. Over time, this fosters greater institutional confidence and encourages all stakeholders to invest consistently in advancing online safety.

- **Cyclical Assessments:** A recurring, time-bound review of systemic risks and advanced obligations, ensuring agility and transparency. Each review cycle enables the regulator to identify emerging risks, de-notify those that no longer meet the threshold, or retain the existing list when no revision is warranted.
- **Participatory Policymaking:** Multi-stakeholder consultations, with industry, civil society, experts, and the public, are built into every review stage. Voluntary risk assessments and transparency reports by DSPs inform regulatory decisions while avoiding unnecessary compliance burdens.

- **Permanent and Independent Regulator:** Implementing online safety laws effectively requires heightened technical and regulatory capacity to assess upcoming issues, organize extensive consultations and assimilate input, and enforce mandates in a rights-respecting manner. We believe this is best achieved when a permanent, well resourced and independent regulatory body is established with the objective of implementing the law.
- **Facilitative and Effective Enforcement:** The regulator ensures compliance through proportionate, risk-aligned penalties rather than blanket withdrawal of safe harbour protections. Where concerns arise, a structured review focuses on substantive outcomes, not mere formal alignment. DSPs may demonstrate that they do not materially contribute to the identified systemic risk, or that their safeguards, though different in form, are equally effective. This approach preserves accountability while enabling innovation and timely course correction.

01

Overview

This chapter aims to:

- Set context on the need for a systematic and targeted approach towards online safety
- Provide an overview of our first service classification report
- Highlight the need for developing a more comprehensive framework that reimagines digital service regulation

1.1. Need for Systematic and Targeted Digital Service Regulation

India's digital ecosystem has evolved rapidly, creating new opportunities as well as increasingly complex online safety challenges. The regulatory framework has also evolved in response, with successive amendments and rules seeking to address emerging issues.² However, this largely incremental approach has resulted in a framework that does not always account for the diversity of digital services, the different risks they pose, or the varying regulatory responses these risks may warrant. This points to the need for a more systematic reassessment of digital service regulation from an online safety perspective, moving **beyond a one-size-fits-all approach towards a more nuanced and targeted framework**.

A reassessment of India's online safety framework must begin with the existing legal architecture and its ability to incentivize digital service providers (DSPs) to adopt safety-oriented practices while serving Indian users. The primary legislation governing DSPs in India is the Information Technology Act, 2000 (IT Act). Its underlying architecture reflects the policy priorities of the time of its enactment, when the focus was on enabling digital adoption and the growth of digitally enabled commercial activity, including by providing intermediaries with safe harbour from liability for third-party content. While this framework played an important role in facilitating the growth of the digital ecosystem, the nature and diversity of digital services have since evolved considerably, raising questions about whether the existing architecture is sufficiently equipped to address contemporary online safety challenges.

Definition and Scope related Challenges: The IT Act is built around the concept of "intermediaries", reflecting an earlier phase of the internet and a digital ecosystem considerably less diverse than the one we see today. The proliferation of services ranging from social media and online gaming to OTT content streaming and cloud storage highlights the limitations of this broad categorization, which does not adequately account for the different functions digital services perform or provide a clear basis for tailoring online safety obligations to the risks they pose.³

Moreover, the traditional conception of an intermediary does not fully account for services that combine multiple functions or play a more direct role in shaping user experiences and interactions. These include social media services that curate and recommend content, AI-based chatbots that generate responses, and OTT platforms that distribute content directly to users. As digital services increasingly integrate such functions, a framework centred primarily on the intermediary concept struggles to address the distinct safety risks that arise from their design and functionality,

² For instance, [Information Technology Rules 2021](#) introduced under the [IT Act, 2000](#) and consequent amendments to these rules. Recent updates to India's IT Rules (2021) also recognize this challenge and introduced new categories of intermediaries such as Social Media Intermediaries (SMIs), Significant Social Media Intermediaries (SSMIs), and Online Gaming Intermediaries (OGIs) with tailored obligations.

³ Section 2(1)(w) of the IT Act, 2000 defines intermediary as "any entity that receives, stores or transmits an electronic message or provides an electronic service with respect to an electronic message." This overly broad definition covers all types of digital services making it difficult for the regulation to be targeted and ultimately necessitating ad hoc amendments to the IT Rules for additional obligations for specific categories or types of intermediaries.

including risks of addictive engagement, distorted perceptions of reality, and exposure to harmful content.⁴

In this context, there is a need to adopt a **functionality-focused approach** to develop a nuanced foundational structure for online safety regulation. As digital services pose different levels of safety risk, there is also a need for a classification framework that groups together services offering similar functionalities and presenting similar risks.

Inadequacies in the Policy Architecture: Beyond the challenges posed by the broad definition of “intermediaries”, the existing framework also lacks a systematic approach to identifying, prioritizing and addressing online safety risks. The introduction of the IT Rules, 2021 sought to strengthen the obligations of intermediaries and introduce additional requirements for certain categories of services. However, the framework continues to lack a clear basis for determining which online safety risks warrant regulatory intervention and what responses would be most appropriate.

For instance, the IT Act and Rules enumerate a range of harms that digital service providers (DSPs) must endeavour to address, but provide limited guidance on how these harms should be assessed and prioritized, how compliance should be evaluated, or how regulatory responses can be tailored to address them effectively.

Taken together, these limitations point to the need for a more coherent policy architecture that brings together service classification, risk assessment and proportionate regulatory intervention. In particular, such a framework would need to provide:

- A well thought out and rational way to classify different digital services and identify the class of DSPs that are most relevant to mitigate online safety risks;
- A framework to understand what online safety risks are, and which ones require regulatory interventions;
- Guidance on regulatory interventions that broaden policy thinking beyond content moderation and individual content-blocking orders to achieve safety objectives; and
- A robust, rights respecting, modern framework to identify non-compliances and address them in a constructive manner.

We believe that the next iteration of Indian digital service regulation requires solutions to some of these challenges that work well in the Indian context, while still fostering an attractive investment climate for digital innovators.

⁴ These risks have been recognized by the Australian eSafety Commissioner. See [Online Safety Advisory](https://www.esafety.gov.au/newsroom/blogs/ai-chatbots-and-companions-risks-to-children-and-young-people), <https://www.esafety.gov.au/newsroom/blogs/ai-chatbots-and-companions-risks-to-children-and-young-people>.

Other jurisdictions around the world, particularly the European Union⁵ and the United Kingdom⁶, but also Australia, have experimented with legislation designed to achieve some of these objectives, and while their approaches may not be directly transposable to India, there are important lessons to be learnt from their experiences that could inform Indian policy thinking on the subject.

Drawing on these lessons, this report brings together desk research, cross-jurisdictional analysis, expert and stakeholder interviews, and sustained internal deliberation and stress-testing to develop recommendations on the key elements that should form the next generation of India's online safety regulation.

1.2. Building on Our Earlier Research

Rationale for the first report

Our first report, [*A Framework for Intermediary Classification in India*](#), was published in December 2022 to examine how India's intermediary framework could better account for the diversity of digital services and enable more proportionate regulation. At the time, India had only just begun exploring sub-classifications within the broader category of "intermediaries". While the IT Rules, 2021 introduced specific categories, including social media intermediaries and significant social media intermediaries, the framework continued to subject services with substantially different functionalities and risk profiles to similar obligations. This created the potential for unnecessary compliance burdens on services for which certain obligations were not as relevant or proportionate to the risks they posed. For instance, the broad definition of "intermediary" covered infrastructure providers such as internet service providers, while the definition of "social media intermediary" could extend to discovery platforms such as dating and matrimony apps.

These limitations pointed to the need for a more systematic, functionality-based approach to classification, one that could better account for the diversity of digital services and their varied impacts on users and society. This became the subject of our first report. Our aim was to propose a flexible and accommodative mechanism that could be designed to include both evolving types of digital services and changing regulatory and policy objectives.

Highlights and key takeaways from the research

Published in December 2022, the research mapped the specific-purpose categories established under the IT Rules and found that, apart from certain differentiated requirements for gaming

⁵ EU DSA categorises intermediaries into conduit, caching, and hosting services, acknowledging their varying levels of control over content. Hosting services are then further divided into online platforms and online marketplaces, with marketplaces facing specific regulations due to risks associated with e-commerce. Further, platforms and search engines with over 45 million monthly active users in the EU are designated as Very Large Online Platforms (VLOPs) or Very Large Online Search Engines (VLOSEs), subjecting them to stricter regulatory obligations due to their significant impact on public discourse and the digital economy.

⁶ The United Kingdom's Online Safety Act (UK-OSA), mandates differing levels of accountability based on the risk to users, particularly children.

services and large social media platforms, the framework largely treated intermediaries alike despite significant differences in their functionality, risk profiles and potential systemic impact.

Recommendations

Drawing on international approaches, the report then proposed a new classification framework to categorise digital services based on specific factors such as functionality and user base. Its key recommendations were:

- **Functionality-based, tiered classification:** Classify digital services into tiers based on their functionality, recognizing that different types of services pose different regulatory risks. The report identified “communication” functionality as an important basis for distinguishing higher-risk services from lower-risk services.⁷
- **Risk-based and proportionate obligations:** Tailor regulatory obligations to match the level and type of risk posed by each service. Stricter obligations can be prescribed for higher-risk services, such as large social media platforms, and lighter obligations for lower-risk services, such as email and video communication services.
- **Balancing safety, innovation and economic growth:** Exempt lower-risk services, including infrastructure services such as caching and conduit services, as well as micro and small enterprises and start-ups, from certain regulatory obligations to avoid imposing disproportionate compliance burdens.
- **Participatory and consultative approach:** Encourage industry participation in developing and implementing the framework, and periodically review risk assessment metrics and classification thresholds to ensure the framework remains responsive to evolving services and regulatory objectives.

1.3. The Need for a More Comprehensive Framework

Since the publication of our first report, the digital ecosystem has continued to evolve, with new types of digital services challenging existing regulatory definitions. The growing **integration of AI** into digital services has raised new questions about liability, accountability and user safety, while the emergence of **decentralised platforms**⁸ presents additional challenges for regulatory oversight. More broadly, digital services have evolved beyond simply facilitating interactions between users, with their design, functionality and modes of operation increasingly shaping user experiences and potential harms. These developments, alongside a **global shift towards more**

⁷ Communication services are digital services whose primary functionality enables users to communicate or interact with one another, such as social media platforms and email or video communication services. Intermediaries that do not primarily facilitate communication, such as cloud service providers, search engines and online marketplaces, were treated as a distinct category, with different regulatory obligations from those applicable to communication-focused services.

⁸ Decentralised platforms are those that operate without a central authority, enabling peer-to-peer interactions and transactions among users. They leverage distributed ledger systems for transparency, security, and autonomy. By distributing control and decision-making across the network, these platforms reduce reliance on central authorities, allowing for greater user participation and resilience against censorship or content moderation.

comprehensive online safety regulation, highlight the limitations of a framework built around traditional categories of intermediaries.

In this context, we decided to undertake a **comprehensive reassessment of India's online safety framework**, with a view to developing a more holistic approach that combines a nuanced classification of digital services with clearly defined responsibilities for preventing and addressing the risks they pose. This is particularly important given India's digital economy ambitions and the need to foster an open, transparent and safe digital environment, while balancing user safety, innovation and regulatory efficiency.

Building on our earlier work, this report examines the rationale for different classification criteria and draws on international regulatory approaches, including experiences from the Global South, as well as insights from stakeholder consultations. It explores the shift from broad safe harbour protections towards greater emphasis on service accountability and user safety, and examines how factors such as platform design, content moderation, AI integration, business models and exposure to intersectional harms shape the risks posed by digital services. Based on this analysis, the report proposes a **new framework for classifying digital services and tailoring regulatory obligations to their functions and risk profiles**, alongside a duty-of-care approach to preventing and addressing harm. It also outlines how the framework could be implemented through a participatory and consultative process.

02

Introduction

This chapter seeks to:

- Situate contemporary online safety discourse within the broader history and evolution of technology regulation globally
- Trace the shift from the early internet era's safe harbour approach to the growing emphasis on digital service providers' (DSPs') duty of care
- Explain why a robust classification framework is essential to designing proportionate and effective online safety regulation

Digital regulation has evolved through distinct waves, each reflecting changing policy priorities as technologies, markets and their associated risks have developed.

The First Wave: At the beginning of the 21st century, regulation primarily aimed at enabling the **growth of e-commerce and building trust in digital transactions** to pave the way for digital economies. The premise of an enabling policy was to exempt internet “intermediaries” from liability for third-party content posted by users on their platforms. Given the volume of user-generated content hosted by these services, safe harbour protections generally hinged on the intermediary’s knowledge of specific unlawful content, subject to the applicable legal requirements. This approach was reflected in the **broad similarities across safe harbour frameworks** adopted by different jurisdictions. Prominent examples include *Section 79 of India’s IT Act* and *Section 230 of the US Communications Decency Act (CDA)*, both of which provide protections against liability for third-party content.

Second Wave: As digital markets matured, regulatory priorities expanded to address the collection, sharing and processing of personal information. This gave rise to the proliferation of data protection legislation across jurisdictions. Despite differences in scope and implementation, these laws **share common underlying principles and regulatory architectures**, including requirements governing the processing of personal data, individual rights and organizational accountability. Prominent examples include the *EU General Data Protection Regulation (GDPR)*, the *UK GDPR*, *California’s Consumer Privacy Act (CCPA)* and more recently, *India’s Digital Personal Data Protection Act (DPDP Act)*.

Third Wave: About half a decade from the enactment of the first prominent data protection legislation, we see a distinct effort⁹ to consider policies that advance online safety. These combat different online harms and risks such as **proliferation of disinformation, exposure to illegal content, children’s online safety** etc.¹⁰ Unlike the earlier waves, this has not produced a comparable degree of convergence in regulatory design. Responses have taken different forms, ranging from broad, cross-cutting legislation, such as the *EU’s Digital Services Act (DSA)* and the *UK and Australian Online Safety Acts*, to more targeted legislative proposals, such as the *US Kids Online Safety Act (KOSA)*. In India, similar policy objectives have been pursued through changes to the intermediary framework, including the *IT Rules, 2021*, and advisories issued by the Ministry of Electronics and Information Technology (MeitY).

Interestingly, unlike the first two “waves”, online safety regulation has seen far less convergence in legislative design and regulatory approaches. Jurisdictions differ both in the harms they prioritize and in the responsibilities they expect digital service providers (DSPs) to assume for addressing them.

In the above context, a central objective of this study has been to examine different policy design

⁹ Though we recognize that this is ongoing, and indeed, yet to commence in many parts of the world.

¹⁰ While this third “safety” wave has taken different forms in different parts of the world, it has also been accompanied by allied and more proactive policymaking around issues of antitrust and competition in digital markets, policies governing cybersecurity, and those that apply to emerging technologies like AI.

approaches with a view to identifying a framework that would be effective for India.

2.1. Defining What We Understand as “Online Safety”

Before examining how online safety frameworks should address harm, it is important to define what constitutes an “online safety” issue and distinguish it from other forms of digital harm. The growing integration of digital technologies into everyday life has created significant opportunities, but also new pathways through which harm can occur. Not all digital harms, however, are necessarily online safety concerns. Distinguishing between different types of harm, their underlying causes and the role digital services play in enabling or exacerbating them is essential to determining which issues warrant an online safety response.

Each of the following issues reflects a distinct category of harm:

A	Digital marketplaces engaging in self-preferencing or preferential ranking practices that disadvantage small and independent businesses.
B	Consumers purchasing goods through an online marketplace without adequate information regarding quality, and lacking effective redress mechanisms for defective or deficient products.
C	Users of a prominent social media platform being exposed to sustained volumes of disinformation and manipulative user-generated content.
D	The absence of effective mechanisms to secure the timely removal of non-consensual intimate imagery (NCII) from a video-sharing platform.
E	A data breach on an insurer’s online platform resulting in the public disclosure of citizens’ sensitive health and identity information.
F	An AI chatbot generating unsafe, manipulative, or harmful output in the course of direct interaction with a user.

While each of these issues results in harm experienced by users and warrants regulatory attention, they are analytically distinct and call for different regulatory responses.

In cases C, D and F, harm arises from platform design choices and governance mechanisms that enable, amplify, generate, or fail to adequately moderate or filter content presented to users. By contrast, in cases A, B, and E harm arises from anti-competitive conduct, deficiencies in consumer protection and redress frameworks, and failures in personal data security practices, respectively.

A further distinction lies in the **nature of the harm**. In C and D, harm is mediated through user-generated content (UGC) and may be characterised as user-to-user harm, even where platform design plays an amplifying role. In F, harm is caused by content generated by the service itself in direct response to a user, rather than by another user - but it replicates the experiential and harm characteristics of user-to-user interaction. In A, by contrast, harm stems from marketplace conduct such as self-preferencing, rather than from interactions between users.

These distinctions necessitate differentiated and often specialized regulatory responses. **Competition law is central in addressing A; consumer protection frameworks are better suited to B; and data protection regulation provides the appropriate safeguards in E.** Questions of AI model development, training, and accountability that lie behind a case such as F are, in principle, best addressed by a dedicated AI governance framework. However, the harm itself arises from direct interaction with the service and can be meaningfully addressed through the service's own design - output safeguards, disclosure, escalation - in the same way that platform design choices can help address the harms in C and D. It is this component of F - not the governance of the underlying AI model - that should fall within the scope of online safety regulation. The challenge, therefore, is to define online safety in a way that captures harms arising from digital service interactions without subsuming distinct regulatory concerns into a single framework.

This distinction provides the starting point for defining the scope of “online safety” in this study. A universally accepted definition or taxonomy of online safety has yet to emerge, and establishing one may be inherently difficult. We therefore adopt a practical approach to identifying the harms that should be addressed through a broad, general-scope online safety law. For the purposes of this study, we consider online safety issues to be those that:

- Cause **harms to end-users** of digital services (or indeed, the public at large);
- Emanate out of **user interactions** - including conversational interactions between a user and an AI system - and/or **design and features of digital services**; and
- **Cannot be adequately addressed through specialized regulatory frameworks** such as data protection, consumer protection, digital competition, or AI governance etc. (discussed in *Chapters 5 and 6*)

Further, though there can be different articulations of such harms, we find it useful to refer to the World Economic Forum’s (WEF)’s “Three Cs” framework¹¹ which clubs online harms into three broad categories:

- **Content:** Harms arising from exposure to harmful or problematic online material
- **Conduct:** Harms arising from harmful behaviours enabled by digital technologies
- **Contact:** Harms arising through interactions with others online

¹¹ World Economic Forum, “Creating a Common Language to Address Online Harm,” 4 August 2023, <https://www.weforum.org/stories/all/common-language-online-harm-typology/>

The production, circulation, and consumption of child sexual abuse material (CSAM), for instance, is a *content*-related harm. Unhealthy habituation tending to addictive behaviours is an example of a *conduct*-related harm. Meanwhile, *contact* harms are illustrated by practices like online grooming or harassment, where perpetrators initiate and maintain abusive relationships via digital services. Cyberbullying and cyberstalking are other examples of contact related or contact-adjacent harms.

Alongside such typologies, jurisdictions use outcome-focused approaches to define the harms their online safety frameworks seek to address. The UK Online Safety Act, for instance, places particular emphasis on protecting children and other users from specified online harms, while the EU Digital Services Act addresses risks to fundamental rights and broader societal interests (e.g., gender-based violence, public health). A detailed comparison of international online safety frameworks is provided in *Chapter 4*.

Despite differences in scope and design, these frameworks increasingly recognise that online safety cannot be addressed solely by requiring digital service providers (DSPs) to respond to unlawful content *after* it appears. As digital services have evolved from platforms primarily understood as intermediaries into services whose design, features and operation actively shape user experiences, regulatory approaches have increasingly placed responsibility on providers to anticipate and mitigate foreseeable harms. **This marks a shift from a predominantly reactive approach centred on safe harbour protections towards more proactive obligations, including duties of care.**

2.2. From Reactive to Proactive: The Fundamental Shift in Online Safety Frameworks

Early regulation of digital services centred on safe harbour protections, under which DSPs are held liable for illegal content only upon acquiring **actual knowledge** of its presence. This approach was grounded in the assumption that DSPs function primarily as **passive conduits** with limited visibility into, and control over, user activity. Imposing liability in such circumstances risked creating an unreasonable litigation burden and incentivizing the over-removal of lawful content. While this rationale continues to hold in relation to instances of illegal user-generated content, regulatory thinking has since evolved.

Today, there is growing recognition that modern digital services **play an active role in shaping user experience** and, by extension, exposure to online harms. Rising concerns about such harms, particularly those affecting children, have driven increasing political and public demand for greater platform accountability and, in some cases, regulatory interventions targeting access to digital services. For instance, Australia became the first country to introduce a ban on social media use for individuals under 16 in late 2024, followed by France, which introduced a similar restriction for users under 15 in early 2026.¹²

¹² Agence France-Presse. 2026. "France Passes Bill to Ban Social Media for Under-15s." *The Guardian*, January 27, 2026. <https://www.theguardian.com/world/2026/jan/27/france-social-media-ban-under-15s>.

As such harms persist, regulators increasingly view digital services not as neutral/passive conduits but as potential **enablers of online safety risks which they must proactively mitigate**. This shift has led to the emergence of a new species of digital regulation that recognized that DSPs owe a **duty of care (DoC)** to their users.

Importantly, unlike safe harbour protections, which can be broadly extended across the digital ecosystem, **DoC frameworks come with higher compliance requirements**, such as setting up grievance redressal mechanisms, proactively moderating content, and publishing transparency reports, among others. These obligations **cannot be applied uniformly across all services** as they may stifle innovation by over-burdening DSPs with compliance mandates or become ineffective in addressing service-specific risks.

This has important implications for the design of online safety frameworks. In the first and the second wave, it was possible and viable to devise policy designs that would apply **in a broad based manner** and potentially to **all digital services**. One can observe this in the way safe harbour applies broadly to all intermediaries, and the way data protection obligations cover every entity that processes personal data.

However, online safety obligations cannot apply uniformly to *all* DSPs. A **classification framework** is therefore essential to ensure that obligations are targeted, proportionate and effective. In our view, a well-designed classification framework is a foundational step towards developing a broader online safety framework. *Chapters 3 and 5* explore this in greater detail.

Designing such a framework requires difficult judgements about why and how digital services should be classified. These judgements must be grounded in coherent reasoning and produce outcomes that are legally defensible. A framework may be vulnerable to legal challenge if it treats similarly situated services differently without adequate justification, or imposes onerous obligations on low-risk services that are disproportionate to the risks they pose. Seen another way, these concerns reflect the broader principle against arbitrariness in policymaking that is increasingly relevant in regulation within modern democratic contexts.

This study therefore seeks to develop a classification philosophy and framework that is both modular and sufficiently granular to support an effective, proportionate online safety regime.

03

Axes of Classification

This chapter aims to:

- Identify the approaches to digital service classification adopted across policy and legal frameworks in India and other jurisdictions
- Provide a framework to assess the strengths and limitations of the three identified approaches: functionality-based, size-based and risk-based

Classification frameworks form a foundational element in the regulation of digital services worldwide. These frameworks help establish thresholds for identifying which DSPs must comply with specific regulations, and which DSPs fall below the regulatory threshold. *For example, in India social media intermediaries with more than 5 million registered users are subject to additional obligations.* Despite the diversity of digital services and their unique features, grouping similar services into defined classes is both essential for effective legal oversight, yet complex to achieve. **This chapter provides a theoretical basis to understand different approaches and their relative merits.**

Through our global comparative study, we examined the different approaches adopted across legal systems to identify the classes of digital services that are subject to regulatory mandates. Legislations across jurisdictions broadly adopt three different approaches to classification (or a mix of these) to identify high-risk classes of digital services:

1. Classification by **functionality**
2. Classification by **size**
3. Classification by **risk**, based on the propensity of a digital service to give rise to specific harms

Each of these approaches reflect distinct regulatory philosophy and objectives, with each offering unique advantages and facing specific limitations. Understanding these approaches is thus crucial while designing a framework that is both effective and adaptable to the evolving digital ecosystem.

In the following sections, we examine each of these approaches along with their advantages and limitations.

3.1. Classification by Functionality

Functionality-based classification frameworks bundle together digital services that provide similar services to users, such that the end-use at the user/consumer level is roughly comparable. ***The underlying assumption of these frameworks is that digital services that have similar functionality and features have comparable impact/risk propensity and must therefore be regulated similarly in one “class”.***

Traditionally, this approach distinguishes between services based on the degree of visibility, access, control and oversight DSPs exercise over content and user interactions. These factors, in turn, shape the nature and extent of the obligations imposed on them.

Such frameworks enable policymakers to conceptualize functionality-specific interventions. *For instance, this approach allows for the creation of a distinct category for infrastructural or conduit services - such as internet service providers and domain name registries - that attract lighter safety obligations, as is the case under the EU DSA (and earlier under the EU E-Commerce Directive).*¹³

¹³ Articles 11-14 of the EU DSA, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065>.

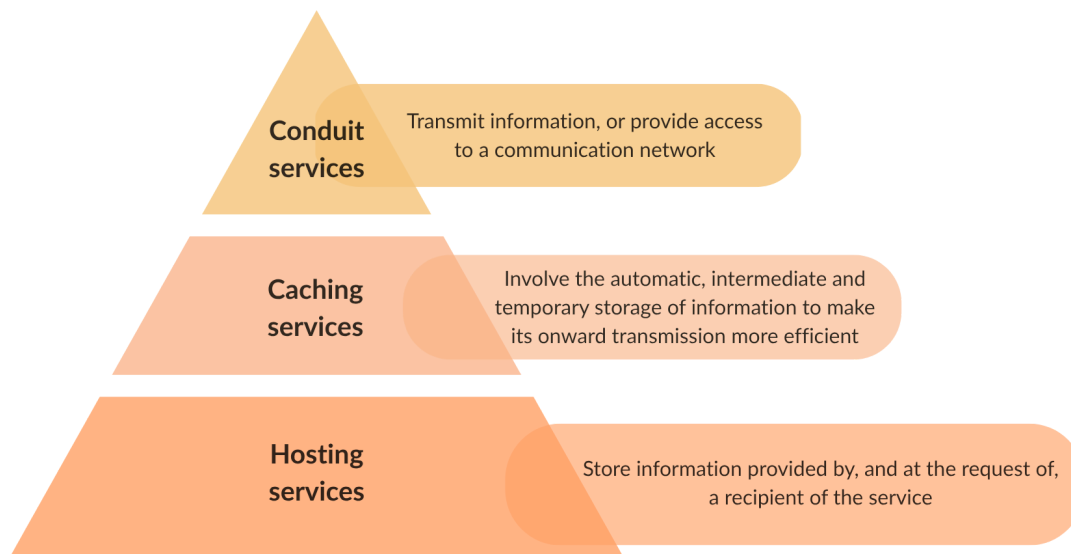


Figure: Functionality-based classification under the European E-Commerce Directive, 2000

Functionality based classification can take **different forms** depending on the intent and the structure of the underlying regulatory architecture. Here are two possible models:

A. Tiered functionality model

This model uses a tiered classification structure, with further sub-classifications and additional obligations introduced at each successive tier. *An example of this approach is the EU-DSA which first creates separate classes for caching, conduit and hosting services,¹⁴ and then creates further sub-categories within hosting services - making room for online platforms and online marketplaces within this class.*

Such tiered classification not only allows creation of differentiated responsibilities for different classes, it also allows creation of **cumulative obligations by layering baseline obligations with additional requirements for specific sub-classes**.

For instance, under the EU DSA, social media platforms such as Instagram and X are subject to the baseline obligations applicable to intermediary services, including transparency reporting and requirements to designate points of contact, subject to applicable exemptions. As hosting services, they must also provide mechanisms for users to notify them of illegal content. As online platforms, they face additional obligations, including enhanced transparency reporting, requirements relating to their interface design, and measures to protect users. Finally, platforms designated as Very Large Online Platforms (VLOPs) must comply with additional obligations, including assessing and mitigating systemic risks associated with their services.

¹⁴ Article 12, 13 and 14 of the EU E-commerce Directive, 2000, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32000L0031>.

This approach could be well suited for general scope legislations (such as the Indian IT Act) that have broad applicability and objective, as opposed to more targeted interventions (such as the US-KOSA) which can afford to directly address the DSPs that are relevant to its subject matter, without first classifying the rest of the digital economy.

B. Flat or Horizontal model

Functionality-based classification can also be approached in a different manner, with parallel or independent classes. In this model, the obligations for different classes are **not incremental but independent**, with obligations looking different between each independent class.¹⁵ Australia's Online Safety Act, for instance, classifies services into eight categories like search engine services, social media services, app distribution services, hosting services,¹⁶ etc.¹⁷ with a clear objective of targeting function-dependent obligations that are relevant for each type of service.¹⁸ Each class is subject to a different set of responsibilities and obligations through industry codes adopted for that class. This model also allows for adding new categories within the existing classes to account for new types of services, without necessarily having to rethink the base structure.

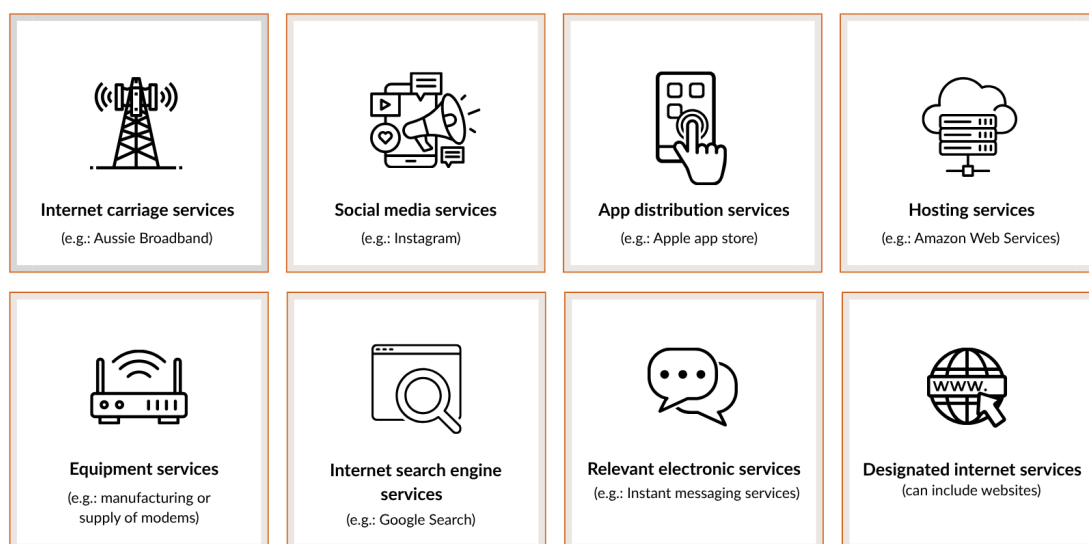


Figure: Functionality-based classification under the Australian Online Safety Act, 2021

While Australia's framework is a classic example of this approach, legislation designed to address specific issues, such as New York's SAFE for Kids Act, follows a similar logic. By identifying the

¹⁵ For instance, the industry codes for six different categories of digital services under Australia's OSA. See: Register of industry codes and industry standards for online safety, <https://www.esafety.gov.au/industry/codes/register-online-industry-codes-standards>.

¹⁶ Schedule 3 of the Australian Online Safety Code, <https://www.esafety.gov.au/sites/default/files/2023-06/Schedule-3%E2%80%933Hosting-Services-Online-Safety-Code-%28Class-1A-and-Class-1B-Material%29.pdf>: "hosting services" include two types of services: first-party hosting and third-party hosting services.

¹⁷ Section 134, Subdivision A, Division 7 of the Australian Online Safety Act, 2021, <https://www.legislation.gov.au/C2021A00076/latest/text>.

¹⁸ eSafety Commissioner. n.d. "Register of industry codes and industry standards for online safety." Australian Government. <https://www.esafety.gov.au/industry/codes/register-online-industry-codes-standards>.

particular categories of digital services targeted by the legislation, such laws enable a more narrowly defined and purpose-specific application of regulatory obligations.

Pros, Cons and Additional Considerations Functionality-based Classification	
Pros	Cons and Additional Considerations
- As explored in <i>Chapter 5</i>, functionality serves as a reasonably strong proxy for underlying risk potential and therefore can yield more effective outcomes in targeting obligations to high-risk DSPs. - Conceptually, this can be an intuitive approach, as it reflects how end-users experience and derive utility from a given service. It also keeps the end-user at the centre of policymaking, which we consider a valuable principle.	- For a functionality-based model to remain sustainable amid technological change, it is important to avoid defining service functionalities by reference to specific features or technologies available today. Instead, definitions should capture the underlying utility of a service from the end-user’s perspective, rather than the particular technologies through which that utility is delivered. <i>For instance, defining a category as “high-privacy closed-group messaging” rather than “end-to-end encrypted messaging” would allow it to accommodate different technologies that serve the same purpose.</i> Another way to future-proof the framework is to allow policymakers to periodically update and expand the list of recognized functionalities to account for emerging technologies and new service offerings. - This approach does not account for the size or scale of digital services, treating services with the same functionality alike regardless of their reach. Consequently, even where functionality-based classification is adopted, it is often complemented by size-based criteria to further differentiate obligations.

3.2. Classification by Size

Size-based classification frameworks differentiate digital services using metrics that indicate their size or serve as proxies for their user base and reach. *The underlying assumption is that “larger” digital service providers are a regulatory priority given their reach, market influence¹⁹ and potential to generate widespread or systemic harm.* Their scale can amplify the spread of harmful content through virality and network effects, increasing the likelihood that harm will affect large numbers of users.

Crucially, the “size” of a digital service can be measured in several ways, including: (i) total registered users; (ii) active users over a specified period; (iii) users as a proportion of the relevant population; (iv) market share among user base; and (v) revenue generated over a specified

¹⁹ Ofcom. 2024. “Why size and risk matter in our approach to online safety.” Ofcom. <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/why-size-and-risk-matter-in-our-approach-to-online-safety>.

period.²⁰ Frameworks may use either or any combination of these. For instance, under the Thai Royal Decree on digital platforms, “digital platform services” that exceed a specified revenue²¹ or average monthly user base²² threshold must make a detailed submission²³ providing detailed business information to the Electronic Transactions Development Agency (ETDA)²⁴ prior to commencing operations.²⁵

A size-based framework may be overlaid atop another framework to create more precise categories, potentially enabling scale-proportionate obligations and compliances. For instance, the Indian IT Rules 2021 first identify social media intermediaries based on their functionality, and then create a separate category of significant social media intermediaries (SSMIs) for those with more than 5 million registered users in India, subjecting them to additional due diligence requirements.²⁶ Similarly, the EU DSA classifies online platforms and search engines with more than 45 million average monthly active recipients in the EU as Very Large Online Platforms (VLOPs) or Very Large Online Search Engines (VLOSEs), respectively.²⁷ These high-reach services are subject to enhanced obligations, illustrating how size-based criteria can be used to differentiate regulatory requirements within a broader class of services, rather than imposing the same level of obligations on all services.

Pros, Cons and Additional Considerations Size-based Classification	
Pros	Cons and Additional Considerations
Enables regulators to prioritize oversight and resource allocation towards larger entities that likely impact more users and have a	Determining the size thresholds for operationalizing size based classifications is not easy. There is disagreement²⁸ about how best to measure a platform’s size – whether by user base, revenue, or both. Some argue that both metrics are important: user base reflects a platform’s real reach and the risks of scale,

²⁰ Goldman, Eric, and Jess Miers. 2021. “Regulating Internet Services by Size.” Competition Policy International, (May), 4-5. <https://competitionpolicyinternational.com/wp-content/uploads/2021/05/3-Regulating-Internet-Services-by-Size-By-Eric-Goldman-Jess-Miers.pdf>.

²¹ Gross income from operating a digital platform service in Thailand of more than THB 1.8 million per year if the Operator is a natural person or THB 50 million per year if the Operator is a legal entity

²² Average monthly active users (AMAU) in Thailand of more than 5,000

²³ The decree mandates both domestic and international digital platform services to notify the Electronic Transactions Development Agency (ETDA) of their operations and adhere to relevant obligations.

²⁴ ETDA is a government agency set up in 2011 under the supervision of the Thai Ministry of Digital Economy and Society. It plays a major role in promoting, supporting and developing electronic transactions (e-transactions) or online transactions in Thailand.

²⁵ Section 8 of the Thai Royal Decree on the Operation of Digital Platform Service Businesses that are Subject to Prior Notification, 2022, <https://www.etcha.or.th/getattachment/Regulator/DigitalPlatform/law/Clean-Royal-Decree-on-DP-Corrected-1.pdf.aspx>

²⁶ Rule 2(v) of the [IT Rules, 2021](#) – The 5 million threshold was specified through a MeitY notification, <https://www.meity.gov.in/writereaddata/files/Gazette%20Significant%20social%20media%20threshold.pdf> on 25.02.2021

²⁷ Article 33 of the [EU-DSA](#), https://www.eu-digital-services-act.com/Digital_Services_Act_Article_33.html.

²⁸ Bal, Meghna, and Shweta Venkatesan. 2024. “Platform Accountability and the Diffusion of Harm Online.” Esya Centre, no. 213 (November). <https://www.esyacentre.org/documents/2024/11/28/platform-accountability-and-the-diffusion-of-harm-online>. The working paper argues that content moderation measures by larger platforms often results in users migrating to smaller platforms that are lesser known and attract less visibility/scrutiny.

more pronounced effect on average user experiences.

- At the same time, it allows smaller players to operate with fewer burdens, promoting a more permissive regulatory environment that encourages growth of new and emerging services.

while revenue shows its market influence and resource base.²⁹ Relying on the user base can result in a regulation that over-emphasizes scale and potential of virality over market influence or vice-versa.

- While each indicator has strengths in relation to specific regulatory objectives, there is no consensus on a single measure that is universally preferred. Consequently, size-based thresholds can sometimes produce arbitrary outcomes that do not align with the underlying regulatory objective or accurately reflect a platform's influence or risk profile. This can lead to inconsistent regulatory treatment of otherwise similarly situated services.³⁰
- This approach **accords little regulatory attention to smaller service providers**, potentially allowing illegal or harmful content to go unchecked on these services. While some harms grow in severity as they attain higher-reach,³¹ others such as CSAM, cyber bullying, online grooming, doxxing, etc. can cause serious real-world harm irrespective of their virality. A size-based classification framework may therefore overlook smaller services that nonetheless facilitate or contribute to severe online harms, undermining the effectiveness of the broader regulatory framework. An overreliance on size-based criteria could, for instance, fail to adequately account for platforms such as 4chan, which have been reported to host and facilitate the spread of harmful or illegal content.

3.3. Classification by Risk

Risk-based classification frameworks rely primarily on the **potentiality of the underlying risk itself as a classifying parameter** to create different categories of digital services. Risk can be understood as the propensity of certain kinds of harm that digital services may expose individual users and/or society to.³² ***The underlying assumption of this approach is that digital services that pose similar risks can be regulated together as one “class”, irrespective of size or functionality.***

²⁹ Cho, Clare Y.; Zhu, Ling. 2023. “Defining and Regulating Online Platforms.” Congressional Research Service, R47662. (August). <https://www.congress.gov/crs-product/R47662>.

³⁰ For instance, Pinterest has around 95 million users worldwide which is close to Snapchat that has around 100 million users and even more than Netflix that has only around 74 million. In terms of posing safety risks, Pinterest would not pose the same safety risks in terms of harmful content that a social media platform or an OTT platform would.

³¹ Such as disinformation and online hate speech.

³² This understanding of risk is guided by our analysis of existing literature on adverse outcomes on online platforms; it relies on the “risk of harms” phrasing to define such outcomes, https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf

While all three approaches seek to identify DSPs that warrant greater regulatory scrutiny, functionality- and size-based classification rely on proxies to identify services that are likely to pose particular risks. By contrast, classification by risk eliminates the need to identify a good proxy for the targeted risk - it **directly classifies DSPs based on their actual susceptibility to a given risk**.

For instance, the UK's Online Safety Act imposes additional obligations on services in relation to content harmful to children, alongside broader duties concerning illegal content.³³ Similarly, Thailand's Royal Decree clubs digital platform services based on specified risk factors or their potential impact, including risks to financial and commercial security, and imposes additional obligations, such as risk assessment and the adoption of mitigation measures.³⁴

This approach is different from a functionality based classification in that the **resultant class in a risk based approach could include platforms with different functionalities**. For instance,

- *Where a risk based classification is designed to address the risk of CSAM spread, a video sharing platform with a large user base as well as a digital file storage service - which drastically differ from each other in terms of functionality - would potentially fall in the same category. Following this logic, the application of a 2023 advisory for protection of children from CSAM issued by the Indian National Human Rights Commission (NHRC) was extended to services such as Dropbox, Drive, and Cloudflare etc., along with social media.³⁵*
- *Similarly, in assessing disinformation risks, open-feed social media platforms with large user bases may be placed in the same risk category as closed, high-privacy messaging services. Although their utility and modes of propagating information differ significantly, a strictly risk-based framework could classify them together. However, the measures used to mitigate these risks must account for differences in the underlying technologies and their functional constraints, as interventions feasible on open-feed platforms may be impractical or inappropriate for private messaging services.*

Risk-based classification is also gaining prominence in the regulation of rapidly evolving technologies such as AI. Many jurisdictions have adopted frameworks that impose stricter obligations on high-risk AI systems while easing compliance requirements for low-risk applications, illustrating how regulatory obligations can be calibrated to the risks posed by different technologies and use cases.³⁶

³³ Chapter 4, Part 3 of UK's Online Safety Act, 2023.

³⁴ Section 18 (2) of the Thai Royal Decree on the Operation of Digital Platform Service Businesses that are Subject to Prior Notification, 2022, <https://www.etda.or.th/getattachment/Regulator/DigitalPlatform/Law/Clean-Royal-Decree-on-DP-Corrected-1.pdf.aspx>.

³⁵ See: National Human Rights Commission's Advisory for Protection of the Rights of Children against Production, Distribution and Consumption of Child Sexual Abuse Material, https://nhrc.nic.in/sites/default/files/Advisory%20on%20CSAM_Oct2023.pdf.

³⁶ Paul, Regine. 2023. "European artificial intelligence "trusted throughout the world": Risk-based regulation and the fashioning of a competitive common AI market." Wiley Online Library 18, no. 4 (December): 1065-1082. <https://doi.org/10.1111/rego.12563>.

Pros, Cons and Additional Considerations | Risk-based Classification

Pros	Cons and Additional Considerations
• Direct alignment with policy objectives: Arguably, classification by risk is the most direct approach in translating the policy objective into tangible form. Given that it does not rely on a proxy for the underlying risk, it theoretically bears the potential of identifying <i>all</i> relevant DSPs embodying the risk, without the problem of undercoverage.	• Difficulty keeping pace with evolving risks: During our consultations, several stakeholders cautioned against relying too heavily on risk as the sole basis for classification as risk profiles in digital technologies can change rapidly, making it difficult for a regulatory framework to keep pace. • Complexity of implementation: Unlike functionality- or size-based approaches, risk-based classification requires an assessment of the risks associated with individual services. Effective implementation therefore depends on robust risk-assessment methodologies, as well as adequate regulatory capacity and resources. • Challenges in assigning classification responsibility: If service providers are required to self-classify, they must invest significant effort in assessing their own risk profiles, while regulators must verify the accuracy of their assessments. If classification is undertaken by a regulator, the process likewise requires substantial expertise and resources. These challenges are particularly pronounced because risk-based classification depends on a substantive assessment of each service's risk profile, rather than on relatively objective or readily observable parameters such as functionality or user base.

Regulations focused on promoting online safety could theoretically opt for any of the above frameworks or a **combination/hybrid arrangement** to evolve sophisticated classification frameworks. *For instance, both Singapore and Thailand³⁷ classify on the basis of risk as well as size. Similarly, the EU-DSA has retained the original functionality-based classification created under the E-commerce Directive in addition to its size-based classification framework.*

3.4. Ad hoc models

While a systematic classification framework can provide greater regulatory certainty for businesses and users and enable more effective regulation, some jurisdictions, such as the United States, have not adopted any comprehensive framework for classifying digital services at the federal level.

³⁷ Section 18 (1) of the Thai Royal Decree on the Operation of Digital Platform Service Businesses that are Subject to Prior Notification, 2022, <https://www.etda.or.th/getattachment/Regulator/DigitalPlatform/law/Clean-Royal-Decree-on-DP-Corrected-1.pdf.aspx?lang=th-TH>.

Absence of a Digital Service Classification Framework: The U.S. Case Study

- No federal level classification framework exists for digital service platforms in U.S. law. Through the category of “interactive computer services”, Section 230 of the Communication Decency Act 1996 casts a wide net over virtually every entity or digital service that operates in the internet ecosystem — from hosting and caching services to social media platforms, search engines, and AI technology developers.³⁸ At a broad level, the strategy of regulation in the US is negative, not positive, i.e. the intent behind Section 230 is to **protect** digital services with a broad shield, and not to proactively impose obligations tied to online safety or allied objectives. As a result, federal level legislation has sidestepped the need to introduce a framework to classify different “interactive computer services”.
- In response to public support for regulation to further objectives such as privacy, child safety, and protection from censorship and lack of new legislation at the federal level, states in the US have passed their own legislation regulating digital services and providers. State-level legislation, such as New York State’s SAFE For Kids Act,³⁹ furnish examples of more stringent obligations. Signed into law in June 2024, the SAFE For Kids Act prohibits social media companies that employ “addictive algorithms” from sending minors late-night app notifications, among other obligations. Interestingly, the Act applies to “*social media companies where the addictive feed is a significant part of their platform, and not on online services that provide such feeds as ancillary features.*”⁴⁰
- Even in the absence of a uniformly applicable federal classification framework per se, **standalone legislations bear the hallmarks of a classification structure often relying on functionality, size, or risk based parameters** to determine their application. For example, laws enacted in Florida and Texas⁴¹ employ size-based classification schemes to determine the applicability of obligations that restrict when and how social media platforms can remove offensive content or “deplatform” users.

3.5. Inevitability of Classification and the Paradox in Avoiding It

We argue that while devising an appropriate classification framework may appear daunting, it is an essential exercise that cannot be avoided. As digital technologies become increasingly embedded in everyday life, regulation must evolve to reflect the changing digital ecosystem. This requires

³⁸ Brannon, Valerie C.; Holmes, Eric N. 2024. “Section 230: An Overview.” Congressional Research Service, R46751. (April). <https://www.congress.gov/crs-product/R46751>.

³⁹ Stop Addictive Feeds Exploitation (SAFE) for Kids Act, S7694A, New York State Senate, 2023, <https://www.nysenate.gov/legislation/bills/2023/S7694/amendment/A>.

⁴⁰ Stop Addictive Feeds Exploitation (SAFE) for Kids Act, S7694A, New York State Senate, 2023, <https://www.nysenate.gov/legislation/bills/2023/S7694/amendment/A>.

⁴¹ Florida Senate Bill 7072, <https://www.flsenate.gov/Committees/bills/summaries/2021/html/2345> and Texas House Bill 20, <https://capitol.texas.gov/BillLookup/History.aspx?LegSess=872&Bill=HB20>.

policymakers to establish clear criteria for determining *who exactly* must be the target of their interventions. Regulating without clear boundaries is sub-optimal, and when repeated across multiple instruments, only adds to uncertainty and complexity. At worst, such an approach would continually face legal challenges by DSPs that contest their classification into a higher-obligation class citing arbitrariness and insufficiently robust rationale of differentiation from their peer services.

This presents a paradox: while elaborate classification frameworks may appear unnecessarily complex, oversimplified approaches can create greater uncertainty and regulatory inefficiency. A well-designed framework, with clear and defensible distinctions between services, is therefore better placed to serve the interests of users, DSPs, innovators and policymakers alike.

04

Evaluating Diverse Duty of Care Approaches

[Status as of September 2025]

This chapter aims to:

- Map online safety frameworks and digital service classification models across jurisdictions to draw lessons for India
- Compare classification approaches, examining their design choices, implementation challenges and trade-offs

Reader Notes and Methodology

This chapter analyses a range of key jurisdictions, selected bearing in mind representation from:

- a. Major digital markets of the world;
- b. Jurisdictions that have attained a mature stage in digital service regulation; and
- c. Jurisdictions with developing digital regulation frameworks from the more relatable Asian contexts.

The cross-jurisdictional analysis thus covers perspectives from the EU, UK, USA, Canada, Singapore, and Thailand. The chapter contains our top-line findings and analysis spanning the themes of (1) **regulatory strategies and enforcement mechanisms**, (2) **digital service classification approaches**, and finally (3) **political economy considerations**.

For each jurisdiction, we have undertaken desk research and in many cases, interviews with relevant domain experts and stakeholders, to understand the underpinning regulatory philosophy behind the classification framework and the specific obligations applied to different classes of digital services. We also examined broader political, economic, and social conditions that shape the regulatory approach, by investigating how factors such as governance, market dynamics, political economy considerations and regulatory capacity influence the design and enforcement of classification systems.

Thematic Insights

A. Regulatory Strategies and Enforcement Mechanisms

Duty of Care as a Concept: The concept of *duty of care* has increasingly become a fundamental aspect of online safety frameworks around the world. Moving away from purely *reactive models of regulation* - where intervention by digital platforms was a legally binding requirement only after illegal content was lawfully flagged (that is to say, once the resultant harm had likely occurred), jurisdictions are now adopting *proactive measures* aimed at preventing harm before it materializes. This approach is centered around a DSP's duty of care that it owes to users to prevent their exposure to online harms. Countries are now embedding these principles into legislation, most notably under the EU-DSA, which requires platforms to proactively conduct risk assessments and implement transparent content moderation policies and mechanisms for user grievance redressal. Similarly, the UK-OSA requires regulated services to comply with Codes of Practice that articulate DoC principles into actionable steps. Even the US, which has historically avoided federal level proactive regulation of digital services, has seen state-level DoC legislations, and now the Kids Online Safety Act (KOSA) under consideration at the federal level. This trend represents the evolution of regulatory thought focusing on DSP's own responsibility to keep their services safe, regardless of liability arising from third-party content.

Divergent Approaches to DSP Obligations: Despite having embraced the philosophy of DoC, the

jurisdictions we studied have taken divergent approaches in translating the DoC into actionable legal mandates for DSPs. These range from process and governance-focused requirements - such as risk assessments and mitigation plans - to more prescriptive design interventions, like Codes of Practice that aim to prevent unwanted adult contact with minors on a platform. There are also interesting design features in regulatory regimes from Thailand and Singapore that are worth noting.

Increased Demand for Regulatory Capacity: The growing adoption of process-focused interventions and complex compliances requires regulators to continually monitor and assess platforms' risk management and overall compliance levels, creating a need for greater regulatory capacity. This demands higher capacity to handle the increased volume of work, more complex investigations, as well as a need for specialized resources, expertise and infrastructure.

Jurisdictions have taken several measures to address these capacity challenges. For instance, UK's OfCom has been additionally assigned as the entity in charge of enforcing the UK-OSA. Its duties and powers have been expanded significantly for enforcement of the UK-OSA, such as preparing and issuing codes of practice⁴² and giving guidance on what material is harmful. The EU has also attempted to solve for the increase in regulatory attention by instituting two new regulatory bodies: the European Board for Digital Services and the Digital Services Commission. In Australia too, we see collaborative and a near-consortium approach between sectoral regulators in addressing online safety.

Recurrent Themes in Policy Priorities: Despite differences in policy design, the objectives of various online safety frameworks show signs of convergence. In the West, there is broad political consensus on prioritising legislation addressing children's online safety - a focus central to the UK OSA, and a driver of legislative action in the US (at both federal and state levels) and in Australia and France, etc. Likewise, the impact of platform practices on democratic processes - including the role of disinformation in shaping public discourse - is a recurring concern. Codified as one of the four systemic risks that VLOPs must address under the EU DSA, this issue has also been addressed through standalone legislative proposals in Canada. Having said that, our research into *how* these priorities were enumerated returns no findings as to objective standards. We acknowledge that policy priorities in each jurisdiction are a result of political incentives and considerations, with no jurisdiction having publicly adopted a systematic process to determine the goalposts of their safety legislation.

B. Approaches to Digital Service Classification

Broadly, we identify two species of legislation - **general scope statutes** that look to provide a baseline framework to solve for *multiple* objectives (such as the EU-DSA, AU-OSA); and **issue-specific legislation** that are conceptualised to respond to one or few clear policy objectives(s) (such as UK-OSA, KOSA, SAFE for Kids Act etc. which focus on children's safety).

⁴² Section 41 of the UK OSA, 2023, <https://www.legislation.gov.uk/ukpga/2023/50>.

From a classification perspective, *general-scope laws* have tended to adopt *functionality-based classification designs*. Generally, we also find these to be more *comprehensive* in their reach - attempting to classify the broader digital economy into categories that can then be subject of appropriate substantive interventions. This is understandable since this approach allows the regulatory framework to pursue different objectives under a single framework. *For instance, this design can be used to set different levels of obligations for different classes, or target different functionalities with separate species of obligations.*

Alternatively, certain jurisdictions employ *specific-purpose* instruments with narrowly defined objectives that address specific risks, online harms or unique challenges within the digital ecosystem. Eg: UK-OSA, KOSA, SAFE for Kids Act and Singapore's Online Safety Act (SG-OSA). In each of these cases, we see that the classification system is typically a hybrid between risk, functionality, and size based models designed to ring-fence the class of DSPs that are most relevant to the *specific risk* sought to be addressed. As a result, if the main objective of the legislation were to be expanded to include different safety issues, the classification system may not yield satisfactory results.

Problem of Over-Broad Classes: A key learning from our focused research into classification approaches is the problem of over-breadth. While it appears tautological, a classification framework that does not adequately serve its *ration-d'etre* (which is to meaningfully differentiate between different DSPs) tends to be ineffective in aiding proportionate and successful outcomes in regulation. The problem is more evident at the extreme end of the spectrum where there is no categorisation at all, when policymakers need to re-invent the “applicability” wheel each time a new safety measure is sought to be introduced.

Though less immediately evident in other jurisdictions, classification frameworks that introduce over-broad classes are vulnerable to the same issue. Our analysis of EU-DSA's structure, with VLOPS as the centre piece, illuminates this issue. The definition of “Online Platforms” (and therefore VLOPS) is so broad that the resultant category of DSPs includes services that offer dramatically different functionalities - which no reasonable person would regard as similar to each other. *For instance, services like WhatsApp, Amazon, Google Maps, Bing, and X Videos are all found in the same class, and are thus meted out the same regulatory treatment.* As a result, several DSPs have challenged their classification as VLOPS,⁴³ and otherwise complain that what is expected of them by virtue of the designation is inappropriate, ineffective, or unnecessary. Therefore, while an overall functionality-overlaid-with-size approach has some clear advantages, policymakers in Indian settings would benefit from introducing granularity in key focus classes.

Convergence around Size Metrics: We observe convergence around the idea that larger platforms should be the starting point - if not the primary focus - of safety legislation. Virtually all models we studied either employ size-based classification criteria as the sole basis of applying

⁴³ Chee, Foo Yun. 2025. “Amazon asks court to scrap EU tech label, arguing it poses no systemic risks.” *Reuters*, June 12, 2025. <https://www.reuters.com/sustainability/boards-policy-regulation/amazon-asks-court-scrap-eu-tech-label-arguing-it-poses-no-systemic-risks-2025-06-12/>.

safety mandates, or introduce size as the key determinant in defining the high-risk or focus categories. While there are some clear and convincing reasons that validate this approach, we caution against overindexing on size as a primary risk determinant (*explored in detail in Chapter 5*), which risks leaving smaller DSPs that could contribute to the overall risk paradigm outside the regulatory radar.

C. Political Economy Considerations

Frameworks across jurisdictions reflect varying political and economic motivations. For instance, the DSA's functionality-based classification reflects the EU's political goals of harmonizing digital safety standards within the EU while providing for oversight on the largest players with substantial influence. Similar to this, other jurisdictions also approach digital service regulation in ways that adapt to their political and economic environments.

For example, the US retains a largely permissive stance with Section 230 of the Communications Decency Act (CDA), prioritizing protection for platforms from liability over regulatory intervention, signifying its decades-long, pro-industry *laissez faire* approach. This has been informed by a political economy that favors free-market principles and minimal state interference, while allowing tech players significant operational and innovative flexibility, even as specific state-level legislation begins to impose targeted regulations around privacy and child safety. Crucially, seen from an international geopolitical lens, the US - being the home of the largest DSPs today - has historically played a role in advocating for friendlier regulatory climates for these DSPs in foreign markets.⁴⁴

On the contrary, in Singapore, where political stability and regulated digital spaces are prioritized, the OSA gives substantial discretion to regulators like the Infocomm Media Development Authority (IMDA)⁴⁵. Singapore's outcome-based approach, influenced by the political economy's focus on efficient governance, predefines harmful content categories and promotes strict preventive obligations tailored to the country's specific online risks. These approaches reflect the tension between stringent control of digital content and the flexibility needed for diverse platform functionalities. Moreover, self-regulation, seen in frameworks like Thailand's Royal Decree, reflects a strategy for jurisdictions with limited regulatory resources to oversee compliance effectively. The Decree allows platforms to self-classify based on metrics like size, risk, and impact, with government guidance on compliance obligations. This method recognizes Thailand's need to mitigate risks without overextending regulatory reach, aligning with a political economy that must balance regulation with digital innovation to attract foreign investments.

⁴⁴ Hwan Lee, Moon. 2025. "Digital Regulation Is No Longer Just Domestic Policy as Korea and US Clash Over New Law." Tech Policy Press, August 8, 2025.
<https://www.techpolicy.press/digital-regulation-is-no-longer-just-domestic-policy-as-korea-and-us-clash-over-new-law/>.

⁴⁵ The IMDA develops and regulates the infocomm and media sectors.



European Union

Overview

The Digital Services Act (EU-DSA, an EU law enacted in August 2022) is often seen as a landmark online safety legislation, and a benchmark against which other approaches are measured. The EU-DSA sought to modernize the regulatory framework governing risks emerging from digital services and importantly, to harmonize standards of online safety and platform accountability across EU member states. It supersedes and expands on the EU Electronic Commerce Directive 2000 (EU E-Com Directive)⁴⁶ by seeking to create a safer and more transparent online environment and addressing key issues such as illegal content, user safety, and platform accountability. The law is applicable to a broad spectrum of digital services, such as social media platforms, online marketplaces, search engines, cloud service providers, content delivery networks and internet service providers. Key obligations include periodic assessments of the “systemic risks” enumerated under the Act; submission, review, and binding implementation of

⁴⁶ European Commission. 2024. “Digital Services Act: Questions and Answers.” Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/faqs/digital-services-act-questions-and-answers>.

mitigation plans; disclosures on algorithmic functioning; transparency in content moderation decisions; and stricter controls on targeted advertising. The EU-DSA has also been the subject of public debate transcending beyond traditional tech policy considerations, and engaging geo-political dynamics between the EU and US (where most digital services impacted by the EU-DSA are based) given differences in their respective tech regulatory philosophies.

Classification Philosophy and Approach

The EU-DSA adopts a flagship classification approach that uses functionality and size as classification metrics, in a multi-tiered model. This system builds on the single tier classification under the EU E-Com Directive with the primary objective of identifying a class of **very large digital services** that are mandated to manage stipulated systemic risks emanating from their operations.

At the **primary level**, the EU-DSA retains the functionality based classification for all intermediaries, by demarcating **conduit, caching, and hosting services** from the EU E-Commerce Directive. This tier recognizes that DSPs have varying degrees of visibility, access, oversight and control over the content/information they transmit, process, store or host, and hence should be customised for each of these classes accordingly.⁴⁷

At the **second level**, the EU-DSA's design rightly seeks to differentiate between key functionalities of digital services having recognised (extensively in the recitals) the need for distinct regulatory treatment. However, the final articulation is complex, non-intuitive, and inconsistent, which leads the intended clean "tier" model to suffer from a degree of distortion.

- The EU-DSA introduces a sub-class of hosting services termed as "**online platforms**". This design feature is a consequence of regulatory recognition that "hosting" is a broad definition that undesirably includes various digital services which only "host" content as an ancillary function, rather than this function being the primary service afforded to users.
- **Marketplaces** or "*online platforms allowing consumers to conclude distance contracts with traders*" is recognised as a *type* of online platforms (therefore, being a subset of that class) which enable e-commerce, and hence warrant some additional considerations while assigning obligations under the EU-DSA. This was welcomed by consumer protection groups who have long advocated for stronger oversight of online marketplaces to protect buyers from fraudulent sellers and unsafe products.⁴⁸
- Finally, **online search engines** is yet another pseudo-class⁴⁹ within the EU-DSA. Unlike marketplaces, online search engines are a subset of the *larger universe of intermediaries*, rather than being a subclass of hosting services or online platform services.

⁴⁷ Recitals 43 and 44 of the E-Commerce Directive 2000, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32000L0031>.

⁴⁸ Turillazzi, A., Taddeo, M., Floridi, L., & Casolari, F. 2023. "The Digital Services Act: an Analysis of its Ethical, Legal, and Social Implications." *Law, Innovation and Technology*, 15(1), 83–106. <https://doi.org/10.1080/17579961.2023.2184136>

⁴⁹ Schmid, Gregor. 2022. "Digital Services Act – an overview." TaylorWessing. <https://www.taylorwessing.com/fr/insights-and-events/insights/2022/11/digital-services-act-ein-ueberblick>.

At the **tertiary and final level**, the EU-DSA achieves the central objective of creating the focus class of **“Very Large Online Platforms (VLOPs)”** and **“Very Large Online Search Engines (VLOSEs)”**. This level employs a pure-play size based classification approach with the threshold for the “very large” categories being 45 million monthly active users in the EU.⁵⁰ The EU-DSA’s recitals state a clear intention of assigning services having a larger user base with heightened regulatory obligations, commensurate to their *“larger reach and a greater impact in influencing how recipients of the service obtain information and communicate online”*.⁵¹

Obligations, Enforcement Mechanisms and Implementation

The EU-DSA introduces a host of obligations for different classes of DSPs with a clear focus on preventing online harm to users, and heightening user agency and ability to hold DSPs accountable. Significantly, the EU-DSA requires VLOPS and VLOSES to also counteract 4 pre-decided systemic risks stipulated within the text of the law. In order to achieve this, the EU-DSA relies on stipulating what can be understood as **“obligations of conduct”** where the primary legal obligation is for VLOPS and VLOSES to undertake specified procedural steps (such as risk assessment, independent audits etc) with a view to manage systemic risks. This is an important distinction from **“obligations of result”** which could include product-prescriptive requirements etc., where the online safety framework stipulates **directly** the *outcome or product feature* that DSPs need to achieve/adopt.

A key process mandated under the EU-DSA is the risk assessment and mitigation plan submission DSPs need to undertake, which upon the regulator’s review and revisions can be given binding effect.⁵² While intended to allow flexibility and innovation in how DSPs adopted safeguards, initial experience has revealed challenges in incentivising meaningful changes to make digital services safer. In the opinion of harsh critics, process-based enforcement mechanisms act as *ceremonial structures*⁵³ failing to yield effective outcomes.

Further, models that require mandatory review and approval of VLOPS’ mitigation plans also spotlight the need for immense regulatory capacity⁵⁴ to be able to consolidate, review and evaluate submissions made by regulated entities and ultimately ensure compliance both with the letter of the law, as well as with the larger regulatory objectives which is trickier in the absence of credible measurable metrics. To address this, the EU has constituted Digital Services Commissioners in each EU member state, which supervise the maintenance of transparency

⁵⁰ Article 33(1) of EU-DSA.

⁵¹ Recitals 41, 57 and 79 of EU-DSA.

⁵² Cobbe, Jennifer. 2020. “Algorithmic Censorship by Social Platforms: Power and Resistance.” Springer Nature Link. <https://link.springer.com/article/10.1007/s13347-020-00429-0>.

⁵³ Edelman, Lauren B. 1992. “Legal Ambiguity and Symbolic Structures: Organizational Mediation of Civil Rights Law.” American Journal of Sociology 97, no. 6 (May): 1531-1542. <https://www.journals.uchicago.edu/doi/abs/10.1086/229939>; See also: Jiang N, Han R. 2024. “The regulation of content moderation in the DSA: can we trust the due process?” Forensic Res Criminal Int 12, no. 4 (September): 231. <https://medcraveonline.com/FRCIJ/FRCIJ-12-00425.pdf>.

⁵⁴ Allen, Asha, and Ophélie Stockhem. 2022. “A Series on the EU Digital Services Act: Ensuring Effective Enforcement.” Centre for Democracy & Technology. <https://cdt.org/insights/a-series-on-the-eu-digital-services-act-ensuring-effective-enforcement/>.

reportings, risk assessments, mitigation strategies, etc., and the European Board for Digital Services, which ensures consistent compliance practices across the EU.⁵⁵ Given these structural demands, the DSA requires significant regulatory capacity at both national and EU levels.

Finally, the EU-DSA has also championed other secondary obligations such as compelling access of key data and information to researchers, to reduce information asymmetry in the occurrence and patterns of online harms. Another notable feature is the empowerment of “trusted flaggers” in content moderation by services, which helps strengthen the voice of stakeholders with greater local context to contribute to overall online safety outcomes.

Key Takeaways

It is pertinent to note that the DSA is a **supranational legislation**, hence its design is influenced by EU-specific compulsions that may not be suitable to be directly transposed to national legislation serving different realities. *For instance, the EU-DSA can perhaps side-step the question of defining categories of illegal or harmful content, since it presumes this intervention will be handled individually by member states.*

- **Strong focus on multi-tier functionality based classification:** Emerging frameworks could benefit from a clearly articulated functionality-based classification, that recognises that different types of digital services have distinct roles and levels of risk associated with online harms. A functionality-first multi-tier approach also helps classify the broader diverse economy into more comprehensive and cohesive categories, creating room for specific interventions to target different relevant classes. EU-DSA’s recitals sharply outline the rationale for differentiation amongst key functionalities, which we believe is a positive feature of the legislation and can guide the application of the EU-DSA optimally when grey-areas or edge-cases arise. This is ideal for general scope legislations, where the objective is not to focus merely on one or few risks, or types of DSPs.
- **VLOPS as an over-broad class:** Given that VLOPS employs a size based classification overlaid on an already broad class of “online platforms”, the resultant class includes vastly different services that a reasonable user would not liken to each other. This prevents regulators from being able to target more effective purpose specific regulatory mandates, given the push back such initiatives would receive from other members of the same class that (likely rightly) should not be subject to the same expectation. Indeed, this has already triggered litigation by various VLOPS claiming to be unduly notified as such, aside from general feedback of VLOPS being an insufficiently cohesive class. This issue has been explored more in *Chapters 5 and 6*.
- **Focus on large service providers:** This size-based design feature has drawn criticism for its disproportionate focus on larger players. Critics argue that by focusing primarily on size as a metric to delineate high-risk platforms, smaller and emerging platforms with similar risk

⁵⁵ European Commission. n.d. “The enforcement framework under the Digital Services Act” <https://digital-strategy.ec.europa.eu/en/policies/dsa-enforcement>.

profiles could evade the same level of regulatory scrutiny as their larger counterparts, benefiting from lower compliance costs.⁵⁶ This design could unintentionally foster an uneven playing field, where large platforms face significant obligations while smaller platforms, despite posing comparable risks, are not pushed towards safer practices. This is heightened by the EU-DSA's wholesale exemption accorded to small business DSPs from many of its safety oriented obligations. At the same time, this approach can even act as a barrier to and disincentive for growth for smaller platforms as scaling efforts under the DSA would trigger more burdensome obligations.⁵⁷

- **Need for high regulatory capacity:** Moreover, a non-prescriptive process-based regulatory approach can only work well if there is meaningful oversight and guidance from the regulator. The lack of sufficient guidance on actual implementation of the EU-DSA, specifically relating to conducting risk assessments and independent audits, can lead to inconsistent compliance and ineffective regulation.⁵⁸ Although the European Commission has adopted the Delegated Regulation on the Performance of DSA Audits (DRA)⁵⁹, it does not provide standard definitions, methodologies, or benchmarks for how audits should be conducted. Since process-based obligations are by-design devoid of prescriptive mandates, each regulated entity could interpret them in vastly different ways, often subject to their resources, and the absence of specific guidance could result in regulatory gaps. One must hedge against the broader exercise morphing into “superficial checkbox exercise” (*transparency-washing* or *ethics-washing*) that does not shed significant light into the state of the digital ecosystem similar to reporting experiences under the Indian IT Rules, 2021.⁶⁰

⁵⁶ Ofcom. 2024. “Why size and risk matter in our approach to online safety.” Ofcom. <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/why-size-and-risk-matter-in-our-approach-to-online-safety>.

⁵⁷ European Commission. 2020. Impact Assessment part 1, 348. <https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act>.

⁵⁸ Loi, Michele. 2023. “Making sense of the Digital Services Act: How to define platforms' systemic risks to democracy.” Algorithm Watch. <https://algorithmwatch.org/en/making-sense-of-the-digital-services-act/>.

⁵⁹ European Commission. 2023. “Delegated Regulation on Independent Audits under the Digital Services Act.” <https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act>.

⁶⁰ AI Now Institute. 2023. “Algorithmic Accountability: Moving Beyond Audits.” AI Now Institute. <https://ainowinstitute.org/publication/algorithmic-accountability>.



The United Kingdom

Overview

The Online Safety Act (UK-OSA) is a legislation aimed at enhancing the safety of users on digital platforms, with a particular focus on children and vulnerable individuals.⁶¹ Enacted in 2023, the Act imposes duties which require providers of regulated services to identify, mitigate and manage the risks of harm from content and activity that is illegal and harmful to children. It creates a new duty of care⁶² for online platforms to take proactive steps to prevent the dissemination of harmful and illegal content, especially child sexual assault material (CSAM) and implement robust content reporting and removal procedures.

To operationalise this, it identifies three categories of digital platforms: user-to-user services, search services and pornography platforms and further classifies them into three categories - Category 1, 2A and 2B - based on their user base and risk profile. At the time of writing this section, the substantive obligations are conceptualized within Codes of Practice that are under development by the designated regulator (OfCom) via extensive consultations.

⁶¹ Department of Science, Technology and Innovation. 2024. "Guidance - Online Safety Act: Explainer." UK Government. <https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer>.

⁶² Latham & Watkins. 2024. "UK Online Safety Act 2023: A primer on the new law for relevant service providers." Latham & Watkins (London) LLP. <https://www.lw.com/admin/upload/SiteAttachments/UK-Online-Safety-Act-2023.pdf>; See also: Thompson, Marcelo. 2016. "Beyond Gatekeeping: The Normative Responsibility of Internet Intermediaries." *Vanderbilt Journal of Entertainment and Technology Law* 18, no. 4. <https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1159&context=jetlaw>.

Classification Philosophy and Approach

The UK's classification framework is grounded in a functionality-based approach, provided under the Electronic Commerce (EC Directive) Regulations 2002.⁶³ These regulations mirror the EU's E-Commerce Directive (EU-ECD) by distinguishing between DSPs based on their functions: conduit, caching, and hosting services.

Building on this foundation, the UK-OSA introduces a parallel classification regime for DSPs, with a particular emphasis on safeguarding children.⁶⁴ UK-OSA's obligations are only applicable to **user-to-user services**⁶⁵ (such as communications platforms, video-sharing platforms, online gaming platforms, etc.), **search services**⁶⁶ (any internet service that includes a search engine) and **other regulated services**⁶⁷ (digital services providing pornographic content). Each of these classes is subject to baseline duties of care defined under the legislation, emphasizing the protection of children and minimizing the risk of harmful and illegal content.

In addition to this, the UK-OSA introduces a secondary level of classification that categorizes these services into Category 1, 2A and 2B, which carry an additional scheme of obligations.

- Category 1 platforms are user-to-user services that meet certain thresholds (i.e. large and high risk user-to-user services)
- Category 2A platforms are search services and combined services⁶⁸ that meet certain thresholds (i.e. large search services)
- Category 2B are user-to-user services that are smaller but still in scope (i.e. smaller but still high risk).

Here, two factors influence the classification design simultaneously: the **size** of the platform and a subjective assessment of risk that emanates from the **functionality** of the platform. At the time of writing this section, these threshold requirements are in the process of being developed and will be notified by Ofcom through a delegated legislation. These standards will be determined by the number of users of the digital service, its functionalities, and any other factors relating to the service that the Secretary of State considers relevant. By adopting this nuanced classification, the UK-OSA distinguishes between the varying degrees of risk posed by different platforms, such that regulatory obligations are proportionate to the size *and* functionality of covered DSPs.

⁶³ Sections 17-19 of the Electronic Commerce (EC Directive) Regulations 2002, <https://www.legislation.gov.uk/uksi/2002/2013/contents>.

⁶⁴ Supra note 22.

⁶⁵ Section 3 of the UK-OSA, <https://www.legislation.gov.uk/ukpga/2023/50>.

⁶⁶ Section 3 of the UK-OSA.

⁶⁷ Section 4 of the UK-OSA.

⁶⁸ Combined services are regulated user-to-user services that include a public search engine feature. Section 4(7) of UK-OSA.

Obligations, Enforcement Mechanisms and Implementation

The UK-OSA assigns a duty of care to DSPs to proactively prevent harm by undertaking active identification, assessment and mitigation of risks related to illegal content.⁶⁹ Failure to discharge this DoC can result in monetary fines and in some cases, criminal penalties.⁷⁰

The UK-OSA's defining feature - **emphasis on protecting children from harmful and illegal content** - is manifested through its DoC obligations. It provides additional obligations for user-to-user services and search services likely to be accessed by children or if a significant number of users of the service are children.⁷¹ These duties vary with each category of services, with the strongest measures imposed on Category 1 services.

The obligations can overlap with UK-OSA's requirement of addressing **exposure to illegal content (for all users including adults)** since regulated user-to-user services are mandated to conduct separate "illegal content risk assessments" (aimed at harms arising from exposing all users to illegal content) in addition to "children's risk assessment" (aimed at harms arising from exposing children to illegal and harmful content).

This segregated approach to conducting risk assessments allows platforms to design targeted mitigation strategies that address specific risks separately.^{72,73} The UK-OSA requires platforms to take steps to manage and mitigate the identified risks⁷⁴, including by complying with more detailed measures provided in the codes of practice, expressly setting out the mitigation measures in their terms of service⁷⁵ and establishing a robust reporting and complaints mechanism, for users to report illegal content and content harmful to children.⁷⁶

To operationalise the duty of care measures under the UK-OSA, OfCom is currently developing two draft codes of practice at the time of writing this section: Illegal Content Code of Practice⁷⁷ and Children's Safety Codes⁷⁸, which will lay out governance obligations and recommended measures to mitigate risks. These codes are informed by extensive consultations with industry,

⁶⁹ Sections 1(2), 9, 10 and 12 of UK-OSA.

⁷⁰ For example, Sections 69 and 113 of UK-OSA.

⁷¹ For example, Sections 11 and 12 of UK-OSA.

⁷² For example, under the EU's Digital Services Act, there is only a single risk assessment that covers an umbrella of risks, including dissemination of illegal content, negative effects on the exercise of fundamental rights, negative effects on electoral processes and public safety and negative effects related to gender-based violence and the protection of minors.

⁷³ Ofcom. 2024. *Children's Risk Assessment Guidance*, Draft guidance for consultation. Step 2. Ofcom. <https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/284469-consultation-protecting-children-from-harms-online/associated-documents/a6-draft-childrens-risk-assessment-guidance-risk-profiles.pdf>.

⁷⁴ Sections 27 and 29 of UK-OSA.

⁷⁵ Sections 10 and 12 of UK-OSA.

⁷⁶ Sections 20 and 21 of UK-OSA.

⁷⁷ Ofcom. 2023. "Quick guide to illegal content codes of practice." Ofcom. <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/codes-of-practice>.

⁷⁸ Ofcom. 2024. "Quick guide to Protection of Children Codes." Ofcom. <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/quick-guide-to-childrens-safety-codes>.

civil society and regulators, ensuring that they reflect diverse perspectives and practical considerations.⁷⁹

Interestingly, some provisions of the Code borrow from corporate governance practices, emphasizing that direct accountability be placed on senior management to ensure compliance with the Act's requirements.⁸⁰ This governance approach aims to embed a culture of safety and responsibility at the highest levels of organizational leadership, and create internal incentives to prioritise proactive harm prevention and ongoing risk management.

Enforcement mechanisms under the UK-OSA are characterized by OfCom's oversight and review. OfCom is also empowered to open investigations into DSP's suspected non-compliances resulting in penalties, which further creates incentives for meaningful compliance by DSPs.

Key Takeaways

- The backbone of the OSA is its emphasis on the “statutory duty of care” approach, which acknowledges the importance of building incentives for safety-oriented thinking *within* DSPs, rather than such considerations purely being an *external* incentive of regulatory compliance.⁸¹ Incorporating governance-style compliance frameworks can also instill accountability in digital platforms without being overly prescriptive.
- Another noteworthy feature of the UK-OSA approach is the reliance on comprehensive Codes of Practice. This design feature allows the primary legislation to remain principle based and focused on desired outcomes, whereas the Codes of Practice can embody more substantive regulatory measures, which can then be periodically updated from time to time. OfCom's approach in conducting systematic and wide public consultations on the measures is also instructive.
- From a classification design perspective, UK-OSA is unique in having multi-dimensional classes that account for different size and risk profiles. We however note that this structure may be effective given the UK-OSA's limited objectives, and may not be scalable when the number of risks that a legislation targets increases.
- Another noteworthy aspect is the limited scope of the law - focused primarily on the dissemination of illegal content and harmful content to children - which presents both advantages and drawbacks. On the one hand, this targeted approach allows for a clear and manageable regulatory focus. However, such a narrow focus fails to address broader systemic risks such as the detrimental impact of harmful content for adults. The previous iterations of the OSA aimed to include “legal but harmful content” for adults, such as

⁷⁹ Ofcom. 2023. “Consultation: Protecting people from illegal harms online.” Ofcom. <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/protecting-people-from-illegal-content-online>.

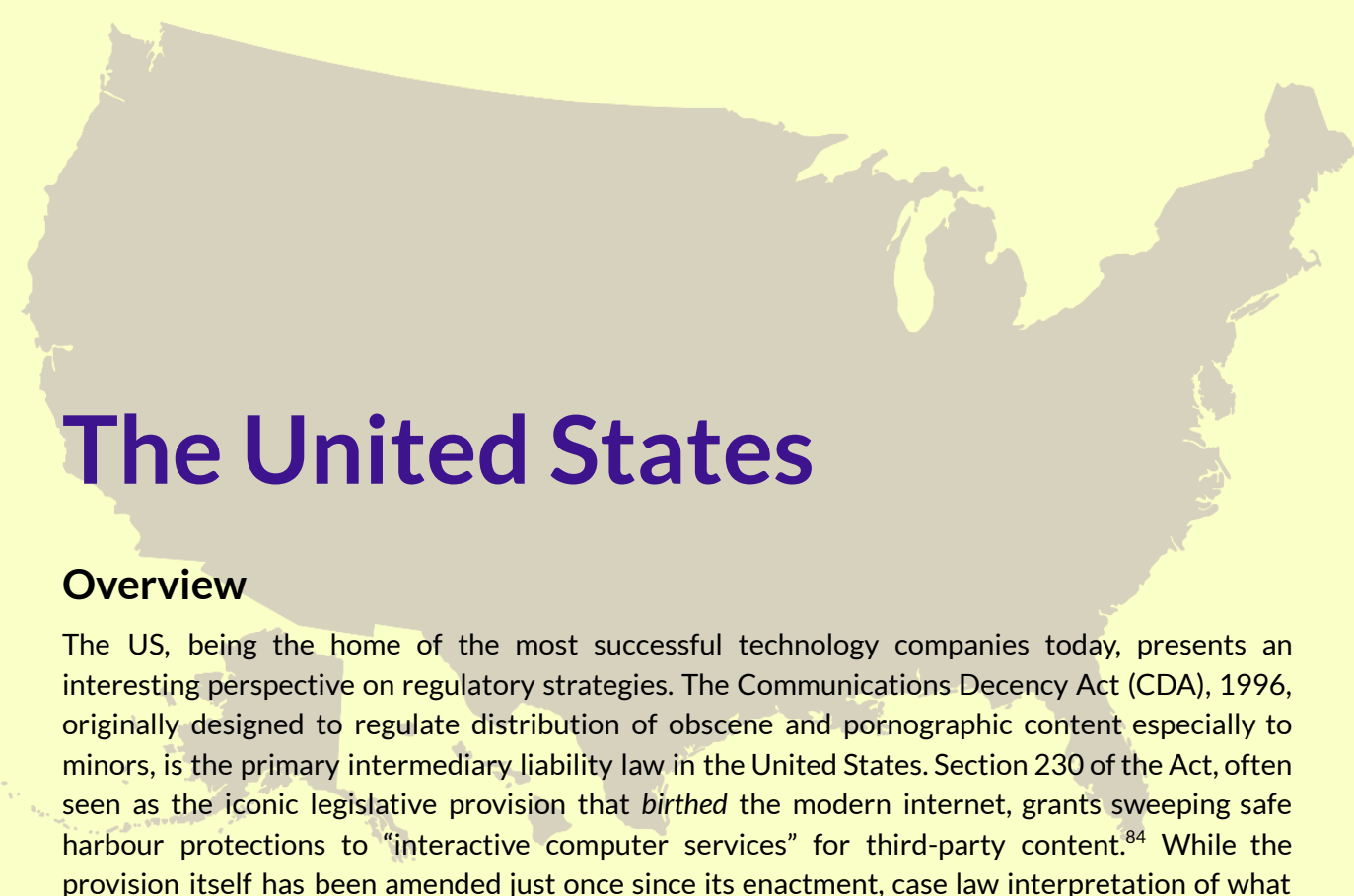
⁸⁰ Sections 72(7), 103 and 110 of UK-OSA.

⁸¹ Kira, Beatriz, and Laura S. Mendes. 2023. “A Primer on the UK Online Safety Act.” Verfassungsblog. <https://verfassungsblog.de/a-primer-on-the-uk-online-safety-act/>.

misogynistic abuse, promotion of self-injury, misinformation, etc.⁸² However, following widespread stakeholder criticism and concerns about implications for free speech, this category was dropped due to its vague definition and the risk of government overreach in regulating legitimate speech.⁸³

⁸² MacCarthy, Mark. 2022. "U.K. government purges "legal but harmful" provisions from its revised Online Safety Bill." Brookings Institute.
<https://www.brookings.edu/articles/u-k-government-purges-legal-but-harmful-provisions-from-its-revised-online-safety-bill/>.

⁸³ Harbinja, Edina. 2021. "The UK's Online Safety Bill: Safe, Harmful, Unworkable?" Verfassungsblog.
<https://verfassungsblog.de/uk-osb/>.



The United States

Overview

The US, being the home of the most successful technology companies today, presents an interesting perspective on regulatory strategies. The Communications Decency Act (CDA), 1996, originally designed to regulate distribution of obscene and pornographic content especially to minors, is the primary intermediary liability law in the United States. Section 230 of the Act, often seen as the iconic legislative provision that *birthed* the modern internet, grants sweeping safe harbour protections to “interactive computer services” for third-party content.⁸⁴ While the provision itself has been amended just once since its enactment, case law interpretation of what constitutes “third-party content” has evolved over the decades.⁸⁵

More than the other jurisdictions we studied, risk of litigation and sub-national policy development in the US have had a deeper impact on DSPs incentives to build in safety considerations into their offerings. Multiple US states have introduced their own legislation focused on narrow subjects such as privacy,⁸⁶ child safety,⁸⁷ and censorship⁸⁸. More recently, there have also been attempts at federal level policy interventions that aim to solve specific problems, such as the Take it Down Act.⁸⁹ In 2024, the US Senate passed the **Kids Online Safety Act (KOSA)** aimed at establishing service provider accountability to protect children from online harms; however, it has yet to become law.

⁸⁴ The key provision of section 230 is that “No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.”

⁸⁵ U.S. Library of Congress, Congressional Research Service, *Section 230: An Overview*, by Valerie C. Brannon and Eric N. Holmes, R46751 (2024)

⁸⁶ Bellamy, Fredric D., and Ashley N. Fernandez. 2023. “A new era of privacy laws takes shape in the United States.” Reuters, November 16, 2023. <https://www.reuters.com/legal/legalindustry/new-era-privacy-laws-takes-shape-united-states-2023-11-15/>.

⁸⁷ Lima-Strong, Cristiano. 2024. “States looking to 2024 to pass revised kids’ online safety bills.” The Washington Post, January 2, 2024. <https://www.washingtonpost.com/technology/2024/01/02/children-safety-social-media-california/>.

⁸⁸ Fung, Brian. 2023. “Supreme Court Will Look at New State Laws That Attempt to Control Social Media Content | CNN Politics.” CNN, September 29, 2023. <https://www.cnn.com/2023/09/29/politics/online-content-moderation-social-media-supreme-court-cases/index.html>.

⁸⁹ The Hindu Bureau. 2025. “What is ‘Take It Down Act’ signed by Donald Trump to combat ‘revenge porn’?” The Hindu, May 21, 2025. www.thehindu.com/news/international/what-is-take-it-down-act-donald-trump-to-combat-revenge-porn/article69596373.ece.

As it currently stands, digital service regulation in the US is not contained in a single legislative instrument (from which classification philosophies can be gleaned). Rather, the landscape varies both across different states as well as for different services, such as social media platforms and search engines, that may be regulated specifically in certain instances.

Classification Philosophy and Approach

No single classification framework exists for digital services in US law. Through the category of “interactive computer services,” Section 230, CDA casts a wide net over virtually every digital service that operates in the internet ecosystem. These range from caching and conduit services to social media platforms and search engines.⁹⁰ While there are limited exemptions from this principle enumerated in legislation,⁹¹ the intent behind Section 230 is to **protect** digital services with a broad shield, and not to proactively impose obligations tied to online safety or allied objectives. As a result, federal legislation currently sidesteps the need to introduce a framework to classify different “interactive computer services”.

However, if the Kids Online Safety Act (KOSA) is brought into force, it will potentially create a separate class of digital services, which cover any online platform, online video game, messaging application and video streaming service that is “*used, or is reasonably likely to be used, by a minor*”.⁹² This presents an example of a risk-focused classification design. These services are required to meet additional accountability and transparency obligations to prevent children from online harms.

Obligations, Enforcement Mechanisms and Implementation

Obligations under Section 230 are relatively light. These emphasize self-regulation, where platforms are largely responsible for setting and enforcing their own content moderation policies. This is because while the First Amendment prevents the government from infringing an individual’s right to free speech, it does not prevent private entities from specifying rules that govern digital services (and spaces) created by them. Some also argue that content policies and rules themselves should be treated as protected speech under the First Amendment.⁹³ Consequently, digital services are left to establish their own content policies and codes of conduct, which can often restrict speech beyond the powers of the First Amendment.⁹⁴ The incentive for

⁹⁰ U.S. Library of Congress, Congressional Research Service, *Section 230: An Overview*, by Valerie C. Brannon and Eric N. Holmes, R46751 (2024)

⁹¹ Section 230 contains statutory exemptions for criminal law, privacy law, intellectual property law, sex trafficking law, and state law.

⁹² Sec 101. (3) of Kids Online Safety Act, USA, <https://www.congress.gov/bill/119th-congress/senate-bill/1748/text>.

⁹³ David Inserra, “A Guide to Content Moderation for Policymakers,” Policy Analysis no. 974, Cato Institute, May 21, 2024, <https://www.cato.org/policy-analysis/guide-content-moderation-policymakers>.

⁹⁴ Spandana Singh, “Legal Frameworks that Govern Online Expression,” in *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content* (Washington, DC: New America’s Open Technology Institute, 2019), <https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/legal-frameworks-that-govern-online-expression/>

DSPs to build in safeguards comes from the risk of being held liable for enabling online harms in litigation brought by users or governments.

With concerns around Section 230's one-size-fits-all approach being out-of-touch with the realities of today's technology landscape,⁹⁵ there has emerged public support for stronger regulation. Lack of new legislation at the federal level has nudged states to pass their own legislation regulating digital services and providers. Many of these laws are designed to further objectives such as privacy, child safety, and protection from censorship (rather than *shielding* technology providers from liability).⁹⁶ However, constitutional challenges leave several of these state laws in limbo.⁹⁷

KOSA, which seeks to establish duty of care in federal digital service regulation, has also faced opposition in the United States' free speech-first political landscape. Opponents argue that it would encourage technology companies to censor online speech if they are made liable for user-generated content.⁹⁸ However, this position has also been contested since the Act seeks to hold digital services accountable for their "design features", and not user speech – *See the Box below*. Here, design features refer to any feature or component that "will encourage or increase the frequency, time spent, or activity of minors on the covered platform". The proposed law requires services to prevent and mitigate harm to minors by exercising "reasonable care in the creation and implementation of any design feature."⁹⁹ The proposed law also specifies design-related safeguards such as limiting user-minor (or stranger-minor) interaction, and recommender system transparency for minor accounts, including the option to opt out of personalized recommendation systems.¹⁰⁰

Under the robust safe harbour regime, the United States did not have an independent agency for monitoring platform governance. KOSA also refrains from introducing a new independent regulator, but prescribes enforcement via the Federal Trade Commission (FTC)¹⁰¹ and state

⁹⁵ Emily A. Vogels, "56% of Americans Support More Regulation of Major Technology Companies," Pew Research Center, July 20, 2021, <https://www.pewresearch.org/short-reads/2021/07/20/56-of-americans-support-more-regulation-of-major-technology-companies/>.

⁹⁶ For example, New York State's SAFE For Kids Act, discussed below, <https://www.nysenate.gov/legislation/bills/2023/S7694/amendment/A>.

⁹⁷ Kruzel, John. 2024. "US Supreme Court sidesteps dispute on state laws regulating social media." Reuters, July 2, 2024. <https://www.reuters.com/legal/us-supreme-court-set-decide-fate-texas-florida-social-media-laws-2024-07-01/>.

⁹⁸ Burke, Caitlin, Christina Lee, and Jennifer King. 2024. "KOSA's Path Forward: Distinguishing Design from Content to Maintain Free Speech Protections." Tech Policy Press, October 1, 2024. <https://www.techpolicy.press/kosas-path-forward-distinguishing-design-from-content-to-maintain-free-speech-protections/>.

⁹⁹ SEC. 102. Duty of care. KOSA 2025, <https://www.congress.gov/bill/119th-congress/senate-bill/1748/text>; It further clarifies that digital services are liable when it is reasonably foreseeable that the design features contribute to specific kinds of harms, including eating disorders, suicidal behaviour, depression, anxiety, addiction to the service, sexual exploitation or abuse, and physical violence, among others.

¹⁰⁰ SEC. 103 of KOSA 2025 [Safeguards for minors].

¹⁰¹ KOSA prescribes that violations be treated as unfair or deceptive act/practice under the Federal Trade Commission Act. Despite the fact that the Federal Communications Commission (FCC), back in 2016, had defended net neutrality rules that prohibit broadband services from engaging in content discrimination, its authority does not extend to digital services and service providers.

attorney generals.¹⁰² Both can initiate action to platforms liable to pay penalties or damages in violations.

Child safety interventions for different digital services have already been enacted at the state-level. Utah's App Store Accountability Act, for instance, requires app stores to verify users' age and ensure parental control to prevent minors from harmful content. New York State's SAFE For Kids Act, on the other hand, prohibits social media companies that employ "addictive algorithms" from sending minors late-night app notifications, among other obligations. Interestingly, the Act applies to "*social media companies where the addictive feed is a significant part of their platform, and not on online services that provide such feeds as ancillary features.*"¹⁰³

Even in the absence of a classification framework per se, these legislations bear the hallmarks of a classification structure often relying on functionality, size, or risk based parameters to determine their application. For example, laws enacted in Florida and Texas¹⁰⁴ – both of which faced Constitutional challenges at the time of writing – employ size-based classification schemes to determine the applicability of obligations that restrict when and how social media platforms can remove offensive content or "deplatform" users.

Understanding the Difference between Content and Design

It is important to note that, while section 230 provides broad immunity from "*third party content*", case law continues to evolve when it comes to defining precisely what qualifies as third-party content. The operative term in section 230 is "information content provider," defined as "*any person or entity that is responsible, in whole or in part, for the creation or development of information...*"¹⁰⁵

In *Fair Housing Council of San Fernando v. Roommates.com*, a landmark section 230 decision in 2008 that has shaped the prevailing understanding of "third party content," the Ninth Circuit Court of Appeals ruled that Roommates.com, a marketplace for listing rooms for rent, was responsible "in part" for statements posted by users on the platform that violated housing discrimination law in the U.S.¹⁰⁶ The court ruled that the user questionnaire, which prompted users to specify a preference for a prospective renter's gender and sexual orientation, contributed to discriminatory statements on the platform. Interestingly, however, the court also ruled that Roommates.com was not liable for statements published in an open comments section of the questionnaire, since in that instance the design of the questionnaire would not have contributed to the discriminatory content. As search engines and social media platforms

¹⁰² SEC. 109 of [KOSA 2025](#) [Enforcement].

¹⁰³ S7694A of [SAFE for Kids Act](#), New York State Senate.

¹⁰⁴ Florida SB 7072 and Texas HB 20.

¹⁰⁵ "Communications Decency Act," Pub. L. No. 104-104, 110 Stat. 133 (1996), 47 U.S.C. § 230.

¹⁰⁶ U.S. Library of Congress, Congressional Research Service, *Section 230: An Overview*, by Valerie C. Brannon and Eric N. Holmes, R46751 (2024)

integrate AI technology into their services and otherwise develop content designed to increase user engagement, the lines drawn in *San Fernando v. Roommates.com* around “third party content” may continue to shift as courts grapple with new questions about when “interactive computer services” would be responsible “in whole or in part” for the information on their platforms.

Key Takeaways

For policymakers seeking robust and self-contained models of intermediary liability legislation, the U.S. is perhaps unlikely to provide inspiration, as it lacks a detailed and consolidated classification framework nor does it offer pathways towards systematically assessing online safety issues, or conceptualising suitable and consistent countermeasures.

A fragmented regulatory approach has emerged through a number of federal and state-level legislatures operating simultaneously, creating uncertainty for companies and users. This over indexes on individual legislative instruments and allied political will to act on pressing issues instead of setting a harmonizing framework which allows ongoing dialogue between all concerned stakeholders in a predictable and consistent manner.

Finally, as fragmented interventions proliferate, the outcome could be an untenably inconsistent and complex regulatory environment for DSPs that prescribes different mandates across state borders. Not only does this serve to undermine the innovation friendly climate that the US prioritises, but could result in a compelling call for uniform federal standards (similar to how EU-GDPR and EU-DSA sought to standardize rules across member states).



Canada

Overview

Canada's online safety landscape features multiple regulatory bodies, numerous definitions, and divergent policy objectives across at least a half-dozen statutes, contributing to a fragmented approach to platform regulation.¹⁰⁷ In 2024, a broader Online Harms Bill was tabled in the Canadian Parliament, carving out separate responsibilities to prevent online harm for social media platforms. The bill, however, had been shelved as of January 2025.¹⁰⁸ DSPs in Canada also enjoy far less protection from liability and face more stringent obligations than they do in the United States. While a version of safe harbour exists in Canada, the jurisdiction does not have a single legislative vehicle that puts forth liabilities and obligations under a clear classification framework.

¹⁰⁷ Statutes in Canada with implications for platform regulation include the 2020 Personal Information Protection and Electronic Document Act (PIPEDA), the 2012 Copyright Modernization Act, the 2018 Election Modernization Act, the 2023 Online News Act, the 2023 Online Streaming Act, and two proposed omnibus bills, C-27 and C-63, that would address child safety and AI regulation.

¹⁰⁸ Lau, Mandy. 2025. "Canada's Online Harms Bill is Dead (Again): Three Questions to Consider for the Next Round." Tech Policy Press, April 28, 2025.
<https://www.techpolicy.press/canadas-online-harms-bill-is-dead-again-three-questions-to-consider-for-the-next-round/>.

Classification Philosophy and Approach

In lieu of a single federal instrument dealing with DSPs, Canada's legislative landscape mixes and matches different approaches. For example, the 2012 Copyright Modernization Act applies a functional classification scheme to its notice-and-takedown obligations for copyright infringement. Hosting and caching entities are among the functional categories defined in the statute and exempted from liability for infringement. In Canada, DSPs are also identified/defined under the Election Modernization Act (EMA) and Online News Act (ONA). The EMA, which seeks to regulate political advertising, defines "online platforms" as apps and sites whose operators, directly or indirectly, sell advertising space. ONA, on the other hand, creates a separate class of "digital news intermediaries". These intermediaries are defined as online communications platforms that make "news content produced by news outlets to persons in Canada."

The proposed 2024 Online Harms Act, or Bill-63, marks a shift towards broader platform accountability and transparency in Canada.¹⁰⁹ Social media platforms (that exceed a threshold of users¹¹⁰) are regulated entities under the proposed law, with additional obligations to prevent online harms.

Obligations, Enforcement Mechanisms and Implementation

In addition to the notice-and-notice¹¹¹ regime for copyright-infringing material and associated safe harbour protections delineated in the Copyright Modernization Act, DSPs also face obligations under a horizontal data protection law¹¹², and the abovementioned EMA and ONA. The EMA introduces transparency obligations for *online platforms* with respect to political advertising, requiring them to publish a "*registry of partisan advertising messages and election advertising messages*" in a specified window of time around elections. The ONA requires digital news intermediaries to bargain with and compensate Canadian news companies as a condition of hosting Canadian news content.

Finally, if enacted, the 2024 Online Harms Bill, roughly modeled on the UK-OSA, would layer an additional set of obligations on social media companies. These include obligations to build tools for users to flag harmful content, submit a digital safety plan, and take down CSAM content, intimate content shared without consent, and deepfakes.

In addition to the approach of addressing distinct harms through isolated statutes, a second key feature of Canada's regulatory landscape is the existence of strong regulatory entities invested

¹⁰⁹ Tremau T&S Research Team. 2024. "What does Canada's proposed Online Harms Act mean for your platform?" Tremau. <https://tremau.com/resources/what-does-canadas-proposed-online-harms-act-mean-for-your-platform/>.

¹¹⁰ Part I, Section 3(1) of *Online Harms Act, 2024*, <https://www.parl.ca/DocumentViewer/en/44-1/bill/C-63/first-reading> [Regulated Services].

¹¹¹ Canadian copyright law adopts a different approach from the US' notice-and-takedown, requiring intermediaries to forward notices of copyright infringement to their users instead of requiring them to take down the allegedly infringing content.

¹¹² Personal Information Protection and Electronic Documents Act (PIPEDA), 2000 (Canada), <https://laws-lois.justice.gc.ca/eng/acts/p-8.6/>.

with broad powers.¹¹³ Some have criticized¹¹⁴ the 2024 Online Harms Bill for a perceived heavy-handed and overly punitive approach that would endow a newly established Digital Safety Commission with powers to investigate complaints,¹¹⁵ hold hearings,¹¹⁶ and levy fines of up to 8% of gross global revenue (among the steepest in the world), if the bill is enacted.

Key Takeaways

Canada's approach to regulating intermediaries offers both positive lessons and cautionary tales for policymakers in other jurisdictions.

- In particular, one innovative feature of Canada's 2024 OHA - potentially of interest to policymakers - is its proposed "**Digital Safety Ombudsperson**." This position would complement the Digital Safety Commission's formal regulatory responsibilities by informally highlighting issues related to online safety and advocating for the public interest.
- However, stringent regulatory standards are often accompanied by **high compliance and cost burden** for service providers operating in Canada. In response to Canada's 2023 ONA, for example, Meta opted to **exit the Canadian news market rather than comply** with the statute's obligations to bargain with and compensate news providers. Mindful of the ONA and the cautionary tale of overregulation that its story embodies, many have criticized the proposed bill for a similarly heavy-handed approach that would impose excessive obligations, enforced by exorbitant fines, on platform companies.¹¹⁷ Policymakers, especially in growing digital economies, must carefully review such unintended impact of stringent and overly-prescriptive regulatory requirements.

¹¹³ Regulatory entities in Canada with purview over various aspects of platform regulation include the Privacy Commissioner of Canada, the Commissioner of Canada Elections, and the Canadian Radio-Television and Telecommunications Commission (CRTC).

¹¹⁴ Isai, Vjosa. 2024. "Canada Wants to Regulate Online Content. Critics Say It Goes Too Far." The New York Times - Breaking News, US News, April 9, 2024. <https://www.nytimes.com/2024/04/09/world/canada/canada-hate-speech.html>.

¹¹⁵ This can include summoning people to give evidence under oath, asking for documents, and accepting evidence even if it would not be admissible in a court of law.

¹¹⁶ In cases related to NCDII, CSAM, or any other issue about a platform's compliance with the Act.

¹¹⁷ Keller, Daphne. 2023. "Five Big Problems with Canada's Proposed Regulatory Framework for 'Harmful Online Content'." Tech Policy Press, October 24, 2023. <https://www.techpolicy.press/five-big-problems-with-canadas-proposed-regulatory-framework-for-harmful-online-content/>.



Australia

Overview

Australia's Online Safety Act, 2021 is the primary legislation governing online safety and digital service provider responsibility in the country.¹¹⁸ It specifies **eight categories of digital service providers** across the online industry to prevent harm to citizens, and has three unique features:

- a. Provides for an industry-specific code-making process;
- b. Requires voluntary safety expectations to be adopted by digital services; and,
- c. Introduces targeted schemes that focus either on particular types of harmful content or on specific obligations for covered services.

Prior to the introduction of the AU-OSA, Australia offered very limited safe harbour protections under the Copyright Act 1968, only to carriage/internet service providers. While the AU-OSA introduces proactive safety requirements for digital services, it also follows the principle of actual knowledge when assigning responsibility. A service is considered to have “knowledge” of harmful content when complaints are raised by affected users or, in some cases, by the regulator.

AU-OSA is enforced by Australia's e-Safety Commissioner¹¹⁹ which primarily focuses on regulating online harms and promoting online safety. Apart from the e-Safety Commissioner, three key regulators oversee different aspects of Australia's digital landscape. The Office of the Australian Information Commissioner (OAIC)¹²⁰ focuses on privacy and information access rights, Australian Communications and Media Authority (ACMA)¹²¹ regulates communications and media and

¹¹⁸ Australia's online safety framework has evolved through successive legislative reforms. The Enhancing Online Safety for Children Act 2015 initially focused on protecting children online. In 2017, amendments expanded the regulator's remit to address online harms affecting Australians more broadly, reflected in the renaming of the Children's eSafety Commissioner as the eSafety Commissioner and the legislation as the Enhancing Online Safety Act 2015. This framework was subsequently replaced by the Online Safety Act 2021, which consolidated and expanded the regulator's powers and introduced additional mechanisms for addressing online harms.

¹¹⁹ e-Safety Commissioner is the national regulator for online safety in Australia.

¹²⁰ OAIC deals with issues that are covered by The Privacy Act, 1988, <https://www.oaic.gov.au/privacy/privacy-legislation/the-privacy-act>.

¹²¹ ACMA covers misinformation and disinformation for traditional media as well as online platforms.

targets misinformation and disinformation while competition and consumer issues are handled by the Australian Competition and Consumer Commission (ACCC). These independent regulators came together, at the time of writing, for information sharing and collaboration under the Digital Platform Regulators Forum (DP-REG).¹²²

Classification Philosophy and Approach

Under the AU-OSA, digital platforms are categorized based on a **non-hierarchical functionality** based philosophy. Services are classified into eight broad, independent, and parallel categories of industry sectors, which is different from the EU-DSA's pyramid mode of hierarchical classification. The eight categories under the Act are:

- Internet carriage services,
- Social media services,
- App distribution services,
- Hosting services,
- Equipment services,
- Internet search engine services,
- Relevant electronic services (e.g., instant messaging services), and
- Designated internet services (can include websites).¹²³

Some of these categories, like the “relevant electronic services”¹²⁴ category, cover a wide and distinct range of services, which end up including disparate services like gaming, dating, and messaging.

Obligations, Enforcement Mechanisms and Implementation

At the primary level, the AU-OSA's Basic Online Safety Expectations (BOSE) is a scheme that sets out the government's minimum safety expectations of three service providers: social media services, relevant electronic services (eg: messaging, gaming), and designated internet services. These include ensuring the design of digital services prioritizes child safety, minimising unlawful and harmful activity, and having terms of use and processes for enforcing these terms. Regulated DSPs are also required to ensure safe use of service features such as anonymous accounts, encrypted services, generative AI, and recommender systems. This approach specifies government expectations without directly prescribing processes and practices that digital services must

¹²² Digital Platform Regulators Forum, "DP-REG Terms of Reference," Australian Government, last updated July 25, 2024, <https://dp-reg.gov.au/dp-reg-terms-reference>.

¹²³ Section 134, Subdivision A, Division 7, Australian Online Safety Act, 2021, <https://www.legislation.gov.au/C2021A00076/latest/text>

¹²⁴ Under the Act, the definition of "relevant electronic service" encompasses various electronic services that facilitate communication or interaction among end-users. E.g., email, instant messaging, SMS, chat services, online gaming services, etc. [Online Safety Act, Part 1, Section 13A]

adopt.^{125,126,127,128} These expectations are set by the Minister for Communications through a legislative instrument called a Determination, which also includes examples of reasonable steps that service providers can take to meet the identified expectations.

While failure to meet BOSE does not automatically trigger non-compliance penalties, it can result in transparency notices by the eSafety Commissioner. Failure to respond to reporting requirements in turn can result in penalties, including monetary fines. Therefore, compliance with BOSE can be reviewed and mandated in individual instances by the eSafety Commissioner.

The AU-OSA gives the eSafety Commissioner different mechanisms/ schemes to deal with specific types of harmful content. These mechanisms apply to different kinds of digital service providers and set clear responsibilities for content moderation in each case.

Scheme	Scope of Content Covered	Complaints Mechanism
Online Content Scheme <i>Covers ALL eight categories of DSPs</i>	Restricted and illegal content.	Complaints can be made by any Australia based individual/body corporate
Adult Cyber Abuse Scheme; Cyberbullying Scheme; and Image-Based Abuse Scheme <i>Covers social media, relevant electronic services, designated internet services, and hosting service providers.</i>	Different content-based Complaints Schemes addressing: - child cyberbullying, - adult cyber abuse, - non-consensual sharing of intimate images.	Complaints can only be made by individuals targeted or depicted in the content or their representatives.
Abhorrent violent conduct powers <i>Covers ISPs</i>	Material that portrays abhorrent violent conduct as defined under Australia’s criminal code as a person engaging in a terrorist act, murder or attempted murder, or torture, rape or kidnapping of another person.	The Commissioner takes action if it is satisfied that the availability of the material online will cause significant harm to the Australian community.

¹²⁵ Global Network Initiative. n.d. GNI Submission to the Australia Online Safety Act Review. <https://globalnetworkinitiative.org/wp-content/uploads/GNI-Submission-to-the-Australia-Online-Safety-Act-Review.pdf>.

¹²⁶ eSafety Commissioner. 2024. Basic Online Safety Expectations Regulatory Guidance. <https://www.esafety.gov.au/sites/default/files/2024-07/Basic-Online-Safety-Expectations-regulatory-guidance-July-2024.pdf>.

¹²⁷ Reset Tech Australia. 2024. “Statement on the Basic Online Safety Expectations Determination.” Reset Tech Australia. <https://au.reset.tech/news/statement-on-the-basic-online-safety-expectations-determination/>.

¹²⁸ Australian Government. 2023. Australian Government response to the House of Representatives Select Committee on Social Media and Online Safety report. <https://www.infrastructure.gov.au/sites/default/files/documents/australian-gov-response-to-house-of-reps-select-committee-on-social-media-and-online-safety-report-march2023.pdf>.

AU-OSA also establishes a unique collaborative approach between the regulator and industry bodies. The eight classes of digital service providers across the online industry are required to put in place their respective industry codes and standards to protect users from illegal and harmful content. This allows industry stakeholders to align sector-wide online safety efforts.¹²⁹ Once registered by the eSafety Commissioner, industry codes achieve binding force under the AU-OSA and non-compliance can yield penalties for relevant industries. The approach represents an example of co-regulation that involves the industry developing and managing its own interventions, with the government providing the legislative framework to support it.¹³⁰ Critical voices point to under-inclusion of non-commercial stakeholders in creating these frameworks.¹³¹

Key Takeaways

There are some clear lessons from the Australian experience that merit examination:

- **Co-regulation and its importance for India:** The delegated code-making process within the OCS has seen high levels of engagement from the industry¹³² as well as the public¹³³. It is a strong example of multi-stakeholderism that promises the following benefits. *First*, delegating some parts of rule-making can lead to greater consensus and less friction between government and industry. The significant delay in establishing rules under India's data law highlights the need for such a consultative process to yield timely regulatory progress.¹³⁴ *Second*, India's regulatory and resource capacity is relatively low which incentivises such a devolved approach to rule-making. As signals of similar thinking within Indian policymaking discourse, we note that the IT Rules, 2021 introduced self-regulating frameworks for digital media publishers.¹³⁵ The 2023 draft amendments to the IT Rules introduced similar self-regulation of online gaming services in India, proposing that they register with self-regulatory bodies (SRBs) recognized by the Indian IT Ministry.¹³⁶

¹²⁹ Interactive Games & Entertainment Association (IGEA). 2025. "New online safety codes to protect children from harmful content submitted to eSafety Commissioner." IGEA.
<https://igea.net/2025/03/new-online-safety-codes-to-protect-children-from-harmful-content-submitted-to-esafety-commissioner/>.

¹³⁰ Australian Government. 2023. Australian Government Guide to Policy Impact Analysis, The Office of Impact Analysis.
<https://oia.pmc.gov.au/sites/default/files/2023-02/oia-impact-analysis-guide-nov-22.pdf>.

¹³¹ Nottingham, Mark. 2022. "A Safer, More Centralised Australian Internet." Mark Nottingham.
<https://www.mnot.net/blog/2022/09/11/esafety-industry-codes#fnref:group>.

¹³² Online Safety Org. n.d. "Online Safety Codes 2025." <https://onlinesafety.org.au/>.

¹³³ eSafety Commissioner. n.d. "Background to the Phase 1 Standards." eSafety Commissioner.
<https://www.esafety.gov.au/industry/codes/background-to-the-phase-1-standards>.

¹³⁴ Mishra, Ashutosh. 2024. "One year of DPDP Act: Firms in a fix over delayed implementation of rules." Business Standard, August 11, 2024.
https://www.business-standard.com/economy/news/one-year-of-dpdp-act-delayed-rules-hamper-india-s-data-protection-law-124081100299_1.html.

¹³⁵ The constitutional validity of some of the relevant provisions of these rules were challenged by online platforms and digital news publishers, and are yet to be determined.
<https://www.scobserver.in/journal/supreme-court-transfers-challenges-to-it-rules-2021-to-the-delhi-high-court/>.

¹³⁶ Mehrotra, Anay, and Puneet Srivastava. 2024. "Online Gaming Platforms and Self-Regulation." Vidhi Centre for Legal Policy.
<https://vidhilegalpolicy.in/blog/online-gaming-platforms-and-self-regulation/>.

However, jurisdictional complications between centre and states, and lack of consensus within the industry, halted the creation of SRBs.¹³⁷

Finally, despite its advantages, any such framework should be designed to also account for diverse perspectives, particularly those of civil society and independent experts. This can counterbalance industry incentives to prioritise ease of compliance over efficacy of what the codes seek to achieve. While relying on industry-led policymaking, one must be additionally careful to not lose sight of the ultimate beneficiaries of a safety law - the users.

- **Limitations of classification framework:** AU-OSA's classification framework, while unique, presents challenges over applicability to emerging technologies. Given that the individual classes are narrowly described based on present-day functionalities (and potentially soon to be outdated), there is a risk of under-inclusion of newer digital services.¹³⁸ For example, it could be argued that AI chatbots may not fit into the existing categories¹³⁹ as they offer a unique functionality, creating a novel online industry.
- **Over-reliance on affected user complaints:** Another limitation of the AU-OSA is its reliance on complaints to initiate action under content-based complaints schemes (child cyberbullying, adult cyber abuse, and non-consensual sharing of intimate images). These are criticised for placing the primary responsibility of safety on the affected users while simultaneously limiting the ability of a *bystander* to raise complaints. Heavy reliance on complaints-based regulation poses challenges for a country like India due to low levels of digital literacy¹⁴⁰ and awareness¹⁴¹ which means that a significant portion of the population may not be able to access and use the complaints-based mechanisms optimally.
- **Built-in sunset clauses:** The AU-OSA has built-in review obligation, such that the working of the Act is reviewed with a view to update it every three years.¹⁴² Since it was enacted in 2022, the statutory three-year review of the legislation was brought forward to ensure the Act stays up to date. Periodic reviews of technology-related laws can help provide frameworks within which incremental updates/amendments can be effected to ensure regulation keeps pace with rapid technological advancements.
- **Coalition in Regulatory Action:** Finally, a coalition of national regulators such as Australia's DP-REG could solve some challenges around regulating emerging technologies

¹³⁷ Sarkar, Gargi. 2025. "Can Central Regulation Bring Stability To India's Gaming Industry?" *Inc42*, March 1, 2025. <https://removepaywalls.com/https://inc42.com/features/can-central-regulation-bring-stability-to-indias-gaming-industry/>.

¹³⁸ "Reviewing Australia's online safety laws." 2024. The Hon Michelle Rowland MP. <https://minister.infrastructure.gov.au/rowland/media-release/reviewing-australias-online-safety-laws>.

¹³⁹ Deb, Abhik, and Vineet Bhalla. 2024. "What do a minister's threats against Google's chatbot mean for AI in India?" *Scroll.in*, March 5, 2024. <https://scroll.in/article/1064700/what-does-a-ministers-threats-against-googles-chatbot-mean-for-ai-in-india>.

¹⁴⁰ Dattopant Thengadi National Board for Workers Education & Development. n.d. Digital Literacy. N.p.: Ministry of Labour & Employment, Govt. of India. https://dtnbwed.cbwe.gov.in/images/upload/Digital-Literacy_3ZNK.pdf.

¹⁴¹ Government of India. n.d. "Literacy." *Know India*. <https://knowindia.india.gov.in/profile/literacy.php>.

¹⁴² Braue, David. 2024. "eSafety can't keep up with risks of gen AI." *Informationage*, February 15, 2024. <https://ia.acs.org.au/article/2024/esafety-can-t-keep-up-with-risks-of-gen-ai.html>.

like Artificial Intelligence. For example, many online harms have multiple dimensions that engage the jurisdiction of different regulators. A coalition based approach avoids the complex exercise of determining appropriate jurisdiction, while allowing a multi-dimensional focus on evolving countermeasures. This clustering of challenges was evident in DP-REG's joint submission¹⁴³ on an AI discussion paper and a series of working papers on algorithms and the large language models (LLMs) used in generative artificial intelligence (AI).¹⁴⁴ In India, regulators have disagreed over jurisdictional issues in the past.^{145,146} Therefore, such an initiative, bringing together independent regulators backed by distinct legislations, may help India's governance ecosystem to enable coherent regulation.¹⁴⁷ Having said that, the DP-REG initiative was set up in 2022 rendering it too early to measure its efficacy and long-term impact.

¹⁴³ Department of Industry, Science and Resources. 2023. DP-REG Joint Submission to Department of Industry, Science and Resources – 'Safe and Responsible AI in Australia' discussion paper. [https://www.esafety.gov.au/sites/default/files/2023-08/DP-REG%20joint submission to DISR AI discussion%20paper July 2023.pdf](https://www.esafety.gov.au/sites/default/files/2023-08/DP-REG%20joint%20submission%20to%20DISR%20AI%20discussion%20paper%20July%202023.pdf).

¹⁴⁴ Digital Platform Regulators Forum. 2023. "Digital platform regulators release working papers on algorithms and AI." Australian Government. <https://dp-reg.gov.au/news-and-media/digital-platform-regulators-release-working-papers-algorithms-and-ai>.

¹⁴⁵ "Recurring Jurisdictional Turf-war between CCI and Sector Regulators - Towards Solution." n.d. CUTS International. <https://cuts-ccier.org/newsletter/spotlight-15.htm>.

¹⁴⁶ Barik, Soumyarendra. 2023. "TRAI can't regulate OTT platforms like Hotstar, says tribunal. This is why it matters." Indian Express, October 10, 2023. <https://indianexpress.com/article/explained/explained-sci-tech/tra-i-regulation-hotstar-ott-platforms-meity-8970887/>.

¹⁴⁷ Jain, Saumya, and Jyoti Panday. 2024. "Turf Wars and Definitional Ambiguities: OTT Regulation in India." Tech Policy Press, September 25, 2024. <https://www.techpolicy.press/turf-wars-and-definitional-ambiguities-ott-regulation-in-india/>.



Thailand

Overview

Global trends suggest that digital regulation broadly addresses two key areas in different instruments – *platform governance* and *competitive market concerns*. Thailand provides an interesting example, speculated to merge user safety and market competition objectives under its impending Platform Economy Act (PEA).¹⁴⁸ The draft PEA was shared with relevant stakeholders in 2024, and a publicly accessible draft is awaited at this time of our analysis.

Currently, the Thai regulatory framework includes:

- the Electronic Transaction Act (ETA) of 2001; and
- the Royal Decree on the “*Operation of Digital Platform Service Businesses that are subject to Prior Notification*” of 2022 (the Digital Platform Decree, 2022); and
- the Computer Crimes Act (CCA) of 2007 (amended in 2017), which governs online speech and content, and criminalises several forms of online expression.¹⁴⁹

Our analysis focuses on the Digital Platform Decree (DPD), issued in 2022 to protect users and standardise digital service provider operations.¹⁵⁰ The DPD builds on the Electronic Transactions Act (ETA) of 2001, and focuses on regulating digital platform services and operations.

¹⁴⁸ Weeravong, Chinnavat & Partners. 2025. “Thailand: proposed legislation sparks concern over how to balance regulation with innovation on digital platforms.” Global Competition Review. <https://globalcompetitionreview.com/review/the-asia-pacific-antitrust-review/2025/article/thailand-proposed-legislation-sparks-concern-over-how-balance-regulation-innovation-digital-platforms>.

¹⁴⁹ EngageMedia. 2022. “New report highlights impact of the Computer Crime Act on online expression in Thailand.” June 10, 2022. <https://engagemedia.org/2022/thailand-computer-crime-act/>.

¹⁵⁰ Tilleke & Gibbins. 2024. “Thailand Issues Draft Platform Economy Act.” Tilleke & Gibbins. <https://www.tilleke.com/insights/thailand-issues-draft-platform-economy-act/11/>.

Classification Philosophy and Approach

DPD defines “digital platforms services” (DPS) as those connecting business users, consumers, or users for electronic transactions,¹⁵¹ with or without remuneration (e.g. social media, e-commerce, search engines). It excludes platform services already regulated by the Bank of Thailand or the Securities and Exchange Commission as well as platform services provided by the government.

The DPD establishes a notification regime administered by the Thai Electronic Transactions Development Agency (ETDA¹⁵²). It combines size, risk, and impact considerations: size-based thresholds inform which DPS must submit long-form notification to the regulator about business operations. DPS that do not meet the size threshold are required to submit a brief notification to the regulator, and may attract additional obligations if they present particular risks or affect critical sectors. The size-based classification relies on both annual revenue and user base,¹⁵³ but our consultations indicate that there is considerable disagreement - at least at the time of writing - on which metrics to adopt, with stakeholders raising concerns about active user statistics.¹⁵⁴ Moreover, the size-based threshold specified for notification does not apply to services that pose specific risks, such as risk to financial security, risk of public harm, among others. It also does not apply to services that impact state security, health, environment, among other sectors.¹⁵⁵

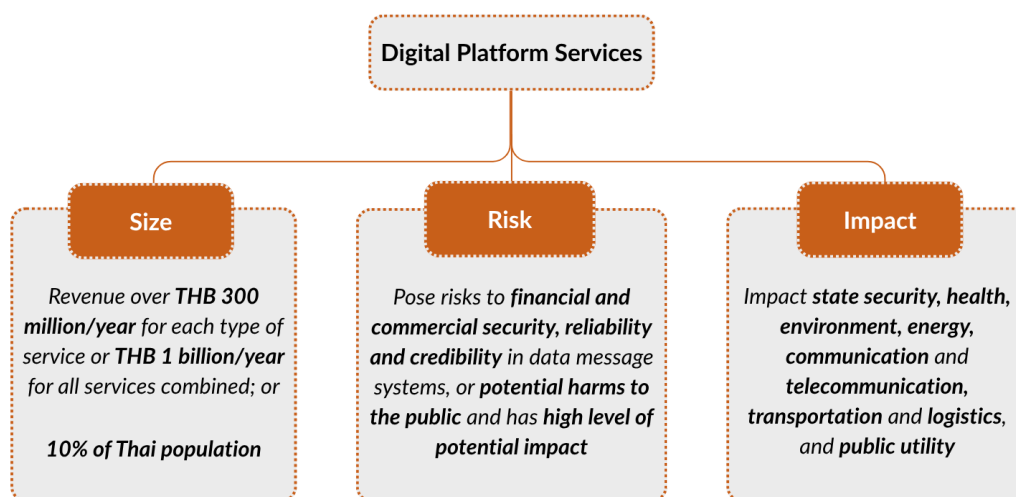


Figure: Classification parameters for notification of DPS under the DPD, 2022

¹⁵¹ “Electronic Transaction means a transaction in which an electronic means is used in whole or in part” – Section 4 of the Electronic Transactions Act, 2001

¹⁵² ETDA is a government agency which regulates electronic transactions under the Thai Ministry of Digital Economy and Society

¹⁵³ DPSs with an annual revenue of over 1.8 million (natural person) and THB 50 million (legal entity) or that have over 5000 AMAU; Average monthly active users (AMAU): DPD defines these as users who actually and frequently use the digital service, and users who log in from different devices are counted as one user.

¹⁵⁴ Some suggest revealing such metrics can be detrimental to investment

¹⁵⁵ Section 18 of Digital Platform Decree, 2022, <https://www.etcha.or.th/getattachment/Regulator/DigitalPlatform/law/Clean-Royal-Decree-on-DP-Corrected-1.pdf.aspx>

Businesses get categorized under these buckets largely through a self-classification process, submitting notifications about their business details to the regulator. If the notified information is found to be incorrect and/or incomplete, the regulator can request resubmission and even suspend services. This **reflects a form of mandatory self-classification approach under the Decree.**¹⁵⁶

The DPD also creates **functionality based classes**. These include:

1. Services that are provided to users for payment – eg: subscription-based content streaming
2. Services that act as intermediaries for offering goods or services – eg: online marketplaces
3. Services that are provided by an operator that has contracts with businesses that offer goods or services to consumers – eg: app stores
4. Online search engines

These classes are **subject to additional transparency obligations** under the DPD.

Search engines are likely specified as a separate class because they may not typically fit under the other three categories, but do enable electronic transactions by enabling discovery of goods and services, and therefore shape online user transactions.¹⁵⁷ In the draft PEA, search engines are explicitly listed as “core platform services” that could qualify as gatekeepers that enable or deny access to other service providers, reflecting Thailand’s focus on gatekeeper regulation.¹⁵⁸

Obligations, Enforcement Mechanisms and Implementation

The **ETDA specifies which platforms under the size, risk and impact buckets are subject to obligations.**¹⁵⁹ Similar to EU-DSA, these include **process-based obligations** like conducting risk assessments, appointing compliance officers, undergoing external audits, and fulfilling any additional requirements prescribed by the regulator.¹⁶⁰ The DPD however refrains from specifying mitigation measures, and allows service providers to **self-regulate based on their risk**

¹⁵⁶ Section 13 of Digital Platform Decree, 2022, <https://www.etda.or.th/getattachment/Regulator/DigitalPlatform/law/Clean-Royal-Decree-on-DP-Corrected-1.pdf.aspx>.

¹⁵⁷ “Thailand Digital Platform Services Law – Compliance Required By 18/11/2023.” 2023. Aust Cham Thailand. <https://www.austchamthailand.com/resources/news/thailand-digital-platform-services-law-compliance-required-18112023>.

¹⁵⁸ Sornchangwat, Poonyisa, and Kwanchanok Jantakram. 2025. “Key Insights on Draft New Legislation Regulating Digital Platforms (Thailand).” *NO&T Asia Legal Review* 98 (March). <https://www.noandt.com/en/publications/publication20250328-1>.

¹⁵⁹ Section 18 of Digital Platform Decree, 2022.

¹⁶⁰ Section 19, 20 and 21 of Digital Platform Decree, 2022.

assessments.¹⁶¹ Non-compliance with DPD obligations could result in THB 100,000 fine or 1-year imprisonment or both.¹⁶²

The draft PEA signals potential expansion of regulatory powers, but its final form remains uncertain. Industry responses to the draft urged caution against prescriptive, “EU-style” obligations - companies have specifically warned about mandatory user verification, real-time content monitoring, automated moderation mandates, and forced transparency requirements for algorithms, citing feasibility, privacy, and commercial-sensitivity concerns as well as plausible stifling of innovation and investment in Thailand’s nascent-stage digital economy.^{163,164}

Key Takeaways

The DPD provides the following valuable insights:

- **Regulatory oversight in self-classifying regimes:** The DPD provides a useful example of a mixed approach - notification plus risk-based exceptions - illustrating how a regulator can combine self-classification with supervisory checks. Its practical effectiveness will depend on ETDA’s verification capacity and the clarity of guidance on process-based obligations. Without adequate regulatory capacity to verify accurate representation and ensure consistent compliance, **high risk platforms may self-classify under low-risk categories or undertake only tokenistic remedial measures.** This misalignment of platforms’ incentives with user safety needs **can result in sub-optimal outcomes.**¹⁶⁵
- **Targeting specific risks and priority areas, irrespective of size:** The Decree defines high-priority risks - such as financial security, state security, and health - by setting clear classification parameters. Like the four systemic risks under the EU-DSA, this allows jurisdictions to tailor frameworks to region-specific challenges and policy goals. Notably, size thresholds do not apply when services pose these specified risks, meaning even smaller platforms may fall under the DPD. Our consultations confirm that user safety risks can emerge from smaller or fringe services, making the Thai DPD a useful example of risk-based classification beyond service size.

¹⁶¹ Section 32 of Digital Platform Decree, 2022.

¹⁶² Yangiam, Nopparak, and Sunthorn-opas, Chattong. 2023. “Thailand’s Royal Decree on Digital Platform Services – Update from the Draft to the Recently Enacted Law.” *NO&T Asia Legal Review* 55 (January). <https://www.noandt.com/en/publications/publication20230131-3>.

¹⁶³ Damazo-Santos, Jet. 2024. “Major tech companies raise concerns over EU-inspired draft platform-economy law in Thailand.” MLex. <https://www.mlex.com/mlex/articles/2155734/major-tech-companies-raise-concerns-over-eu-inspired-draft-platform-economy-law-in-thailand>.

¹⁶⁴ Asia Internet Coalition. 2024. AIC Industry Submission on Thailand Platform Economy Act, 2024. <https://aicasia.org/policy-advocacy/aic-industry-submission-on-thailand-platform-economy-act-2024/>.

¹⁶⁵ Cusumano, Michael A., Annabelle Gawer, and David B. Yoffie. 2021. “Can Self-Regulation Save Digital Platforms?” *Industrial and Corporate Change* 30 (5): 1259–1285. <https://www.hks.harvard.edu/centers/mrcbg/programs/growthpolicy/can-self-regulation-save-digital-platforms-david-yoffie>.



Singapore

Overview

Historically, Singapore has regulated online speech through content-specific laws such as Protection from Harassment Act 2014, Protection from Online Falsehoods and Manipulation Act 2019 or POFMA, and Foreign Interference (Countermeasures) Act 2021 or FICA. Both POFMA (which addresses mis- and disinformation) and FICA (which criminalises engagement with non-Singaporeans on political and social issues) have faced criticism for allegedly suppressing online speech¹⁶⁶ and activism.^{167,168}

By contrast, Singapore's Online Safety (Miscellaneous Amendment) Act (SG-OSA) 2022 marks a shift towards a general-scope approach aimed at establishing service provider accountability for addressing user safety concerns.¹⁶⁹ It amends the 1994 Broadcasting Act, bringing "online communication services" (OCS) under its ambit. Among other things, this allows extraterritorial

¹⁶⁶ Human Rights Watch. 2020. "Singapore: Free Expression Restrictions Tighten." January 14, 2020. <https://www.hrw.org/news/2020/01/14/singapore-free-expression-restrictions-tighten>.

¹⁶⁷ Annamalai, Kokila. 2021. "Singapore's 'foreign interference' law will weaken people power." *Al Jazeera*, October 18, 2021. <https://www.aljazeera.com/opinions/2021/10/18/singapores-foreign-interference-law-will-crush-people-power>.

¹⁶⁸ Chng, Kenny. 2023. "Falsehoods, Foreign Interference, and Compelled Speech in Singapore." *Asian Journal of Comparative Law*. <https://www.cambridge.org/core/journals/asian-journal-of-comparative-law/article/falsehoods-foreign-interference-and-compelled-speech-in-singapore/55D1F297B03B6E2545D557EB4EBD6E9B>.

¹⁶⁹ Kaarlep, Agne, Ilana Rosenzweig, and Noah Douglas. 2023. "Platform Regulation in APAC and the EU: A Comparative Overview." *Platform Regulation in APAC and the EU: A Comparative Overview*, September 27, 2023. <https://www.techpolicy.press/platform-regulation-in-apac-and-the-eu-a-comparative-overview/>.

application of the broadcast law to social media services dominant in Singapore.¹⁷⁰ SG-OSA adopts a **risk-based classification framework**, creating a **sub-class of Regulated Online Communication Services (ROCS)**. It prescribes both reactive and proactive obligations focused on preventing eight predefined categories of harmful content, and has a separate online safety code for ROCS. The Act is enforced by the Infocomm Media Development Authority (IMDA).¹⁷¹

Classification Philosophy and Approach

SG-OSA defines “online communication services” as services that enable point-to-multipoint communication (or broadcast-like communication) or allow users to access content over the internet. This is also relevant because, unlike POFMA, SG-OSA explicitly exempts private messaging services.¹⁷² Parliamentary debates suggest that the exclusion was premised on preserving user privacy. The debates raise important questions about the extent to which speech remains private, even on messaging platforms. *For instance, Facebook group chats that are publicly accessible or those that permit unrestricted access could potentially fall within the scope of the Act.*¹⁷³

Within online communication services (OCS), SG-OSA presents a **risk-based classification** by creating a **separate class of regulated online communication services (ROCS) based on the “nature of the effect”** that these have on Singapore users and “different communities”. The Act also indicates **some consideration for the size or the “extent” of impact for identifying ROCS**, which face additional due diligence obligations articulated under the Code of Practice (COP) on Online Safety.

Notably, rather than **specifying a quantitative size threshold**, the SG-OSA empowers the Infocomm Media Development Authority (IMDA) to identify ROCS through qualitative assessment based on the (i) range of all online communication services offered to Singaporean users and (ii) **extent and nature of impact** these services have on users and different communities. This **flexibility in phrasing offers the IMDA significant control over classification**. While arbitrary control is usually faced with some opposition, in 2022, Singaporean Communications Minister confirmed that regulated entities were found to be “receptive” to the proactive obligations proposed under the Code of Practice (COP) during consultations.¹⁷⁴ As of 18 July 2023, the IMDA

¹⁷⁰ Lim, Vanessa, and Davina Tham. 2022. “MPs debate 'egregious' content and privacy concerns under proposed law to enhance online safety.” *MPs debate 'egregious' content and privacy concerns under proposed law to enhance online safety*, November 8, 2022. <https://www.channelnewsasia.com/singapore/online-safety-egregious-harmful-content-children-social-media-josephine-teo-305358>.

¹⁷¹ Government body under the Singaporean Ministry of Digital Development and Information

¹⁷² Section 2A (3)(e) of Online Safety (Miscellaneous Amendments) Act 2022 [Part 1 Amendment of Broadcasting Act 1994], <https://sso.agc.gov.sg/Acts-Supp/38-2022/Published/20221221?DocDate=20221221&WholeDoc=1>

¹⁷³ Lim, Vanessa, and Davina Tham. 2022. “MPs debate 'egregious' content and privacy concerns under proposed law to enhance online safety.” *Channel News Asia*, November 8, 2022. <https://www.channelnewsasia.com/singapore/online-safety-egregious-harmful-content-children-social-media-josephine-teo-305358>.

¹⁷⁴ Ibid.

had identified 6 social media services as ROCS – Facebook, Instagram, TikTok, YouTube, X (formerly Twitter), and HardwareZone.¹⁷⁵

Obligations, Enforcement Mechanisms and Implementation

SG-OSA's reactive obligations apply to all online communication services (OCS), and involve content takedown/blocking orders related to “egregious” or harmful content. Internet service providers (ISPs) may also be ordered to block access to harmful content. The IMDA issues these orders, but the scope of what constitutes harmful content is predefined,¹⁷⁶ potentially limiting arbitrary enforcement. The Singaporean government emphasizes that the objective of these orders is to takedown harmful content “within hours”.¹⁷⁷ In June 2023, IMDA issued its first content blocking orders, requiring Meta to quickly review and remove a Facebook Page and Group containing CSAM, which Meta addressed within 24 hours.¹⁷⁸ ISPs were also asked to block access to curb further distribution.

Additional obligations imposed on ROCS under the code of practice *COP on Online Safety*¹⁷⁹ focus on **preventing adverse outcomes related to six predefined categories of harmful content**,¹⁸⁰ **without prescribing rigid due diligence processes**. For example, services must actively offer relevant safety information to users searching for high-risk key words such as “suicide” or “self harm.” However, measures to achieve this at the service level can be defined by the providers. This aligns with pre-enactment public consultations held in 2022, which reflected an **industry preference for outcome-based obligations that allow for flexibility and take into account different business models and internal processes/policies**.¹⁸¹

COP obligations are divided into three buckets – *User Safety, User Reporting and Resolution and Accountability*. Beyond the safety of all end-users, user safety obligations recognize children as

¹⁷⁵ IMDA: Designated Social Media Services, <https://www.imda.gov.sg/-/media/imda/files/regulations-and-licensing/regulations/codes-of-practice/codes-of-practice-media/list-of-designated-social-media-services-subject-to-the-code-of-practice-for-online-safety.pdf>

¹⁷⁶ Section 45D (1) of *Online Safety (Miscellaneous Amendments) Act 2022* – (a) suicide and self harm, (b) violence and cruelty, (c) sexual violence, (d) CSAM, (e) public health risks, (f) matters of race or religion sentiments, (g) terrorism and (h) any other content that is prescribed by Part 10A as “egregious”

¹⁷⁷ Lim, Vanessa, and Davina Tham. 2022. “Singapore passes law requiring social media sites to block harmful content ‘within hours.’” *Channel News Asia*, November 9, 2022. <https://www.channelnewsasia.com/singapore/social-media-sites-block-access-harmful-content-within-hours-under-new-law-passed-parliament-3057041>.

¹⁷⁸ Rahim, Nisha. 2023. “Meta removes Facebook page, group with child exploitation content after IMDA alert.” *Yahoo News*, June 9, 2023. <https://sg.news.yahoo.com/imda-meta-remove-facebook-page-child-sexual-exploitation-content-083403085.html>.

¹⁷⁹ Infocomm Media Development Authority (IMDA), Code of Practice for Online Safety, Singapore, <https://www.imda.gov.sg/-/media/imda/files/regulations-and-licensing/regulations/codes-of-practice/codes-of-practice-media/code-of-practice-for-online-safety.pdf>.

¹⁸⁰ These are: a. sexual content; b. violent content; c. suicide and self-harm content; d. cyberbullying content; e. content endangering public health; and f. content facilitating vice and organized crime.

¹⁸¹ Ministry of Communications and Information. 2022. “Public Consultation on Enhancing Online Safety for Users in Singapore.” REACH. <https://www.reach.gov.sg/Participate/Public-Consultation/Ministry-of-Communications-and-Information/Public-Consultation-on-Enhancing-Online-Safety-for-Users-in-Singapore>.

particularly vulnerable to online harm. Without prescribing rigid processes, COP requires ROCS to provide parents/guardians access to risk minimising tools such as limited public visibility or location sharing for children's accounts.

Predefined outcomes can shape the processes platforms put in place. For example, COP requires ROCS to let both the reporter and the original poster of flagged content seek a review (the outcome). To enable this, platforms will need to set up an internal grievance system (the process).

Non-compliance

SG-OSA has predefined penalties for OCS, ROCS and ISPs. IMDA can impose penalties up to SGD 1 million on OCS and up to a maximum of \$500,000 on ISPs for non-compliance with content takedown/blocking orders. ROCS may face additional penalties up to SGD 1 million for non-compliance with obligations. These penalties, however, are relatively rare. In 2020, amid the global pandemic, the IMDA imposed fines worth over SGD 600,000 on two major telecom providers for internet disruptions, viewed as non-compliance with the code of practice on Telecommunication Service Resilience.¹⁸² **While hefty, these fines appear to be used sparingly, with not too many reported instances. Fewer instances may also be indicative of the effectiveness of monetary penalties as a deterrent to ensure compliance.** The Singaporean Communication Minister has even described the penalties as a “*slap on the wrist*” compared to non-compliance measures in other jurisdictions.¹⁸³

In terms of regulatory design, the SG-OSA allows some room to claim exemption. **Service providers are exempt if they can prove that it was not reasonably practicable to do more than what was done to meet their obligations, or if there was no better practicable means available.** For instance, when telecom providers faced penalties in 2020, M1 (Singaporean telco) was only charged for one of its two instances of disruption. The IMDA clarified that the exempted incident was a “first of its kind” equipment disruption that could not have been reasonably foreseen or prevented by the provider.

Key Takeaways

Policymakers should consider the following elements of SG-OSA in designing their safety frameworks:

- **Limits to power and legal ambiguity:** While the SG-OSA grants significant power to the IMDA, it mitigates the risk of overreach by predefining categories of harmful content that

¹⁸² Baharudin, Hariz. 2021. “IMDA Fines M1 and StarHub \$610,000 in Total for Internet Disruptions During Circuit Breaker.” *The Straits Times*, June 8, 2021.

<https://www.straitstimes.com/singapore/imda-fines-m1-and-starhub-610000-in-total-for-internet-disruptions-during-circuit-breaker>.

¹⁸³ Channel News Asia. 2022. “Singapore Passes Law Requiring Social Media Sites to Block Harmful Content ‘within Hours.’” CNA, November 9, 2022.

<https://www.channelnewsasia.com/singapore/social-media-sites-block-access-harmful-content-within-hours-under-new-law-passed-parliament-3057041>.

is subjected to takedown or blocking. For instance, Indian IT Act and Rules mention broad categories of harmful content such as "obscene" or "defamatory" content, SG-OSA seeks to narrower categories such as suicide and self-harm promotion or content promoting violence or extreme cruelty.

- **Outcome-based regulatory approach:** Our stakeholder consultations revealed that highly prescriptive compliance processes increase the burden on digital service providers and demand significant regulatory capacity. In this regard, Singapore OSA's outcome-based approach could be instructive. Rather than prescribing processes or design features, it sets clear goals for service providers. For example, the online safety code for ROCS recommends design changes to protect minors but stops short of mandating specific solutions. In practice, however, jurisdictions may benefit from balancing process requirements with clear outcomes. The EU DSA, for instance, mandates risk assessments for VLOPs and VLOSEs, while also specifying systemic risks that these service providers must actively prevent.

05

Considerations in Designing an Online Safety Framework

Drawing on international regulatory approaches, stakeholder consultations and supporting research, this chapter identifies key principles that should inform an effective online safety framework. It covers:

- Regulatory objectives
- The evolving nature of online harms
- Functionality as a proxy for risk
- The philosophy of obligations
- Implementation strategies and practices

5.1. Objectives an Online Safety Framework Should Pursue

The central objective of an online safety framework is to **protect end-users from harms arising from their interactions with digital services, including interactions with other users online**. A broader goal is to generate spillover benefits for the wider public in an increasingly digital society - preventing online harms from resulting¹⁸⁴ in adverse real-world consequences.

Our research shows that different jurisdictions have approached this by recognising DSPs' duty of care towards their users' safety, which has led online safety frameworks to impose proactive obligations on DSPs to address various safety risks. Landmark legislations such as the EU-DSA and UK-OSA have taken differing approaches and are still in the early stages of implementation. In India, online safety has also been a key policy concern, with discussions including potential reforms of the existing IT Act, 2000. Therefore, a key objective of our research has been to identify pathways for India to promote online safety by establishing **clear, proportionate and future-proof obligations for DSPs**.

We believe that the following objectives are central to online safety policymaking, which our framework (*detailed in Chapter 6*) optimises for:

1. Risk-centred Approach to Targeted Solutions

Designing online safety regulations presents two significant challenges:

- ***Prioritising the right problems to solve***

Online safety risks lie on a diverse spectrum, differing from each other in their severity, nature, and other underlying factors. Further, the digital realm is dynamic where new forms of harms evolve that may warrant regulatory attention. Therefore, for an online safety framework to be future-proof, it must **establish systems that enable periodic/dynamic identification of priority problem statements**, as well as guidance to policymakers in **delineating which problems indeed require regulatory interventions**. We believe this provides a more sustainable approach rather than merely having a few safety issues baked into legislative text.

- ***Identifying solutions that work well for the diverse digital economy***

The digital economy comprises a vast diversity of services that offer different functionalities to users. At a high-level, there is hardly a common thread that binds *all* services together in their technology, utility, or ability to mitigate online safety risks. This leads to a situation where introducing regulation (even where required) in a way that accommodates this diversity becomes near-impossible. Therefore, another important

¹⁸⁴ Banaji, Shakuntala, and Ram Bhat. (n.d.). "WhatsApp Vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India." London School of Economics.
<https://www.lse.ac.uk/media-and-communications/assets/documents/research/projects/WhatsApp-Misinformation-Report.pdf>.

objective of an online safety framework must be to **classify the digital economy** to clarify the responsibility of each distinct “class” of service towards the broader objective of online safety.

2. Service Provider Accountability and User Empowerment

While online safety risks often arise from user interactions, the functionality and design of digital services can play a catalytic role in enabling, amplifying, or perpetuating harms. Regulation should therefore **clearly define service providers’ duty of care and hold DSPs accountable** when this responsibility is not met. This requires setting out specific responsibilities for different types of services: for many, only baseline obligations may be necessary, while higher-risk services may warrant additional, targeted measures.

Seen from another perspective, a key aspect of service provider accountability is to **empower users** to be able to hold DSPs accountable. This includes effective grievance redressal mechanisms, user reporting and transparent content moderation policies. Such measures can act as important distributed safeguards, particularly relevant for global majority contexts, to avoid relying solely on regulatory action to address risks.¹⁸⁵

3. Clarity and Predictability in Policymaking

The limitations of over-broad definitions of service provider classes are evident from the Indian technology regulatory landscape. *For instance, the wide definitions of the terms “intermediaries” and “social media intermediaries” under the IT law¹⁸⁶ means that the legal framework lays down the same obligations for functionally distinct digital services like social media, online dating, online messaging and video calling services.* This can create ambiguity around service provider responsibilities in different contexts and may be inadequate to address specific risks.¹⁸⁷ Beyond tightly defining service classes, an effective online safety framework should also clearly articulate its approach to the following:

- **Policy objectives** – the specific goals that all stakeholders should work towards.

¹⁸⁵ Tavishi and Shobhit S. 2025. “Platform Transparency under the EU’s Digital Services Act: Opportunities and Challenges for the Global South.” Centre for Communication Governance.
<https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/platform-transparency-under-the-eus-digital-services-act-775.pdf>.

¹⁸⁶ Section 2(1)(w) of the IT Act, 2000 defines an intermediary as “any person who, on behalf of another person, receives, stores, or transmits electronic records, or provides any service related to such records”. This definition encompasses a wide range of entities, including telecom service providers, network service providers, and various digital platforms irrespective of the amount of control they exercise over the data they are transmitting. Similarly, IT Rules 2021 define a social media intermediary as “an intermediary which primarily or solely enables online interaction between two or more users and allows them to create, upload, share, disseminate, modify or access information using its services”.

¹⁸⁷ For instance, content moderation obligations outlined for Significant Social Media Intermediaries may not be possible to implement for high-privacy services like encrypted messaging, or cloud service providers. They may also not be relevant for online dating apps or matrimonial websites that are also considered social media as per the broad definition of “social media intermediaries” used in the IT Rules.

- **Authority** – the principal policymaking bodies responsible for developing and implementing policy.
- **Process** – when and how policies will be developed, ensuring they are deliberative rather than reactive.
- **Predictability and guidance** – clarity on the likely content and direction of future safety requirements, enabling DSPs to anticipate and prepare for compliance.

In developing our recommendations, we have kept these questions in mind and adopted a modular structure that classifies digital services by functionality to ensure clarity of responsibility. This approach promotes predictability by signalling to DSPs the safety objectives that must be met when enabling or rolling out specific functionalities.

4. Future Ready Regulatory Design

Finally, we recognize that technology evolves rapidly and associated risks grow more complex with this evolution.¹⁸⁸ When policymakers design interventions keeping in mind technology architectures that are prevalent today, it is possible (and likely) that some of their assumptions will be out of touch when technology architectures evolve. *For example, the content moderation mandates under the IT Rules may not be appropriate for increasingly popular DSPs that offer federated architectures.*

- Thus, it is important to set **clear, technology-neutral safety standards** for each class of service such that they are **scalable and adaptable to new technologies** over time. This approach **signals to innovators the intent and impact of safety measures** they must keep in mind when developing new services, though the form in which they embed safeguards may look different when compared to traditional technologies of today. *(Explored with an illustration later in this Chapter)*
- On another note, future proofing also means creating systematic avenues for policymakers to **review and re-assess the efficacy of regulation** in the context of evolving technology and changing online safety risks, allowing for timely realignment when needed.¹⁸⁹ *This can be met through mandatory review obligations in legislation, or sunset clauses that force deliberate reconsideration of policy approaches.*

5.2. Understanding the Divergent Nature of Online Safety Risks

Good policy hygiene dictates that the first step to identifying sustainable interventions is to clearly identify the problem statement and its contours. Therefore, we begin by assessing the

¹⁸⁸ eSafety Commissioner, Australian Government. 2025. "AI chatbots and companions – risks to children and young people." eSafety Commissioner. <https://www.esafety.gov.au/newsroom/blogs/ai-chatbots-and-companions-risks-to-children-and-young-people>.

¹⁸⁹ Challenges and risks posed by Generative AI, see, Berger, Ilana, and Tim Torres. 2025. "What is Trust and Safety? A Comprehensive Guide." Alice September 20, 2024. <https://alice.io/blog/what-is-trust-and-safety-comprehensive-guide>.

spectrum of different online safety risks that a potential online safety framework is intended to address.

Online safety risks are best understood as the **propensity for users to experience harms** in their usage of digital services. In turn, online harms denote a vast spectrum of adverse experiences, damage or loss that that users can be exposed to. While most online harms emerge from user behaviour (via content, conduct, or contact that is problematic),¹⁹⁰ harms are numerous and diverse. To illustrate the diversity within the broader universe of harms, some attributes that can be used to characterise or differentiate them include:¹⁹¹

- **Nature/origin:** Harms may arise from content (e.g., *violent imagery*), conduct (e.g., *cyberstalking*), or contact (e.g., *grooming*).
- **Manner of manifestation:** Some harms remain virtual (e.g., *cyberbullying*), while others can translate into physical, real-world consequences (e.g., *online radicalization leading to violence*).
- **Impact:** Harms vary in severity, scale and the nature of impact. *For e.g., misinformation which has a large-scale impact may be less traumatizing individually than harms like cyberflashing or cyberstalking that target individuals but cause severe personal trauma.* Some harms disproportionately target **vulnerable communities** - such as exposure to harmful content for minors or identity-based targeted bullying of women, LGBTQ+, persons with disabilities. Other harms are more **generalized** and targeted at society at large (e.g. large scale disinformation during elections, public health emergencies, etc.).

While the potential for individual harm can be shaped by any combination of these factors, another important dimension emerges at a broader view: the **capacity of certain online harms to persist and spread beyond isolated incidents**. This is the essence of “**systemic risk**”. Such risks rarely stem from a single piece of content or a lone act of misconduct; instead, they arise from the typical ways in which users - and bad actors - interact with the core functionalities of digital services.

For example, the spread of hate speech on social media platforms is not a result of a few isolated comments, but a “systemic risk” due to how users’ harmful behavior is magnified by platform functionalities - such as algorithmic amplification, engagement-based ranking, open-group formation, low-friction sharing, and absence of source credibility indicators (e.g., fact check labels, community

¹⁹⁰ Child, Daniel, John-Orr Hanna, Adam Hildreth, and Julie Inman Grant. 2023. “Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms.” World Economic Forum. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf; Better Internet for Kids, European Union. (n.d.). “Key Online Safety Risks.” European Union. <https://better-internet-for-kids.europa.eu/en/learning-corners/teachers-and-educators/key-online-safety-risks>; Government of United Kingdom. 2020. “Online Harms White Paper.” Government of United Kingdom. <https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper#the-harms-in-scope>.

¹⁹¹ We recognise that there is no widely (let alone, universally) accepted typology or taxonomy of online harms. The delineation of the attributes below is non-exhaustive, and intended only to demonstrate the wide diversity and possible points of classification of different individual harms.

notes).¹⁹² *The risk is embedded not in any single piece of content, but in the predictable ways these features are used to spread hate speech at scale, making the harm persistent and structurally embedded in the platform's overall design and use patterns.*

From the analysis above, we derive the following principles that are central to our approach:

- Online harms are wide ranging and diverse. However, a sub-set of online harms are persistent and exhibit phenomenon-type characteristics, creating **systemic risks**.
- In developing safety-focused regulatory mandates for platforms, policymakers may find it useful to prioritise these systemic risks as the core problem to address, even as they build in baseline provisions to manage sporadic or isolated harms that do not rise to the level of systemic risk, reducing the need for targeted regulatory interventions in response to one-off incidents.
- Each systemic risk is unique and often requires a tailored approach to address it effectively. Policymakers should consider its specific causes and manifestations when designing countermeasures. A one-size-fits-all approach, if used at all, should be a deliberate choice rather than the default.

5.3. Determinants of Harm

5.3.1. Functionality as a Suitable Proxy for Risk of Harm

As diverse as online harms are, not all digital services are equally prone to them. Each service's risk profile is shaped by the unique intersection of harm factors outlined above. **We argue that the primary determinant of risk profile is the set of functionalities a DSP offers its users.** Since harms largely stem from user behaviour, functionalities - what a service enables or restricts - directly influence the likelihood of specific risks arising.

For example, consider a spread-based harm such as disinformation: DSPs that enable high-reach interaction are far more susceptible to the harm rather than DSPs that provide food-delivery or ride-hailing apps. Therefore, it is possible to identify a **convincing level of correlation between functionalities offered by DSPs and the profile of harms they may be susceptible to.**

Crucially, these harm factors can also be **mitigated by focusing on the design, use, and management of the functionality** and features. *For example, if a platform faces risks that can scale rapidly, such as misinformation, it can minimize this risk by adjusting functionalities that limit sharing, introduce friction, etc. Similarly, if a platform disproportionately exposes minors to harm, it can mitigate this risk by restricting anonymous interactions or disabling direct messaging to minors.*

¹⁹² Akers, John, et al. 2018. "Technology-Enabled Disinformation: Summary, Lessons, and Recommendations." *Technical Report*, Paul G. Allen School of Computer Science & Engineering, University of Washington, (December): [10.48550/arXiv.1812.09383](https://arxiv.org/abs/10.48550/arXiv.1812.09383); Ceylan, Gizem, Ian A. Anderson, and Wendy Wood. 2023. "Sharing of misinformation is habitual, not just lazy or biased." *Proceedings of the National Academy of Sciences* 120, no. 4 (January). [10.1073/pnas.221661412](https://doi.org/10.1073/pnas.221661412).

Functionality, therefore, offers a practical basis for assigning regulatory obligations and targeting high-risk services. **We propose that it serve as the core unit of focus for any online safety framework.**

Current approaches often regulate entire services as indivisible units, leading to sub-optimal, one-size-fits-all obligations. *The EU-DSA, for example, groups vastly different services together under the VLOP category - treating TikTok, Wikipedia and Amazon Store alike despite their very different risk profiles. While the mandated risk assessments can surface service-specific risks and mitigation measures, the overbroad classification makes tailored interventions difficult.*

Our framework recognises that digital services are more often than not presented to users as a bundle of individual functionalities, and therefore, we must focus on these functionalities as the centre of gravity for any regulatory analysis. *For example, an online gaming service provider that allows users to chat in closed groups must be recognised as constituting two different functionalities - gaming and private messaging.* A functionality specific approach, while offering benefits in targeting obligations more precisely, also prevents DSPs from being unfairly burdened with obligations that are inapplicable or irrelevant to the functionalities they offer.

Finally, by using functionality as the predominant metric of classifying digital services, we overcome the challenges associated with classification by risk and classification by size explored in Chapter 3.

This leads us to the following considerations that guide our classification framework:

- Functionality is the optimal proxy for assessing online harm potential, and can serve as the basis for classifying DSPs to ensure proportionate online safety obligations.
- The fundamental unit of focus for an online safety framework should be specific functionalities, not the DSP or platform as a whole. This approach reflects the reality of today's multi-modal platforms, which bundle diverse functionalities under one roof.

5.3.2. Other Determinants of Harm

The potential for harm arising from a digital service is also influenced by a range of intersecting factors. Though **functionality is the prime determinant**, there are also other factors like size or type of user base and business model that contribute to a services' risk propensity. It is therefore crucial to understand the **interplay of the following modulating factors** when defining service provider responsibilities:

- **Scale:** Due to their scale and reach, harms on larger platforms affect far more users than on smaller DSPs. It is therefore understandable that policymakers worldwide often begin by imposing safety obligations on large platforms, given both their greater capacity to implement changes and their potential for a much larger impact on overall user safety.

Equally, it is untrue that large platforms *alone* present high risk potential. Smaller platforms often operate outside regulatory scrutiny, but can be regionally popular or nonetheless be used to perpetuate high-severity harms.¹⁹³ **In this sense, size often amplifies harm; it does not create it.**

- **Service Design and Affordances:** While platform functionality determines the core actions a user can perform, platform design and affordances determine how these actions are presented, encouraged and mediated, through interface choices, default settings, algorithmic systems, etc. **Design governs the user experience and behavioral incentives that are layered on top of platform functionality.**

*For instance, two messaging platforms A and B can present vastly different risk potential given differences in service design. Platform A which restricts group size and limits one-click sharing of multimedia links to a single group at a time is likely to present lesser risks of disinformation spread as compared to Platform B which allows 1,00,000 users in a single group with no restrictions or friction on sharing material with multiple high-reach groups. **Therefore, even with similar functionalities, platforms with different designs can pose vastly different levels of harm.***

- **Nature of User Base:** Digital services predominantly used by vulnerable or targeted groups, such as children, women and other gender minorities, religious minorities, etc., face heightened risks - that is also a digital reflection of heightened vulnerabilities in the real world.

For example, social media platforms that primarily cater to or are used by children or adolescents may need to shoulder a greater burden in moderating harmful content or limiting unwanted stranger contact as opposed to other platforms - owing to uniquely heightened risks of these harms to children, as compared to adults. Further, online dating platforms for queer men are often used as a gateway for cyberbullying, non-consensual intimate image sharing, doxxing, or even harms that perpetuate into the offline world, like extortion, criminal intimidation, etc. which could call for greater safety interventions that are relevant to this context.¹⁹⁴

Therefore, platforms serving at-risk communities may need stronger safeguards, moderation, and support to counter disproportionate harms.

From the analysis above, we derive the following insights which are central to our approach:

- While adopting a functionality based approach to an online safety framework is

¹⁹³ Ofcom. 2025. "Helping Small Services Navigate the Online Safety Act." 2025. Ofcom. January 24, 2025. <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/helping-small-services-navigate-the-online-safety-act>.

¹⁹⁴ NDTV News Desk. 2025. "UP Man Loses Rs 1.4 Lakh In Grindr Date After He Is Filmed During Sexual Act." NDTV, February 2, 2025. <https://www.ndtv.com/india-news/up-man-loses-1-4-lakh-in-grindr-date-after-he-is-filmed-during-sexual-act-7615194>; Express News Service. 2025. "Second Gang Busted in 5 months: 4 held in Noida for extorting people using gay-dating app." *Indian Express*, May 20, 2025. <https://indianexpress.com/article/cities/delhi/second-gang-busted-in-5-months-4-held-in-noida-for-extorting-people-using-gay-dating-app-10016637/>.

beneficial, it must not be the *only* factor that should be considered.

- Online safety frameworks should incentivise good design practices in DSPs as a core means of achieving broader policy objectives.

5.4. Philosophy of Obligations

5.4.1. Delineating Safe Harbour Preconditions and Duty of Care

The Indian IT Act's current architecture is based on intermediary liability frameworks from the late 1990s and early 2000s, which aimed to promote open access to the internet and the growth of digital markets.¹⁹⁵ These frameworks also emerged partly in response to DSPs facing litigation over the speech and behaviour of their users.¹⁹⁶ Therefore, early digital regulation centred on safe harbour, shielding service providers from liability for individual user posts, which appear on their services in massive volumes and are near-impossible to pre-screen or monitor.

Under this model, service providers enjoy safe harbour from liability for individual pieces of illegal content, provided they lack actual knowledge of such content. This protection typically depends on meeting certain preconditions, such as acting expeditiously upon receiving official notice of illegal content. Some jurisdictions introduce further refinements. *The EU ECD, 2000, for instance, defines preconditions based on the type of intermediary, while the US CDA, 1996 includes a Good Samaritan clause that encourages voluntary content moderation without fear of losing safe harbour (see Chapter 4).*

Safe harbour protections were crucial in preventing DSPs from being overwhelmed by legal claims and from resorting to preemptive censorship to protect themselves. At that time, DSPs were largely viewed as **passive intermediaries**, not as services designed to actively encourage content dissemination or sustained user interaction. The early internet ecosystem was characterised by relatively limited, notice-board-like functionalities rather than the highly interactive, user-driven platforms that dominate today.¹⁹⁷

However, a lot has changed since. The internet has evolved from primarily functioning as a digital notice board with limited agency in content discovery, to platforms with complex algorithms and recommender systems that actively shape user experience. As a result, many now believe that DSPs are not just passive intermediaries but have greater control over how users experience their

¹⁹⁵ Schmon, Christoph, and Haley Pedersen. 2022. "Platform Liability Trends Around the Globe: From Safe Harbors to Increased Responsibility." Electronic Frontier Foundation. https://www.eff.org/tr/deeplinks/2022/05/platform-liability-trends-around-globe-safe-harbors-increased-responsibility?utm_source=chatgpt.com.

¹⁹⁶ Hoboken, Joris v., and Daphne Keller. 2019. "Design principle for intermediary liability laws." Transatlantic Working Group. https://www.ivir.nl/publicaties/download/Intermediary_liability_Oct_2019.pdf.

¹⁹⁷ Dhawan, Sachin. 2024. "How the World's First Major Safe Harbour Law Came to Be." The CCG Blog. <https://ccgnludelh.wordpress.com/2024/03/27/how-the-worlds-first-major-safe-harbour-law-came-to-be/>.

digital space.¹⁹⁸ These services have achieved significant scale and influence, making it increasingly important for them to take responsibility for keeping users safe.¹⁹⁹

Since the notice board analogy no longer holds strong, much like any other business that offers services to consumers, there is growing recognition that DSPs must discharge a **duty of care to protect their users from harm**. As noted in *Chapter 4*, different jurisdictions have specified different obligations for service providers to meet this duty of care. India, too, has recognised this duty by introducing proactive obligations under IT Rules in 2021. However, these obligations remain linked to safe harbour protections offered under the parent act. In practice, this means that if DSPs fail to fulfil their DoC obligations under the Rules, their safe harbour can be negated.²⁰⁰ This effectively reduces these platform-wide, proactive obligations geared towards outlining service provider responsibility to additional preconditions to safe harbour.

Treating DoC obligations as additional precondition to safe harbour protections creates both conceptual inconsistencies and practical limitations.

A. Conceptual Inconsistencies

- Safe harbour is associated with **third-party liability** for individual pieces of content. To claim safe harbour, DSPs only need to make content-level interventions such as reactively taking down illegal content upon official notification. Duty of care on the other hand is concerned with **service providers' proactive responsibility** (first party liability) to protect users from harm through **systems-level interventions**. Since safe harbour is claimed for individual pieces of content hosted by intermediaries, withdrawing it in response to **systems-level negligence** creates a **violation-remedy mismatch**. It overestimates the scope of a safe harbour framework in addressing harms that do not arise out of individual pieces of content, while offering no meaningful incentive to improve underlying systems. *For instance, widespread disinformation on social media platforms is not a mere summation of individual false posts but a product of systems-level failure to address how bad actors exploit functionalities (eg: open content feeds) and design (eg: built-in virality mechanisms).* In such cases, liability structures should incentivise service providers to improve their systems-level ability to uphold information integrity at scale.
- It is also important to acknowledge that while safe harbour preconditions apply only to illegal content, DoC obligations define service provider responsibility which often concerns content and conduct that may be **lawful but harmful**. Linking DoC

¹⁹⁸ Vidhi Centre for Legal Policy. (n.d.). "An Integrated Framework for the Study of Digital Platforms." Vidhi Centre for Legal Policy. https://vidhilegalpolicy.in/wp-content/uploads/2020/10/200924_Platforms-Report_Final.pdf

¹⁹⁹ Machado, Caio C., and Thais H. Aguiar. 2023. "Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care into Intermediary Liability Models." Cambridge University Press. <https://www.cambridge.org/core/journals/business-and-human-rights-journal/article/abs/emerging-regulations-on-content-moderation-and-misinformation-policies-of-online-media-platforms-accommodating-the-duty-of-care-into-intermediary-liability-models/26C58F>.

²⁰⁰ Press Trust of India. 2021. "Twitter Loses 'Safe Harbour' Shield in India over Non-Compliance to IT Rules." *Deccan Herald*, June 16, 2021. www.deccanherald.com/india/twitter-loses-safe-harbour-shield-in-india-over-non-compliance-to-it-rules-997970.html.

non-compliance to the loss of safe harbour risks framing online safety solely through the lens of the illegality of underlying content, without providing clarity on how services should address harmful speech and behaviour that does not cross the illegality threshold.

B. Practical unviability

- When DoC obligations are framed as additional preconditions to safe harbour, the only lever left for policymakers is to **weaken or dilute these protections**. This is a suboptimal and extreme step, not only because it can have serious free speech implications (see *Box below*), but also because it fails to address systems-level negligence. Since weakening safe harbour does not incentivise DSPs to proactively monitor patterns of online harms, the next step might be expanding the scope of what constitutes illegal speech – yet another extreme and sub-optimal measure.
- Diluting safe harbour protection would also open the floodgates to **unpredictable and inconsistent forms of liability** for service providers.

Consider two platforms that each fail to fulfil a safety obligation, such as publishing clear terms and conditions of usage or appointment of a local PoC. A platform with, say, a hundred isolated reports of CSAM might face criminal liability under the Indian POCSO Act, while another where health misinformation is rampant due to systems-level negligence could go unpunished owing to lack of clear illegality or missing penalty structures. This not only undermines consistency in enforcement but also fails to hold platforms accountable in a meaningful way.

Box: Why Linking Duty of Care to Safe Harbour Undermines Free Speech

Our consultations with experts revealed broad agreement on the necessity of preserving strong safe harbour protections for third party liability. Safe harbour is a crucial enabler of online free speech. If DSPs are perpetually at risk of losing their legal protections for user-generated content (UGC), they may adopt overly cautious measures and create overly censored digital spaces. Having said that, it was also clear that threats of safe harbor withdrawal/dilution creates strong incentives for platforms to comply (or indeed, over-correct) with regulatory mandates. We understand that the risk of DSPs' employees facing criminal action for user-generated illegal content presents a "red-line" that most DSPs take great care to avoid, even if it comes at the risk of suppressing legitimate speech. Therefore, while safe harbour withdrawal is a potent incentive for compliance, we believe it is a blunt instrument that simultaneously creates undesirable censorship in content moderation.²⁰¹

Even without the looming threat to safe harbour, content policies and moderation practices have displayed a tendency to disproportionately silence voices owing to lack of understanding of

²⁰¹ Devadasan, Vasudev. 2023. "Conceptualising India's Safe Harbour in the Era of Platform Governance." *Indian Journal of Law and Technology* 19 (1): 8-50. <https://repository.nls.ac.in/cgi/viewcontent.cgi?article=1420&context=ijlt>

cultural contexts.²⁰² Newer safety features, not fully optimised or filtered for errors, may also unfairly block speech. As platforms iron out their safety measures in rapidly evolving and risk-prone digital environments, regulation must ensure it does not unintentionally encourage unmindful censorship.

Our analysis points to the following key drivers of online safety obligations:

Modern regulatory approaches to platform liability must **distinguish between two fundamentally different types of violations – third-party content/conduct based violations and first-party DoC violations**. For the former, safe harbour should remain ironclad as long as the service provider has not violated any removal orders related to illegal content. In contrast, for first-party DoC violations that reflect systems-level negligence, a separate penalty structure should be articulated **based on severity and impact of the violation**. The intent of this structure should be to incentivise DSPs to internalise safety considerations into their design and operations.

5.4.2. Application of DoC Obligations under an Online Safety Framework

Building on our thinking that online safety risks primarily arise out of user interaction, DoC obligations from an online safety perspective should primarily focus on **service providers that enable user interaction**.²⁰³ While DSPs that do not offer user interaction services also have a duty to keep users safe, they are often more susceptible to sector-specific harms that are best addressed through sector-specific pathways – *see Box below*.

Further, not all services that facilitate user interaction have the same level of **oversight or control over user content, conduct and contact**. DoC obligations must therefore vary depending on the degree of *reasonable intervention* a service provider can potentially exercise. The primary objective of the classification framework therefore must be to group services into distinct classes in a way that ensures each class has **necessary but proportionate DoC obligations**.

To start with, bare minimum obligations needed to protect users - that can be deployed without causing unduly onerous burdens - can be broad based. *For example, an obligation to publish clear terms and conditions of service can serve as a basic safeguard for users and is easily implementable by all DSPs, whether or not they enable user interaction and therefore could be applied to the broader service categories.*

On the other hand, an obligation to take proactive steps to prevent exposure to specific kinds of abusive or harmful content, similar to Rule 3(1)(b) of the IT Rules, requires greater oversight and

²⁰² Wilson, Richard A., and Molly Land. 2021. "Hate Speech on Social Media: Content Moderation in Context." *Connecticut Law Review* 53 (1). https://digitalcommons.lib.uconn.edu/law_review/449/

²⁰³ While safety concerns traditionally emerge from user-to-user interactions, the advent of AI has created a novel category of standalone services that mimic human interaction (quasi-user interaction).

control over user-generated content. Therefore, even within the broader class of services that enable user interaction, the level of reasonable intervention for providers of different services could be different (also acknowledged in the IT Rules, which calls for “reasonable efforts” by intermediaries).²⁰⁴

Take ISPs, for instance. They enable user interaction by providing access infrastructure but have minimal visibility or control over what users consume or share online. Their role is closer to that of an electricity provider: essential to making things run, but not responsible for how people use the power. Subjecting ISPs to the same obligations as social media or messaging services would therefore be both disproportionate and ineffective in mitigating user harm.

A **narrower class** of services may thus be defined for those with greater control over user interaction.

Finally, obligations should adopt a **tech-sensitive approach** that recognises the differences in underlying technologies within the same service/functionality class. Despite functional similarity, differences in underlying technology can significantly impact the level of oversight and control available to service providers. *For instance, an end-to-end encrypted messaging platform is designed to ensure complete user privacy, including from the service provider. As a result, providers of encrypted services have theoretically no visibility into user interaction on their platforms, unlike providers of non-encrypted private messaging services. Therefore, obligations could look different for different underlying tech.*

While not all contexts justify exemptions or differentiated treatment based on underlying technology, some may. For instance, obligations that introduce friction in content dissemination, aimed at curbing the spread of problematic content on messaging platforms that allow wide-shareability, can serve as consistent safeguards across both encrypted and unencrypted platforms. However, obligations that require providers of encrypted services to monitor content or trace the origin of harmful material could undermine the very nature and intent of offering such services.

Finally, hard-coding regulations based on existing tech is detrimental to the overall growth of digital markets. It can disincentivise alternative technical approaches to providing the same functionality and limit service providers to status quo, even when there is clear market demand for innovation. *For instance, growing user demand for decentralised control over content practices has spurred the development of federated social platforms. These platforms are typically moderated by users or communities themselves and challenge traditional, centralised control over content moderation, requiring a reimagining of how content-related obligations are applied.* Therefore, our approach of setting end goals allows creation of multiple technical avenues and options for compliance, encouraging innovation rather than setting every new innovation on course for non-compliance.

²⁰⁴ Information Technology (Intermediary Guidelines) Rules, 2011, Rule 3(1) - Due Diligence by an Intermediary.

Box: Treatment of Sector Specific Harms and Sectoral Interventions

A horizontal online safety framework serves as a foundational “base layer” of regulation for the digital economy, establishing general principles and minimum standards for user safety and platform accountability across all digital services. This is especially important given the rapid integration of digital technologies across sectors and the emergence of new risks that cut across traditional regulatory boundaries. However, sector-specific risks-such as financial frauds, or consumer protection require tailored interventions reflecting the unique risks and operational realities of each sector.

When thinking of **harms that are specific to certain domains or harms emanating from digitalizing pre-existing “traditional services”** like banking, commerce, ride-hailing etc. that are already regulated by sectoral regulations, we believe that established sectoral regulators, like the RBI for digital payments or the Ministry of Consumer Affairs for E-commerce, are best placed to design and enforce targeted obligations. *For instance, the consumer protection regulations of the Ministry of Consumer Affairs that have been expanded to include specific rules for e-commerce²⁰⁵ or the Motor Vehicle Aggregators Guidelines, 2020²⁰⁶ issued by the Ministry of Road Transport and Highways.* These relevant ministries possess a better understanding of the risk profiles of these digital services (e-commerce and online cab aggregators in the highlighted examples) and the nature of intervention required, making a case for them to lead on the regulatory front. Existing sectoral regulations are already designed keeping in mind the unique considerations of that sector. *For instance, financial regulations have evolved to include stringent measures for consumer protection and fraud prevention.²⁰⁷*

The **role of a horizontal safety framework** is to, therefore, enable sectoral regulators to address harms within their domains, while insulating purely digital services (such as social media, video-calling, or cloud storage) from unnecessary overlap.²⁰⁸ At the same time, such a framework provides a common platform for regulators to coordinate on cross-cutting harms - like CSAM or disinformation - that require broad-based regulatory interventions in addition to sector-specific rules.

Legal experts consulted for this report also opined that the horizontal online safety legislation should focus on “core informational” risks (or, online safety risks, in the phraseology of this report), and that other hybrid or sectoral issues are better handled by sector-specific regulations. *For example, a general scope IT legislation may not be best suited to address niche or*

²⁰⁵ Consumer Protection (E-commerce) Rules, 2020 (India)

²⁰⁶ Motor Vehicle Aggregators: Guidelines under Motor Vehicles Act, 1988
https://morth.nic.in/sites/default/files/notifications_document/Motor%20Vehicle%20Aggregators27112020150046.pdf

²⁰⁷ For example, Reserve Bank of India's Notification on “Customer Protection – Limiting Liability of Customers in Unauthorised Electronic Banking Transactions”, RBI/2017-18/15DBR.No.Leg.BC.78/09.07.005/2017-18

²⁰⁸ Similar to the EU Rules on fair access to and use of data (Data Act), 2023. See, 2 ATIK, Can. 2023. “Horizontal intervention, sectoral challenges: Evaluating the data act's impact on agricultural data access puzzle in the emerging digital agriculture sector.” *Computer Law & Security Review* 51. <https://doi.org/10.1016/j.clsr.2023.105861>.

*specific phenomenon such as user-to-user messaging platforms being used for drug trafficking, which is better addressed under the jurisdiction of the narcotics law (the sector specific framework). In our assessment the horizontal and vertical regulations should be complimentary to each other to fill any legal vacuum.*²⁰⁹

By clearly delineating the scope of the horizontal law and sectoral interventions, our framework aims to minimize regulatory overlaps, gaps, and conflicts that can arise when multiple authorities claim jurisdiction over the same issues. It seeks to promote regulatory certainty for digital service providers, clarifying both their baseline obligations and any additional requirements from sectoral regulators.

5.4.3. Form and Approach to DoC Obligations

Duty of care can become an abstract concept in policymaking if obligations are not tailored to address specific, and often diverse, online safety risks.

As discussed in earlier sections, **systemic risks pose the greatest threats to user safety** in terms of both scale and severity. DoC obligations tailored to address these risks must therefore be adequately specific, **offering service providers clear, actionable guidance on minimising potential harm to users**. While the specific objective of preventing systemic risks makes these obligations outcome-oriented, it is important that the processes DSPs adopt to achieve this objective are grounded in stable principles, thus reducing ad hocism and unpredictability in policymaking.

This challenge can be mitigated by introducing standardization within the obligation framework – such as by **specifying categories or “families” of obligations that remain consistent across systemic risks**. For instance, one such family could be “transparency”, which requires service providers to share clear information about mitigation strategies adopted to address systemic risks, both well-understood and entirely novel (the Box below highlights the benefits for both DSPs and regulators of using a principles-based approach to address evolving systemic risks).

Finally, advanced DoC obligations should be grounded in a robust understanding of how specific functionalities and design features enable systemic risks, and which mitigation strategies are best suited to address them. This requires going beyond traditional transparency reporting to include deeper, more structured assessments. **Risk assessments** were intended to serve this purpose under the EU-DSA (see Chapter 4). Based on a review of the first round of risk assessments

²⁰⁹ For instance, in 2022, India's central bank introduced data localization guidelines, requiring all digital service providers in the payments ecosystem to store data locally, going beyond the data processing regulations then housed under the IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011. The guidelines aim to protect sensitive payment data (such as Permanent Account Number (PAN), Aadhaar Number, customer account details, etc.) by requiring such data to be stored in India or deleted from overseas systems within a strict timeframe of 24 hours if the data is processed outside India. While addressing such niche and specific risk profiles, it would be helpful to leverage existing regulatory structures to minimize overlap and confusion while ensuring user safety. See HT Tech: India's data localization decision leaves global digital payment firms in a bind, <https://tech.hindustantimes.com/tech/news/india-s-data-localisation-decision-leaves-global-digital-payment-firms-in-a-bind-story-gJv4p56ZmNFt84NxU8WE5I.html>; DSCI: Sectoral Privacy Guide - Banking, <https://www.dsci.in/sectoral-privacy-project/wp-content/uploads/2023/08/DSCI-Banking-Privacy-Guide.pdf>.

submitted by VLOPs, however, there appears to be broad consensus that this model faces critical shortcomings.²¹⁰ Despite increasing compliance requirements for DSPs, many risk assessments failed to fully capture the interplay between service design and harm, or to meaningfully evaluate the effectiveness of existing safeguards.²¹¹

Box: Potential Benefits of Introducing Standardized Families of Obligations	
For the Regulator	For Digital Service Providers
● Reduce scope of abstraction in designing countermeasures for individual systemic risks. ● Avoid unintended consequences related to novel solutions and the regulatory burden of “reinventing the wheel” each time. ● Prevent policy panic and helplessness when faced with new risks.	● Improved understanding of what type of voluntary safeguards may be expected <i>ex-ante</i> or what may be required <i>ex-post</i> to prevent systemic risks from developing. ● Predictable regulatory responses once a systemic risk is identified.

To be effective, risk assessments must be based on predefined parameters that make them measurable across comparable services. Mandating them as a standalone compliance, and as an end goal in itself risks reducing them to a checkbox without meaningful benefits, similar to the Indian experience with transparency reporting.²¹² As a result, our recommendations move away from mandatory risk assessments and mitigation plan submissions as a compliance goal. Having said that, we recognise the value of leveraging DSPs’ existing capacity to conduct risk assessments and develop mitigation plans.

To balance both these considerations, our recommendations propose the submission of **voluntary risk assessments and mitigation plans** as a way for DSPs to engage with the policymaking process, and form a key avenue in which industry stakeholders are able to highlight:

- What they believe the most pressing systemic risk are;
- Factors and contours of the risks, and how they manifest and persist;
- Countermeasures they see as most effective.

²¹⁰ DSA Civil Society Coordination Group. 2025. “Initial Analysis on the First Round of Risk Assessments Reports under the EU Digital Services Act,” An initiative of the DSA Civil Society Coordination Group.
<https://cdt.org/wp-content/uploads/2025/03/RA-Report-Assessment-Report.pdf>.

²¹¹ Kraemer, Jordan, and Tech Policy Press. 2024. “On Social Media, Transparency Reporting Is Anything But Transparent.”
<https://www.techpolicy.press/on-social-media-transparency-reporting-is-anything-but-transparent/>.

²¹² Kraemer, Jordan, and Tech Policy Press. 2024. “On Social Media, Transparency Reporting Is Anything But Transparent.”
<https://www.techpolicy.press/on-social-media-transparency-reporting-is-anything-but-transparent/>.

Such submissions would **incentivize realistic, detailed assessments** that can withstand stakeholder scrutiny, while also helping policymakers identify feasible countermeasures and understand the practical challenges faced by different classes of DSPs.

The following principles should guide the design of online safety obligations:

- DSP liability for third-party content is distinct from its own online safety responsibilities, and must be treated independently.
- DSPs that enable user interaction have a higher Duty of Care (DoC) towards users to keep them safe from online harms.
- Failure to meet DoC obligations reflects systems-level negligence and should be remedied through a separate set of penalties, and **not** by eliminating safe harbour immunity for third-party/user violations.
- DoC on different service providers manifest differently, and a well functioning online safety framework must articulate these clearly for each “class” of DSPs:
 - DSPs that enable user interaction have a higher duty than the ones that don’t;
 - DSPs that have greater oversight and control over these user interactions have higher DoC than the ones with lower degrees of control and oversight; and
 - Larger or more dominant services have a greater duty than relatively smaller ones given their scale of impact and influence.

5.5. How Best to Implement?

A resilient and adaptive digital safety framework relies on adequate **state capacity** to implement, monitor, and evolve regulation in step with the digital economy’s rapid changes. As detailed in our recommendations (*see Chapter 6*), this requires not only a dedicated regulatory institution but also structured processes that ensure accountability, transparency, and responsiveness.

A. Periodic assessments

Our framework entails identification of certain systemic risks for formulating additional responsibilities for high-risk services. As these **risks evolve**, periodic assessment will be essential for identifying emerging risks and re-assessing existing ones. Such a regular review cycle can bring **predictability and agility** to the regulatory process, allowing timely adaptation to new technologies, market practices, and harm vectors.

B. Stakeholder consultations

To ensure systematized and transparent policymaking, our framework envisages a consultative approach where regular formalized stakeholder consultations are a core part of the regulatory process.²¹³ These consultations, hardcoded into the framework, can facilitate systemic risk identification and review through structured engagement with the industry, civil society, technical experts, and other relevant actors. Such a **participatory approach** can ensure that diverse perspectives inform the review mechanism and the design of interventions, fostering legitimacy and buy-in.

That said, we acknowledge that tech policy issues can be complex and challenging, requiring **regulatory capacity for conducting high quality consultations** such as those conducted by OfCom on the UK-OSA.²¹⁴ These consultations entailed publication of detailed policy statements prior to consulting with a wide range of stakeholders including industry, civil society, as well as expert bodies and law enforcement agencies.

The Indian approach to public consultations has also evolved. Some Indian regulators like the RBI²¹⁵ and SEBI²¹⁶ have well-articulated processes for conducting in-depth stakeholder consultations through publication of discussion papers, impact analysis, etc. and inviting public comments before a policy is finalized. The future of Indian technology regulation should borrow from these evolved frameworks that set out systematic and well articulated procedural approaches to policymaking.

C. Capacity for risk assessment and targeted obligations

A **dedicated regulator** would have to undertake periodic assessments and organize stakeholder consultations. Permanent staffing and a focused mandate would be essential for building deep expertise and ensuring continuity in such regulatory oversight.

While the specific design, structure, and make-up of this regulator is outside the scope of this study, we use the term “regulator” as a placeholder for the competent authority that is tasked with implementing the Indian online safety framework. While there are different options in structuring such a body, we believe that such an entity must have the necessary resourcing, capacity, and independence in undertaking core policymaking and enforcement activities.

²¹³ Vijay L. Kelkar, Ajay Shah, *In Service of the Republic: The Art and Science of Economic Policy*, Penguin Random House, 2019

²¹⁴ Ofcom. “Consultation: Protecting People From Illegal Harms Online,” December 16, 2024.
<https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/protecting-people-from-illegal-content-online>.

²¹⁵ Dubey, Navneet. 2025. “RBI Introduces New Framing Rules; Public to Get Up to 21 Days to Provide Feedback on Draft Regulations.” *Fortune India*, May 8, 2025.
<https://www.fortuneindia.com/personal-finance/banking/rbi-introduces-new-framing-rules-public-to-get-up-to-21-days-to-provide-feedback-on-draft-regulations/122871>.

²¹⁶ Pti. “Sebi Notifies Procedure to Amend Rules; Mandates Public Consultation.” *The Economic Times*, February 17, 2025.
<https://economictimes.indiatimes.com/markets/stocks/news/sebi-notifies-procedure-to-amend-rules-mandates-public-consultation/articleshow/118328428.cms?from=mdr>.

D. Capacity to undertake investigations and enforcement

Similarly, the regulator would also need institutional capacity to enforce the provisions and mandates of an online safety framework. The ability to effectively investigate potential violations of attendant safety obligations as well as enforcing penalties is the necessary and indeed primary fulcrum which creates compliance incentives in DSPs.

This entails avenues through which violations can be detected on its own motion or pointed out by relevant stakeholders, the ability to assimilate such information into actionable and well framed cases for investigation, the power and technical prowess to summon the right kind of information that helps determine potential violations, and finally imposing and enforcing penalties where violations occur.

Modern online safety frameworks rely on formal investigations into potential DSP violations as the primary way to signal their seriousness to tackle priority issues, and create potent incentives to proactively mitigate risks being investigated.

Enforcement activity of course must respect principles of natural justice, and provide ample opportunities for the investigated DSP as well as other stakeholders to make submissions, as well as require well reasoned and public disclosure of decisions made as a part of the process. OfCom's Guidance on Enforcement under the UK-OSA²¹⁷ serves as a relevant example of processes and procedures that an investigation could follow.

An online safety framework's implementation plan should address:

- Periodic review of existing and emerging risks since online safety risks are dynamic and evolving. Periodic review brings predictability and agility to regulation by allowing timely adaptation to new technologies, market practices, and harm vectors.
- A participatory, multistakeholder consultative approach that ensures diverse perspectives inform the review mechanism and the design of interventions, fostering legitimacy and buy-in.
- A dedicated institution with adequate resourcing and capacity, and focused mandate for regulatory oversight - ranging from policymaking to enforcement.

²¹⁷ Ofcom, "Online Safety Enforcement Guidance" (Ofcom, December 16, 2024), accessed March 10, 2026, <https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/online-safety-enforcement-guidance.pdf?v=391925>.

06

Our Recommendations

Building a Future-Ready Online Safety Framework

This chapter sets out a blueprint for the next generation of online safety regulation in India, outlining a clear, predictable, risk-proportionate and future-ready framework. It supports policymakers in operationalising core principles for identifying systemic risks, classifying digital services and tailoring obligations to their functionality.

Understanding Online Safety Risks



Which online safety problems to prioritize?

Our analysis recognizes that online harms vary in severity, scale, pattern of occurrence, and the countermeasures needed to address them. Our recommended framework aims to provide avenues to identify online safety risks, while suggesting pathways to **address both sporadic and systemic risks** through proportionate policy responses. We also propose a standardized **three-part test** for identification of systemic risks.

Classification of Digital Services



Where should safety interventions be targeted?

From the larger digital economy, identifying the set of services (or functionalities) that should be subject to online safety mandates is a key challenge. We propose a **three-tier, functionality-centric classification structure for digital services** to guide regulation.

Philosophy of Obligations



What should the interventions look like?

An online safety framework must contain both a static set of baseline safety obligations while retaining the ability to nimbly adapt to emerging risks. We propose the adoption of a **functionality-specific, tech-agnostic obligation structure** encompassing **Baseline, Proactive, and Advanced obligations**. This structure is meant to achieve proportionality, provide clarity to DSPs on their responsibilities, and be flexible in responding to a dynamic risk landscape. Crucially, it also **differentiates between safe harbour preconditions and Duty of Care**.

Implementation Roadmap



How should this be implemented?

We outline a blueprint for periodic assessments and consultative implementation to identify emerging systemic risks, provide DSPs with the necessary flexibility, and enable the regulator to keep pace with technological change. We also recommend establishing a permanent, adequately resourced and independent regulator to implement the proposed framework.

6.1. Which problems to prioritize?

6.1.1. The Spectrum of Online Safety Issues

Moving towards a structured and future-ready digital governance regime entails developing a sustainable **regulatory philosophy that addresses the entire spectrum of online harms** with proportionate and effective countermeasures to ensure online safety, rather than an approach that responds merely to **isolated instances (or flash points) of online harms**.

Digital spaces, much like their physical counterparts, enable user interaction — primarily through three key facets: **content, conduct and contact** as discussed in Chapter 2. Most online activity can therefore be understood in terms of **what users say (speech), what users do (behaviour) and how users interact**.²¹⁸ Again, like the physical world, unregulated speech or behaviour can lead to harm, justifying the need for a legal and regulatory framework that mitigates the risk of their occurrence.

However, online harms are complex, vastly different from each other, and there are no clear answers as to which harms to prioritise over others. We begin by acknowledging that online harms exist on a multifaceted spectrum of **severity, scale/reach of impact, and pattern of occurrence**. However, most importantly, these species differ from each other in the **type of countermeasures and remedies** that must be deployed to effectively address them. Online harms can be understood as placing different types of demands on an online safety law. Some harms are complex and demand coordinated, multi-stakeholder responses, often requiring the law to incentivise changes in service design and DSP policies, while others are more straightforward and can be addressed through simpler interventions. *For instance, an individual post embodying illegal speech is perhaps best addressed by targeted content blocking measures, whereas individual content blocking is insufficient to solve information integrity issues such as widespread disinformation activity.*

Similarly, the same harm may also merit different responses based on its pattern of occurrence. *For instance, consider an isolated instance of abusive speech; this may be appropriately addressed through user reporting and grievance redressal. However, the policy response may have to be fundamentally different if abusive bullying was directed to children as a recurring/systematic pattern of conduct. In such a case, the issue shifts from individual content moderation to structural risk mitigation and platform-level safeguards.*

If policy frameworks are to be effective, they should build in suitable flexibility to deal with different parts of the online harm spectrum in ways that optimise their effect in each case. While sporadic or isolated instances could potentially be addressed reasonably using a static framework of policy options - like general guidance on moderation²¹⁹ and individual remedy mechanisms such as content blocking action or user complaints - problems that occur as a phenomenon in recurring

²¹⁸ World Economic Forum. 2023. "Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms." World Economic Forum, August, 2023, https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf.

²¹⁹ Ahead of 2024 Lok Sabha polls, MeitY issued an advisory to social media platforms to remove AI-generated deepfakes from their platforms to prevent misinformation spread and abuse, <https://economictimes.indiatimes.com/industry/telecom/telecom-policy/ahead-of-lok-sabha-election-government-asks-x-meta-to-re-move-ai-driven-deepfakes/articleshow/108292732.cms>.

patterns, require policymakers' toolkits to be expanded. This entails fashioning **advanced measures or advanced obligations that are tailor-made to suit significant and/or severe “systemic risks”**. As we argued in the preceding Chapter, these types of risks should form the core focus both for regulators and service providers when thinking of what *more* can be done.

6.1.2. What are Systemic Risks?

As distinguished from discrete and individual content-level violations, we refer to **broader and persistent trends** of harmful phenomena as “Systemic Risks”.²²⁰ Systemic risks focus not only on the harms themselves, but also on the underlying conditions, platform design choices, and operational practices that make such harms likely to emerge, recur, or escalate over time. This approach shifts the lens from isolated incidents to structural drivers and enables a more forward-looking assessment of when and how harmful outcomes may arise.

Systemic risks are significant in scale, reach, and impact. They may be dispersed across large segments of the population (scale), across multiple service providers (reach), and across different sectors (impact). As a result, they require regulatory responses that move beyond addressing individual instances of harm and instead target the systems that enable them.²²¹

For example, disinformation on social media cannot be meaningfully understood as merely a collection/summation of misleading posts. Falsehoods spread depending on how disinformation actors interact with other users and exploit platform functionalities²²² and affordances²²³ like feeds and messaging to disseminate narratives. Looking at individual falsehoods alone also misses how narratives are systematically amplified to shape opinion, making it hard to recognise their broader impact. It is more useful, therefore, to view disinformation as a systemic phenomenon - one that arises from multiple factors, including the very design of digital platforms.

Given this, regulatory responses, including those currently employed in India, that focus primarily on identifying and removing individual pieces of false content are likely to be inherently limited. India's digital regulatory approach has largely relied on reactive tools, especially content takedowns, which do little to address the structural conditions that enable harm to emerge and scale. Our framework recognises that systemic risks cannot be effectively mitigated through such approaches. Instead, they require **root-cause oriented interventions** that reshape platform design, architecture, and underlying incentives of different stakeholders. Each **systemic risk also**

²²⁰ While we refer to this category of online harms as “Systemic Risks”, modifying the nomenclature to any other expression that conveys their pervasive nature would not change any of our recommendations. For instance, this category could just as well be named “Priority Harms”, etc. We recognise that similar nomenclature is employed by the EU-DSA, but caution readers against substituting the EU-DSA's conception of systemic risks for the ones in these recommendations.

²²¹ Michele Loi, “Making Sense of the Digital Services Act: How to Define Platforms' Systemic Risks to Democracy,” AlgorithmWatch, August 2023, https://algorithmwatch.org/en/wp-content/uploads/2023/08/AlgorithmWatch_Risk_Assessment-DSA.pdf.

²²² Han Luo et al., “Spread of Misinformation in Social Networks: Analysis Based on Weibo Tweets,” *Security and Communication Networks* 2021 (December 16, 2021), <https://doi.org/10.1155/2021/7999760>.

²²³ See: Integrity Institute. n.d. “Political Misinformation.” Focus on Features: A Project of the Integrity Institute. <https://gbw-ii-fellowship.web.app/harm/political-misinfo>; and Integrity Institute. n.d. “Health Misinformation.” Focus on Features: A Project of the Integrity Institute. <https://gbw-ii-fellowship.web.app/harm/health-misinfo>.

demands a tailored regulatory response aligned with the nature of the harm and the platform features that drive it.

Accordingly, systemic risks form the basis for determining whether the online safety law should mandate that DSPs adopt advanced safety measures. By **introducing systemic risk as the threshold for triggering enhanced regulatory requirements**, our recommendations allow such legal obligations to be carefully calibrated and targeted based on the level of risk posed. This approach ensures that policy responses are guided by clearly defined and evidence/reason-based objectives and recurring patterns of harm, rather than ad hoc reactions to isolated incidents.

Having said the above, given the wide variety of online safety risks and the differing dimensions of harm they create, we believe it is **important to devise a principled method to identify those that require heightened regulatory attention**. Without a principled method to distinguish systemic risks from other kinds of risks, stakeholders concerned with online safety are likely to get overwhelmed by the volume and diversity of issues that surface each week, with no clear basis for prioritization. In the absence of such differentiation, limited regulatory attention and resources may be spread too thin or misdirected. At the same time, not every harm requires a novel regulatory response. The key is to develop a framework that helps assess which risks demand structural intervention and which can be addressed through existing tools.

6.1.3. How Do We Identify Systemic Risks?

The online safety laws we studied do in effect prioritize *some* harms over others.²²⁴ However, our research into these choices reveal that the selection of these harms were ad hoc and reflected the political priorities at the time of their formulation. While the harms prioritized by these laws are no doubt compelling, we argue that **any future-ready framework must set out a test or formula that explains the choice of priority harms**, in a way that can enable future selection or deselection as the digital landscape evolves. This is also salutary to constitutional promises of non-arbitrariness and benefits of rational and explainable policy development.

At present, there is no widely accepted mechanism that sets out how online safety risks/harms ought to be prioritized for regulatory attention. The approach we propose is based on our independent analysis of different legal and policy interventions in India and around the world, nature and patterns of online harms and the relationship between digital service functionalities and risks of online harm.

To assess whether the risk of an online harm is systemic in nature (and thus warrants higher degree of regulatory intervention under an online safety framework), we recommend applying a **conjunctive three-part test**.²²⁵

²²⁴ Article 34 of the EU-DSA lists a discrete list of systemic risks that it focuses on, the UK-OSA prioritises preventing exposure to illegal content and children's online safety, etc.

²²⁵ All three conditions must be met/satisfied for the risk to qualify as "systemic"

Test 1	Test 2	Test 3
<i>The Risk is widespread or systematic in its manifestation</i>  <i>Filters out sporadic and isolated harms, thereby justifying the need for regulatory intervention.</i>	<i>The Risk is caused or primarily exacerbated by digital services</i>  <i>Establishes that risk is rooted in platform design or digital governance, thereby justifying the need for intervention under an online safety framework.</i>	<i>Vertical regulation is not adequate or effective to address the Risk</i>  <i>Justifies the role of a general digital services regulator rather than a more narrow approach driven by sectoral regulators.²²⁶</i>

TEST 1: The risk is widespread or systematic in its manifestation.

In essence, this test assesses whether a risk either reaches substantial scale (widespread) or exhibits methodical recurrence (systematic), and filters out sporadic or isolated issues that should not form the sole basis of imposing obligations on an entire class of DSPs.

To satisfy this test, a given risk should either be widespread, or systematic, or both. The intention is to exclude isolated or one-off incidents that do not pose broader, long-term, high-impact or commonly persistent challenges linked to platform design or overall functioning.

We achieve this by either testing for:

- **Widespread harm potential** that is substantial in scale or reach, or crossing a normative *de minimis* threshold of numerosity and/or impact. In other words, this test filters out risks that are not felt by a substantial number of users (either as a general population, or a specific subset of users²²⁷),

OR

- **Systematic** describes harm that occurs in a methodical, patterned, organised or persistent manner.

Harms that satisfy Test 1 therefore either manifest on a large scale or are recurrent due to insufficient safeguards. It **establishes the necessity of regulation** by demonstrating that the

²²⁶ We recognize that from a legal and policy hygiene perspective, we must lean towards narrower and 'least-disruptive' options in policymaking, and therefore this third test creates an avenue for deliberate engagement on whether less-intrusive options from a sector-specific perspective are available that would achieve the same end.

²²⁷ For example, a risk will meet this threshold if it impacts a substantial number of children, or specific user groups such as LGBTQ+ users, social minorities etc. even though the affected users/group may only represent a smaller subset of *all* users in the country.

magnitude or severity of the harm exceeds what can reasonably be managed through existing practices deployed by service providers or organic market correction, making state action or regulatory intervention justified and proportionate.²²⁸

TEST 2: The risk is caused or primarily exacerbated by digital services.

In essence, this test requires policymakers to delve deeper into the root cause of a risk, assessing and establishing a relation between the risk and its occurrence over digital services.

This test identifies harms that are significantly heightened when facilitated or amplified through a digital service. In other words, a risk is not systemic simply because it manifests on digital services, however, it satisfies TEST 2 **IF** it is caused/exacerbated by affordances on digital services **or** made significantly more harmful *because* of its occurrence in the digital sphere. Hence, the risk in question should be causally and/or materially exacerbated by digital services - rather than simply being a broader socio-political phenomenon that spills over to the digital realm.²²⁹

In other words, this test identifies situations where digital services are not merely a passive site of risk occurrence, but are exploited to become **enablers, amplifiers, or vectors of risk**. *For instance, disinformation is not a new problem, but online disinformation is far more potent, prevalent and likely harmful than before because of the scale and speed at which it spreads.* Assessing this will require sensitivity to the **marginal** effect caused by digital services, which could be on account of network effects that act as a multiplier, inherent vulnerabilities in certain functionalities etc.

TEST 3: Vertical regulation is not adequate or effective to address the risk

This test is linked to the subject matter, the degree of specialisation and jurisdictional competence of regulators over most relevant stakeholders.

Test 3 identifies harms that are appropriate for intervention *under* an online safety framework. In other words, it enables policymakers to assess whether the solutions to a given risk fall within the scope of a horizontal online safety law, or whether the issue is already, or can be more effectively, governed under a different regulatory regime.

Not all widespread or systematic and technology-enabled risks (that pass Tests 1 and 2) warrant intervention under an online safety framework. **Such an intervention is typically justified only where vertical or purpose-built regulation does not adequately or effectively address the risk.** In making this determination, policymakers should also assess whether obligations under a

²²⁸ Our inspiration for the “widespread” or “systematic” test is rooted in international treaties where such thresholds help separate between sporadic activity and more deliberate, phenomenon-scale/nature incidents. While we acknowledge the difference in contexts where the test is applied, we found the core concept to be well suited and applicable to the purpose of Test#1 here. Having said that, the articulation of the test is merely indicative, and alternative articulations that serve the same purpose will also be compatible with our recommendations.

²²⁹ This does not imply that the “systemic risk” designation requires conclusive causal evidence, directly linking a platform to societal harms. Rather, a reasonable, well-grounded inference based on observed patterns, platform functionalities, and plausible risk dynamics should suffice. The standard here is not as high as legal causation, since we recognize extremely high thresholds can prevent or delay common-sense or good faith regulation.

horizontal online safety law would provide meaningful additional risk mitigation or merely impose compliance burden with limited **marginal benefit**.

*For instance, consumer right violations pose a widespread (often systematic) and tech-enabled risk to end users. However, in the Indian context, these harms are already addressed through a purpose-built consumer protection framework. In addition to the broader principles under the Consumer Protection Act, 2019, the Consumer Protection (E-Commerce) Rules, 2020 specify service-level obligations for e-commerce providers,²³⁰ such as requiring provision of information about refund, discount, warranty to help consumers make informed choices.²³¹ This eliminates the need for additional measures or duplication of obligations under an online safety framework. **Even as new risks emerge, it is important to assess whether existing purpose-built regulation can be adapted to address them.** For instance, while the rise of social commerce²³² introduces new risks, broader principles under the consumer protection framework may be adequate to address them.²³³*

In contrast, circulation of non-consensual intimate imagery (NCII) targeting women in particular is a risk that is both widespread²³⁴ and exacerbated by digital services²³⁵, and is only partially addressed by existing criminal laws.²³⁶ These frameworks are designed to address individual criminal liability for perpetrators, but are not purpose-built to (1) help survivors take immediate and effective steps to limit the spread of NCII, or (2) guide digital service providers in taking steps to prevent and limit the harm caused by NCII dissemination. As a result, criminal law is an inadequate and ineffective countermeasure to the broader phenomenon of NCII dissemination (though they nonetheless are necessary to penalise offenders). Indeed, this is also acknowledged under India's IT Act and Rules framework.²³⁷

²³⁰ The Consumer Protection Act, 2019, No. 35 of 2019, § 2(16) (India): "e-commerce" means buying or selling of goods or services including digital products over digital or electronic networks.

²³¹ Consumer Protection (E-Commerce) Rules, 2020, r. 5(3)(c), Gazette of India, pt. II, sec. 3(i) (India): e-commerce entities are required to provide information relating to return, refund, exchange, warranty and guarantee, delivery and shipment, modes of payment, and grievance redressal mechanism, and any other similar information which may be required by consumers to make informed decisions.

²³² Social media platforms enabling sale and purchase of goods and services. Here, the social media platform uses existing functionalities and features to enable digital commerce through shoppable posts, product tagging and in-app checkouts. Social commerce also includes influencer commerce which further complicates digital commerce, as consumers transition from intended purchases to influenced purchases.

<https://www.forbes.com/sites/jamiegutfreund/2025/04/13/from-e-commerce-to-social-commerce-the-influencer-effect/>.

²³³ Press Trust of India, "'Social Commerce Set for Growth, Concern of Authenticity Looms Large,'" *Business Standard*, April 7, 2024, https://www.business-standard.com/industry/news/social-commerce-set-for-growth-concern-of-authenticity-looms-large-124040700339_1.html.

²³⁴ Ranjana Kumari, "Tackling the Harms of Non-consensual Intimate Images | Hindustan Times," *Hindustan Times*, October 18, 2025, <https://www.hindustantimes.com/ht-insight/governance/tackling-the-harms-of-non-consensual-intimate-images-101760777390725.html>.

²³⁵ "Non-consensual Explicit Imagery," Focus on Features, n.d., <https://gbw-ii-fellowship.web.app/harm/ncEI>.

²³⁶ These laws include Bharatiya Nyaya Sanhita (BNS), 2023, the Indecent Representation of Women (Prohibition) Act, 1986, and Protection of Children from Sexual Offences Act (POCSO), 2012.

²³⁷ Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, r. 3(2)(b); Standard Operating Procedure to curtail dissemination of Non-Consensual Intimate Imagery (NCII) content Under Clause (b) of sub-rule (2) of rule 3 of the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, <https://www.meity.gov.in/static/uploads/2025/11/a2c9500ef5f8b62a43bfc68747de592d.pdf>.

Though it may be **difficult to articulate precise limits and thresholds** for the application of Test 3, it is nonetheless useful to treat the exercise as an inquiry into the subject matter of the risk. Many online safety risks will tend to overlap with other sectors, laws, and regimes, and the determination of whether *existing interventions are enough* will inevitably **entail a degree of subjective reasoning**. To that end, the application of this test should be informed by the findings of the previous two tests, reflecting the fact that these dimensions are interconnected and often overlapping. Even so, analysis under Test 3 remains valuable in determining the appropriate regulatory “fit” for countermeasures, taking into account the role of other competent authorities and the dominant character of the risk. Properly applied, this helps ensure that regulation is targeted and effective rather than duplicative or redundant.

Three-Part Test in Action

See Appendix II

The above three-part test provides a **regulatory threshold** for distinguishing systemic risks from other ordinary online risks and offers a justification for state intervention. In Appendix II, you’ll find a table that runs a few online harms through this three-part test to illustrate which risks are likely to be systemic, and which may not qualify. These examples are indicative, not exhaustive, and are intended to illustrate the kinds of risks that may warrant policy attention under this framework, and the type of arguments that may be made to justify their selection/or deselection as a systemic risk.

The table illustrates how the three-part test distinguishes between systemic risks, such as CSAM and disinformation, which may justify enhanced obligations under an online safety framework, and other harms, such as gaslighting, that do not meet the same threshold.

These assessments reflect our present understanding of how the risks manifest and may evolve over time. Harms that are not currently systemic may, in future, develop in ways that satisfy the conditions of a systemic risk. Conversely, risks that are systemic at a given moment may diminish in scale, relevance, or structural impact. The framework must therefore remain responsive to changing digital realities.

6.2. Where should safety interventions be targeted?

Not every segment of the digital economy can, or should, be expected to address every online safety concern. The ecosystem is vast and highly diverse, with few digital service providers offering identical services or interacting with users in the same way. Indeed, many DSPs operate largely in the background and remain effectively invisible to end users.

While it is clear that some DSPs need to do more to keep the users' digital experiences safe, the critical challenge for any online safety framework is to identify the "class" of DSPs that should be legally mandated to adopt stipulated safety measures.

Most online safety laws attempt to classify or categorise DSPs so that regulatory obligations and oversight can be targeted at those entities most relevant to the policy objective. This, however, is far from straightforward. As discussed in *Chapter 4*, comparative experience shows considerable variation in approach, with no single model emerging as clearly effective. Similar challenges have arisen in recent and proposed Indian legislation, including the Telecommunications Act and draft broadcasting laws, which have been criticised for employing broad and imprecise definitions that risk sweeping in a far wider range of digital services than is necessary or appropriate.^{238,239}

In this section, we propose a classification structure that categorises the broader digital economy into coherent "classes" that are designed to proportionately and effectively target policy interventions to only the most appropriate service providers.

In doing so, we steer away from the current approach under the IT Act which applies a largely uniform framework to all "intermediaries".²⁴⁰ By **classifying services in a modular, technology-neutral and functionality-specific manner**, we hope to prevent the over-regulation of low-risk services and the under-regulation of high-risk ones.

In practice, **DSPs would be responsible for self-classifying their services against the proposed classes**, based on the different functionalities they offer. This classification would then determine the safety obligations applicable to the DSP, subject to regulatory oversight and the ability of the regulator to review or challenge a DSP's classification where appropriate.

Our proposed framework introduces a functionality-based classification structure — i.e., it creates classes of digital services based on the functionalities they offer. We rely on a **functionality-based approach** since digital service functionality is a good proxy for underlying risk potential (see *Chapter 5*), thus enabling the grouping of services that share comparable types and levels of online safety risk. This, in turn, allows safety obligations to be calibrated and tailored to the specific risk profile of each class.

²³⁸ Rentala Chandrashekhar, "A Word of Advice on OTT and the Draft Telecom Bill," *The Hindu*, November 24, 2022, <https://www.thehindu.com/opinion/op-ed/a-word-of-advice-on-ott-and-the-draft-telecom-bill/article66179200.ece>.

²³⁹ Jitendra Awhad, "The Broadcasting Services Regulation Bill: A Threat to Digital Freedom and Innovation," *The Hindu*, March 16, 2025, <https://www.thehindu.com/opinion/the-broadcasting-services-regulation-bill-a-threat-to-digital-freedom-and-innovation/article69334761.ece>.

²⁴⁰ Information Technology Act, 2000, No. 21 of 2000, § 2(1)(w) (India) defines "intermediaries" broadly to mean "any person who on behalf of another person receives, stores or transmits that record or provides any service with respect to that record and includes telecom service providers, network service providers, internet service providers, web hosting service providers, search engines, online payment sites, online-auction sites, online market places and cyber cafes". Having said that, we acknowledge that MeitY has attempted to narrow this definition by introducing classes such as SMIs, SSIMs, etc under IT Rules, 2021.

Functionality-specific obligations when **applied using a modular approach** ensure that similar functionalities are subject to consistent rules, regardless of the format or **combination “bundles”** in which they are offered. This allows DSPs to allocate compliance resources efficiently and enables regulators to focus oversight on higher-risk functionalities, resulting in more targeted and effective regulation.

“Functionality-based” Approach in the Context of Bundled Services

Modern digital services provide multiple distinct functionalities, which can be understood as **“bundled services”**. *For example, one application or website can simultaneously offer messaging, public content sharing, and e-commerce services – each carrying different levels and types of risk.*

In our recommendation, the functionality-specific approach denotes regulatory requirements linked to individual functionalities, rather than viewing services i.e. applications or websites as an indivisible whole. **As a result, safety mandates linked to a given functionality would only apply to a DSP if and when it offers that functionality to users.**

Though this is commonly the intent behind most regulatory approaches, explicit clarity about the scope of certain safety mandates increases regulatory clarity for all stakeholders involved. This also helps create a clear *à la carte* style applicability to regulatory obligations which are determined based on specific functionalities each DSP chooses to offer to users, insulating DSPs that are not offering similar functionalities from mandates that may be impractical or inappropriate for their services.

*For instance, a social media service that provides **e-commerce** and **private messaging** functionalities alongside offering **public content feeds** will fulfill obligations associated with each of these respective categories. The implications of this recommendation will become clearer in the following section where the proposed structure is explained in greater detail.*

Our framework also introduces a **three-tier classification structure**. At each of the three tiers, the objective is to either exclude classified services from obligations that are unnecessary for them, or to vary the manner in which they apply to the different classes.

Ultimately, this yields a framework where light-touch and common-sense baseline obligations²⁴¹ apply to most services, and specific additional responsibilities are tailored to a narrower band of services most likely to pose systemic risks.

²⁴¹ Geared towards service provider accountability.

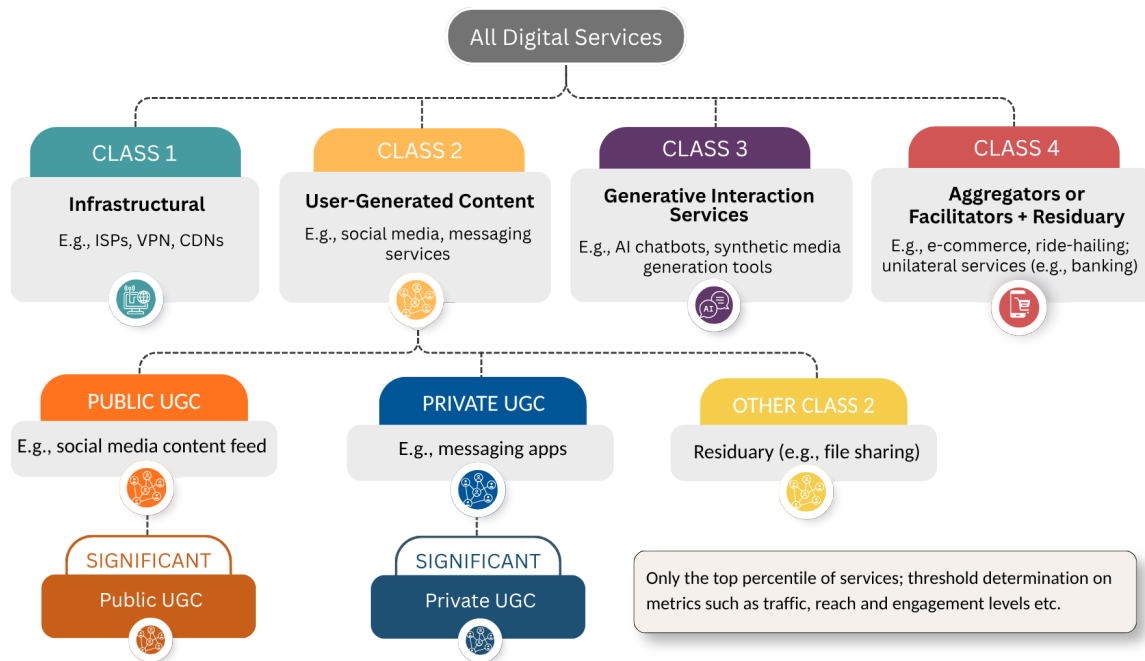


Figure: Proposed classification structure under the envisioned online safety framework

This is a technology-neutral, non-exclusive classification of services and NOT service providers. A single service provider may offer services across multiple categories.

6.2.1. The Classification Tiers

1. **Tier I:** The first tier of the classification framework aims to distinguish between four fundamentally distinct types of digital services on the **basis of the functionalities offered** and the **degree of control** they exercise over the content/interaction.
 - a. **Class 1 (Infrastructural):** Class 1 comprises all infrastructure services like conduit, caching and storage. These services are involved in **enabling transmission, routing or temporary storage of content**, without any control or editorial influence over content being transmitted. In simpler terms, the rudimentary intention in devising Class 1 is to encompass back-end services that are mostly “invisible” to users in their common everyday usage of digital services.

Examples include Internet Service Providers (ISPs), Domain Name System Providers, Content Delivery Networks (CDNs), cloud storage services and caching services for improved data retrieval.

The framework envisages **lighter obligations for infrastructure services**, recognising their limited direct interaction with end-users and seeking to avoid overburdening them with content/conduct/contact specific obligations they are not equipped to implement.

- b. **Class 2 (User-Generated Content)**: This category comprises services that **actively facilitate user interaction, particularly through the creation, sharing, and dissemination of content**.²⁴² The interaction element encompasses both one-to-one and one-to-many interactions between human users. Examples include social media platforms, dating services, video and audio sharing platforms, messaging applications, and video-calling services. Unlike infrastructure services, which simply transmit or temporarily store data without engaging with its content, these services allow users to actively engage with the service to create, modify, share, and interact with content. Today, these services exercise a **higher level of control** over users' digital experiences as compared to infrastructure services. They often moderate, recommend, or organize and curate content, influencing what users see or how they interact.

As these services enable (1) user-generated content, (2) user conduct and behaviour,²⁴³ and/or (3) user-to-user interaction or contact [for the purposes of this report we collectively refer to these as User-Generated Content, or UGC], they are **more prone to be the site of substantial safety risks**. This category includes both UGC services where content/other users are publicly discoverable (**Public UGC Services**) and private or closed-loop user interaction services (**Private UGC Services**). Other UGC services that do not fall within the public/private categories are classified as **"Other Class 2 Services"**.

Social media services such as X and Instagram are examples of Public UGC Services, while messaging services such as WhatsApp and Telegram are examples of Private UGC Services. Services that do not fit the spirit of any of these definitions such as document creation and editing softwares like Microsoft Word or Google Docs, fall within the category of "Other Class 2 Services".

- c. **Class 3 (Generative Interaction Services)**: This class comprises services (or features within a service) whose primary function is to **generate novel, first-party output** - including conversational responses, text, images, audio or video - in direct interaction with a user, rather than facilitating an exchange between users. This includes general-purpose AI chatbots (e.g. ChatGPT, Gemini, Claude), companion and persona-based conversational applications, and services that generate synthetic images, audio or video in response to user prompts. It also covers generative conversational AI features embedded within an otherwise Public UGC, Private UGC, or Other Class 2 service - for instance, an AI chat assistant built into a messaging app. Unlike passive content storage or creation tools (e.g. Google Docs, discussed under "Other Class 2 Services" above), which store or transmit user-authored content, Generative Interaction Services produce content themselves in response to user input.

²⁴² Also known as User-Generated Content (UGC) or user-driven content which is generally understood as any kind of media (text, data, image, audio, video, etc.) or action performed by online digital systems users, published and disseminated as an expressive or communicative effect either in an individual manner or combined with other contributions from the same or other sources. Examples include social media posts, messages exchanged on private chats, customer reviews, user profile on a dating app, avatar created by a user in an online game, etc. See <https://www.emerald.com/insight/content/doi/10.1108/oir-06-2020-0258/full/html>

²⁴³ E.g.: unhealthy addictive behaviour with short form video content etc. See Carr, Naomi, and Brittany Ferri. 2025. "Hijacked by Reels: Insights on Short-Video Addiction and Decision-Making." Recovered. <https://recovered.org/blog/short-video-addiction-impulsive-decision-making>.

The relevant test is not whether the AI feature is offered as a standalone product or bundled within a larger service, but whether the user is interacting directly with the AI. Where the AI itself is the user's primary interlocutor, it constitutes a Generative Interaction Service. Where AI merely supports interactions between users or between a user and content - such as through recommender systems, or automated moderation - it remains part of the host service and is governed by the obligations applicable to that service.

The resulting human-AI exchange is best understood as *quasi-user interaction*: it replicates certain experiential characteristics of user-to-user exchange closely enough (dialogue, personalisation, perceived relationship) to give rise to online safety harms such as inappropriate content, manipulative or emotionally destabilising outputs, and harmful advice - even though no second human user is present or involved.

In our proposed framework, Class 2 and Class 3 serve as the primary gateway for identifying services that warrant targeted regulatory attention, as most online safety risks stem from user-generated content and conduct, or from content generated by a service in direct interaction with a user. Accordingly, these services carry enhanced responsibilities (elaborated below): Proactive obligations apply by default across both classes, and Class 2 services additionally carry a higher likelihood of being subject to systemic risk-specific obligations.

How Do We Treat AI Within Our Framework?

AI raises regulatory questions that both include and extend beyond online safety. It creates two overlapping but distinct categories of regulatory risk: AI governance risks, such as algorithmic bias, explainability and training data practices; and online safety risks, such as suicide or self-harm encouragement, CSAM generation and fraud assistance.

The distinction depends on whether the risk arises from the **user-facing interaction** or the **underlying model**. Online safety governs what users experience when interacting with a service, including the outputs they receive, the safeguards applied, and the controls available to them. AI governance, by contrast, governs how the model was built, trained, evaluated, and made accountable.

Accordingly, an online safety framework should address AI-related risks only where they arise through direct user interaction and can be mitigated through service-level measures such as filtering, disclosures, or user controls etc. Risks whose solution lies in model development fall within AI governance, even if they manifest in individual outputs. **Online safety is therefore not a substitute for a comprehensive AI regulatory framework, which remains better suited to addressing broader questions of model development and accountability.**

Services such as ChatGPT and Gemini, as well as conversational AI features embedded within other platforms (for example, an AI chat assistant within a messaging app), enable users to interact directly with AI-generated outputs. These interactions can give rise to online safety harms such as exposure to inappropriate content, manipulative outputs, or harmful advice. As these risks arise through direct user interaction and can be mitigated through measures such as output filtering, age-appropriate design and disclosures, they fall within the scope of an online safety framework.

But AI is also increasingly embedded within digital platforms through functionalities such as recommender systems, automated moderation, and content ranking. These features may amplify misinformation, facilitate misuse, or increase users' exposure to harmful content. However, in these cases, AI supports the underlying service rather than constituting the primary user interaction. Accordingly, these functionalities do not form a separate service category and remain subject to the obligations applicable to the host service.

Under our proposed framework, conversational AI interfaces are classified as *Generative Interaction Services*, whether offered as standalone services or embedded within another digital service. Other AI-enabled functionalities remain part of the host service. This approach ensures that the online safety framework addresses user-facing harms that can be mitigated through service design, while broader questions relating to model development and AI governance remain within the scope of a dedicated AI regulatory framework.

- d. ***Class 4 (Aggregators or Facilitators)***: This class comprises services that function primarily as **aggregators or facilitators**, connecting end-users with service providers or offerings that are not principally designed for user-to-user interaction. This may include platforms that link users to third-party providers, as well as services that deliver offerings directly without enabling interpersonal exchange. *Typical examples include e-commerce platforms, travel booking applications, ride-hailing services, publishers of curated online content such as OTT streaming platforms, and app stores.*

This class would also cover the *advertising functionality* offered by social media, search engines and other DSPs; insofar as it connects businesses seeking to promote products and services with audiences who may be interested in them. Class 4 also encompasses **unilateral services** offered directly by the provider, such as banking applications or online games without the communication features (Wordle is an example of one such game).

What unites these services is the **limited ability of end-users to interact with one another or generate and share user-generated content**, thereby reducing the scope for users to inflict harm on each other through the service. Typically, risks arising from such services form a different “species” than the ones we commonly identify as online safety issues, and align closely with subjects that are native to the services’ specific sector. *For instance, with*

e-commerce platforms or online food delivery, the dominant issues revolve around consumer protection or food safety, respectively.

The sector-specific nature of the risks associated with these services makes them **more appropriately governed through vertical regulation by the relevant authority under specialised frameworks**, rather than through a general online safety regime. In several cases, mature regulatory frameworks may already exist, such as for payment services or marketplace platforms. In others, a dedicated regulator or tailored framework may still be evolving.²⁴⁴

Given this, it is only appropriate that these services be insulated from additional online safety obligations under our proposed framework. **In our recommendations, an online safety regulator will not be the lead regulator for these services.**²⁴⁵

Role of Sectoral Regulation for Certain Service Categories

The proposed framework intends to make the regulation of digital services more intuitive by acknowledging the role of sectoral regulation for certain service classes. Many services in **Class 4** are **traditional/analog services** like payment, food delivery, ride hailing, retail, etc. **simply being offered through a digital medium**. As these services primarily *digitize* activities that already fall within established domains of governance, it is more appropriate for such DSPs to be regulated principally by the relevant sector-specific authority. *For instance, e-commerce platforms fall under the oversight of the Ministry of Consumer Affairs and the Ministry of Commerce, and are subject to the Consumer Protection Rules. Similarly, payment service providers are governed by regulatory frameworks issued by the RBI.*

Accordingly, subject to baseline light-touch obligations (*outlined in the next section*), these services can continue to be regulated primarily under existing sectoral frameworks. This approach minimises overlap and reduces the risk of jurisdictional conflict between sectoral and technology regulators.²⁴⁶

We also acknowledge that, with growing digitalization of daily life, some services may be difficult to place within the four Tier I classes. For the framework to remain future-ready, it must provide for **residuary** or novel services - including those that may not exist currently.

²⁴⁴ The baseline safety obligations envisaged in this framework applicable to all digital services would help fill this gap.

²⁴⁵ This category, being home to vastly different services, holds potential for further sub-classification as the need may arise, though we have not ventured into the exercise given our focus on “online safety”.

²⁴⁶ For instance the recent Significant Digital Platform (SDP) framework of SEBI created confusion among digital platforms whether they are obligated to be notified as SDP by SEBI. See “Digital Platforms Have the Option to Be Notified as SDP: SEBI,” The Hindu, December 4, 2024, <https://thehindu.com/incoming/digital-platforms-have-the-option-to-be-notified-as-sdp-sebi/article68947689.ece>.

Therefore, under our proposed framework, Class 4 Services will also house all **other digital services** that are neither in the nature of infrastructure services (Class 1), UGC services (Class 2), generative interaction services (Class 3) nor aggregation or facilitation services (Class 4). *For instance, platforms that exclusively enable home automation do not fit within the proposed definitions, and may be treated as a residuary service falling within Class 4.* If in the future, there emerge any systemic risks associated with such services, the regulator is empowered to formulate additional responsibilities in that regard.

Since Class 4 Services are only subject to baseline accountability obligations, we believe this creates innovation friendly outcomes for emerging services (until the need for a specific classification arises).

Forward-looking Note on Agentic Services

In the context of agentic AI services - systems that take actions on a user's behalf (e.g. executing transactions, navigating third-party services, or performing multi-step tasks) rather than generating conversational output - the service's primary function is action-taking rather than content-generation. Its risk profile therefore resembles that of **Class 4 (facilitation of outcomes)** rather than Class 3 (first-party content generation in direct interaction with a user).

Given the early and evolving state of these products, we recommend housing agentic AI services in Class 4 for now. As with other Class 4 services, if in the future, there emerge any systemic online safety risks associated with such services, the regulator is empowered to formulate additional responsibilities in that regard.

2. **Tier II:** At this level, **we divide Class 2 (User-Generated Content) into three sub-categories** to account for the diversity of services it encompasses. For instance, Wikipedia, X and WhatsApp differ significantly in their functionalities, risk profiles and modes of interaction. Distinguishing between these services through defined sub-categories is therefore essential to designing meaningful and proportionate safety obligations.

However, we do not sub-divide Class 3 (Generative Interaction Services) at this stage. Unlike Class 2, which spans several distinct modes of interaction, Generative Interaction Services share a common structure: content is generated by the service in direct response to a user, and the resulting risks arise from that output. We therefore consider a single category appropriate at this stage, while leaving scope for further sub-categorisation if materially different risk profiles emerge over time.

For Class 2, we envisage the following sub-categories:

- a. The first sub-category, **Public UGC Services**, are the proverbial '**public squares**' of the internet. This category is intended to cover **publicly discoverable** content services that

enable creation and dissemination of UGC (like *social media feeds*). In the case of bundled services, it is possible that only a specific feature of the platform qualifies as Public UGC while other features fall within other classes.

*Classic examples of Public UGC services include public social media feeds such as Facebook, X/Twitter, Reddit etc. However, there are less obvious Public UGC functionalities that fall within this class. For instance, on dating and matchmaking platforms, only the explore feed displaying user profiles would be Public UGC, while the direct messaging feature would count as Private UGC (explained below). Similarly, navigation or map services that allow users to add locations and leave reviews would be Public UGC, while the navigation service itself would fall in the Class 4 category.*²⁴⁷

- b. The second sub-category, **Private UGC Services**, covers services that enable **direct, private interactions** between users - similar to conversations that take place behind closed doors in the real world. *Messaging apps, video calling services, private chats and email services would fall in this sub-category.* The messaging functionality offered by an otherwise Public UGC Service (like private chat on a dating platform or a social media platform) would be treated in this sub-category.

How are Public and Private UGC Services Different?

These sub-categories are treated differently given the **wide disparity in users' expectations** of what these services are to be used as, their risk profile (in terms of virality and reach) and the type of measures that can be applied to address the associated systemic risks.

The users of Private UGC Services enjoy a **higher degree of volitionality** in their exposure to certain safety risks associated with the service. Private UGC Service users typically **opt in** or **actively choose** whom to interact with, exercising control over their exposure and expecting a higher degree of privacy.

By contrast, Public UGC Services often expose users to content and interactions without any specific consent or positive act, making users more susceptible to risks such as virality and unsolicited harmful content. These foundational differences in user choice and privacy expectations shape not only the risk profiles of these sub-categories but also the types of regulatory and platform measures that can meaningfully address systemic risks associated with these services.

- c. The third and default residuary sub-category at this tier is the **"Other Class 2 Services"**. The rationale for creating a residuary category at Tier II is to ensure that other

²⁴⁷ Digital navigation services by themselves are in the nature of facilitation services that connect users to real-time mapping/location data and routing directions (specialized service) like other Class 4 services. In a way, these services offer traditional navigation support (location data, routing directions, etc.) through a digital medium.

content-related services like content storage and file sharing services (e.g. Google Workspace suite services like Documents, Sheets & Slides, DropBox, Truecaller, etc.) and content creation services (like Canva) are suitably identified and classified.

We emphasise the importance of creating a residuary class to avoid force-fitting services into the definitions of Public or Private UGC, where they do not naturally belong. In the absence of such a category, safer or lower-risk services may be inadvertently swept into more heavily regulated classes, undermining the intended proportionality of the framework.

An interesting example to consider is the Wikimedia v. Secretary of State litigation, where Wikipedia's providers argued that Wikipedia must not be subject to measures similar to social media platforms, where risks of illegal content discovery etc., are greater than on Wikipedia. They referred to the fundamental design differences that made Wikipedia a misfit to the category, though it technically met the criteria set by UK-OSA's classification design. This instance spotlights the need to make room for residuary classes to avoid unintended over-inclusions.

3. **Tier III:** At this tier, the more significant Public and Private UGC Services are identified as separate sub-categories on the **basis of size and scale**. This sub-classification reflects the fact that large-scale services, with extensive user bases and wide reach, can amplify content-related risks - such as the rapid spread of disinformation - and expose more users to harmful content and behaviour. Consequently, making dominant services safer can yield the most substantial macro-impact of online safety interventions (see Chapter 5).

The precise method for determining the size and scale of services is an important question, but a second-order one. While the detailed calibration of thresholds lies beyond the scope of this report, the objective is to identify a set of services that, taken together, account for the average user experience within a given functionality. The size and scale determination could rely on metrics such as user traffic, revenue or turnover attributable to the service, reach and engagement levels, with sufficient flexibility to account for local or demographic-specific patterns of use.

Is There a Case For a “Significant” Generative Interaction Services Sub-class?

Under our proposed obligations framework (detailed in the next section), online safety obligations are organized into three levels that scale with risk. **Baseline** obligations are minimum accountability requirements, such as publishing terms of service and maintaining grievance redressal mechanisms, that apply to all digital service providers. **Proactive** obligations are closed-list, non-discretionary design and disclosure floors that apply by default to a service category, in addition to baseline obligations. **Advanced** obligations are reserved for services operating at significant scale, where a regulator has identified specific risks that each warrant structural, architecture-level intervention or enhanced public transparency, and apply on top of both preceding levels.

The “Significant” designation for Public and Private UGC Services reflects the fact that scale fundamentally changes how harms arise. Public UGC Services become susceptible to virality, algorithmic amplification, and coordinated inauthentic behaviour, while Private UGC Services enable mass forwarding, network effects, and abuse that becomes systemic rather than purely interpersonal.

Generative Interaction Services do not exhibit the same user-to-user transmission dynamics. Their risks arise primarily from the model's outputs within individual user interactions, where each conversation is a discrete exchange between one user and the model. For most associated risks, such as unsafe advice and inappropriate content, each conversation carries a similar risk profile regardless of user count, and a targeted safety-by-design floor can meaningfully mitigate the harm.

This does not mean scale is entirely irrelevant. A flawed safety filter, a problematic model update, or a jailbreak vector can affect a large number of users, although some of these risks, particularly questions of model-level accountability and evaluation, fall within the remit of a dedicated AI governance framework rather than online safety.

Having said that, a risk warrants *advanced* obligations not only because it is serious or the category has grown large, but where a targeted *Proactive* floor proves insufficient to address the underlying harm mechanism. For Generative Interaction Services, such floors include output safeguards, disclosure that a user is interacting with an AI system, and controls for users reasonably identifiable as minors, elaborated further in the obligations outlined later.

Weighed against these considerations, we do not consider the case for a distinct “Significant Generative Interaction Services” sub-class made out today. This is an assessment of current market and risk conditions, not a structural limit on the framework: should risks of this kind become more clearly evidenced, advanced obligations may be extended to Generative Interaction Services where compelling reasons exist, and the regulator is empowered to formulate a dedicated sub-class and the corresponding obligations through the periodic review process described later in this chapter.

***Note:** While formulating this classification structure, we acknowledge that some services may not fit neatly into the defined classes. The goal of the framework is not to create perfectly airtight categories, but to provide clear guidance on stakeholder expectations and the responsibilities of digital service providers. Where a service may plausibly fall within two or more classes, some degree of judgment will inevitably be required. The structure we propose is designed to narrow the scope of such subjectivity and provide a principled basis for safety-oriented reasoning when resolving borderline cases. It is also designed to preserve regulatory flexibility, allowing new classes or sub-categories to be introduced where emerging services do not fit existing categories, or where new forms of systemic risk require distinct and calibrated obligations.*

6.3. What should the interventions look like?

At its core, our approach to DSPs' obligations recognizes that **different risks warrant different safeguards**. Equally, we also recognise that the level of safety-oriented obligations DSPs are subject to must be **proportionate** and well matched to the DSPs **risk profile and technology architecture**. Further, we recognise that in many cases, DSPs already have a market incentive to build services that are perceived by their user base as safe and trustworthy. However, there are also instances where service design and/or commercial incentives misalign with safety and wellbeing considerations. The role of regulation is to recognise these instances and introduce obligations that offset the misalignment.

In this sense, regulatory intervention gives concrete expression to a service provider's duty of care toward its users. Because regulation is meant to operationalize this otherwise abstract "duty", it is important to identify and articulate its distinct dimensions:

- In the first instance, by virtue of being service providers serving users, DSPs have a fundamental **duty of accountability** - which entails providing clarity about the terms of service provision, and responsibly addressing/responding to user and regulatory complaints.
- Further, given that user interaction - including direct interaction with a conversational AI system - can expose individuals to harm, DSPs have a **duty to safeguard against foreseeable harms**, including by filtering/moderating illegal or abusive content and conduct.
- Finally, in some cases (as explored below), **DSPs may have a duty to protect users from systemic risks**. How DSPs discharge this duty will vary depending on the nature of the systemic risk and the specific functionalities the service offers. Given this complexity, the measures that operationalize this heightened responsibility in relation to systemic risks are described in this report as "advanced obligations".

Accordingly, the model proposes three distinct "levels" of safety measures that an online safety law should contemplate. Drawing on the classification structure outlined in the previous section, **we adopt a tiered framework** in which obligations are calibrated and assigned according to the relevant tier of service.

6.3.1. How are Obligations Matched to Service Tiers?

Baseline <> All Services

Within the proposed three-tier model, we recommend **applying "Baseline" obligations to all services - Class 1, 2, 3 and 4** - to ensure the required minimum accountability of all digital services. These obligations are intended to be light-touch and focused on foundational transparency and user protection measures. At this level, services should be required to publish

clear terms of use, establish accessible grievance redressal mechanisms, and ensure direct channels of accountability to users. Baseline obligations may also include enabling compliance with lawful blocking orders and designating a point of contact to coordinate with competent authorities.

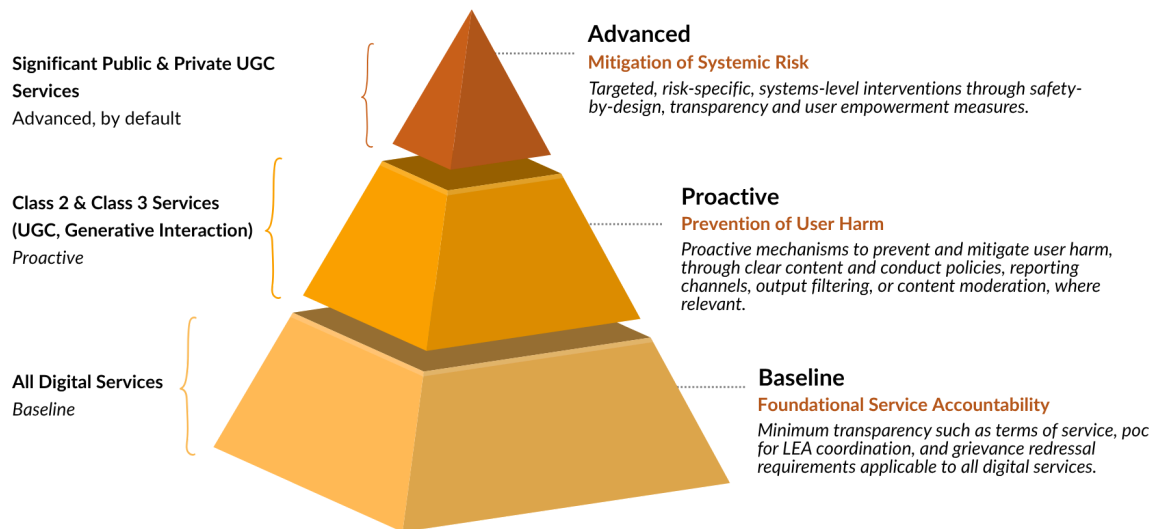


Figure: Three-level framework for risk-specific online safety obligations

Proactive <> All Class 2 and Class 3 Services

(applied in addition to Baseline obligations)

At the second tier (which is essentially all Class 2 services - Public UGC, Private UGC and Other Class 2 services, together with Class 3 Generative Interaction Services), obligations focus on **preventing user harm** whether inflicted by other users or arising from a service's AI-generated outputs, through proactive measures such as safer design, content moderation, and risk mitigation. **“Proactive” obligations are closed-list, non-discretionary product-safety floors**, confined to Class 2 and 3 services, rather than being applied across all services because of their defining role in enabling user interaction.

Rather than a single undifferentiated obligation set, we recommend that Proactive obligations be organised into four tracks corresponding to Public UGC, Private UGC, Generative Interaction Services, and Other Class 2 Services, reflecting how each sub-category’s characteristic risk pathway differs (see Table below). At a general level, these may include obligations to publish clear content and conduct policies, provide simple and “accessible” mechanisms for users to report content or address the specific harms characteristic of that sub-category, and take “reasonable” steps to prevent online harms as defined under the law.

While regulatory guidance is necessary to clarify what constitutes “reasonable” steps or an “accessible” reporting mechanism, overly prescriptive requirements should be avoided, as they may fail to account for differences in how risks manifest across services. For instance, the 2025 amendments to the IT Rules, 2021 sought to address misleading synthetic content and represented a timely regulatory response to emerging risks. While the final rules refined aspects of the original proposal and addressed many concerns flagged by the industry, the earlier draft illustrates an important design concern. It had required Significant Social Media Intermediaries to apply labels covering at least 10% of the visual or audio area of synthetic content, a requirement widely viewed as disproportionate and operationally prescriptive.²⁴⁸ It also did not fully account for the limitations of available detection technologies, making strict compliance challenging in practice.

This example highlights a broader issue: prescriptive, format-specific mandates can create feasibility concerns without necessarily improving safety outcomes. A more resilient approach would define the regulatory objective, such as requiring clear and prominent labelling of synthetic content (as ultimately reflected in the final rules) while allowing DSPs flexibility in how they meet that standard. Outcome-oriented obligations are better aligned with rapidly evolving technologies and reduce the risk of rigid, format-specific mandates becoming impractical or obsolete.

Advanced <> Significant Public and Private UGC (applied in addition to Baseline and Proactive obligations)

Finally, at the highest level, **“Advanced” obligations target specific systemic risks.** Because different risks require different countermeasures, these obligations will vary by both the type of systemic risk and the class of DSP.

For example,

- measures to address the systemic risk of disinformation will differ from those for CSAM, and,
- interventions appropriate for a given systemic risk will differ between Significant Public UGC Services and Significant Private UGC Services, reflecting the distinct risk profiles and modes of interaction inherent to each.

These obligations are designed to address complex, systemic risk phenomena and are therefore likely to require systems-level interventions, with correspondingly higher compliance and resource implications for DSPs. If imposed indiscriminately, they may impose disproportionate burdens and produce counterproductive effects for the digital ecosystem. Accordingly, we recommend that **such obligations apply by default only to Significant Public UGC and Significant Private UGC services**, where mitigation efforts are likely to yield the greatest net benefit given their scale, reach, and impact. Larger providers are also generally better resourced to meet

²⁴⁸ Draft Rule 3(3), Due Diligence in Relation to Synthetically Generated Information, Proposed Amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, in relation to Synthetically Generated Information (India, 2025). <https://www.meity.gov.in/static/uploads/2025/10/8e40cdd134cd92dd783a37556428c370.pdf>.

elevated compliance standards and, in doing so, can establish industry benchmarks that shape broader market practice.

That said, we recognise that smaller services can also pose systemic risks. Accordingly, where compelling reasons exist, the framework allows for these obligations to be extended to other service classes (*Public UGC, Private UGC, Generative Interaction Services or Other Class 2 services*).

What do the advanced obligations look like?

Advanced obligations are designed to address specific systemic risks and will often need to be tailored to the nature of each risk. At the same time, regulatory frameworks must avoid policy vacuums when new or evolving risks emerge, particularly in environments where institutional capacity and policy development timelines are constrained.

To balance flexibility with preparedness, we propose that the online safety law **pre-identify broad “families” of advanced obligations** from which context-specific interventions can be drawn. These can function as a structured toolkit for policymakers, enabling calibrated responses without requiring a redesign of the regulatory architecture each time a new systemic risk arises.

Based on our assessment of industry and regulatory best practices, as well as recurring themes in leading trust and safety literature, we identify **three core families of advanced obligations**:

- Safety-by-design measures
- Transparency obligations
- User empowerment mechanisms

These families can operate as foundational building blocks. They can be used to provide principled guardrails while allowing flexibility in application, thus reducing the risk of overly prescriptive or technically unworkable mandates. At the same time, they can **offer greater regulatory predictability for industry** by signalling the types of interventions that may be deployed as new systemic risks emerge. *Details on these three families are presented in the next section.*

6.3.2. Structure and Design of Obligations

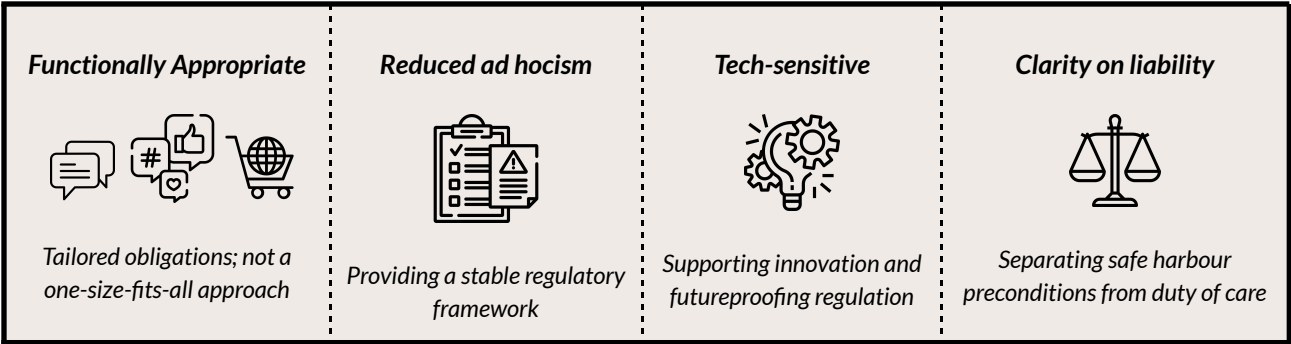


Figure. Principles underpinning our proposed obligations

All online safety obligations, including advanced obligations, should support a proportionate and stable regulatory framework for both regulators and DSPs. This requires ensuring that obligations are neither overly prescriptive nor rigid, but instead clearly signal to DSPs what their duty of care entails across different risk contexts.

Moreover, where a DSP is found to have breached its duty of care (*following a fair investigative process, outlined in the next section*), the appropriate consequence should be the imposition of monetary penalties or mandatory compliance directions. This approach is preferable to withdrawing existing liability protections, such as safe harbour provisions, which should continue to apply to DSPs hosting third-party content.

Finally, safety challenges persist and may even be amplified on services that rely on **newer technologies** (see Chapter 5). The tech-agnostic classification framework proposed in this report provides a proactive way to address these risks. By clearly defining service or functionality-level obligations regardless of the underlying technology, it **signals to innovators the baseline objective of safeguards required when offering equivalent services**. It also provides reasonable guidance to regulators by indicating what end-goals they should expect out of newer innovations. This establishes necessary guardrails to prevent policy vacuums where innovation outpaces regulation.

For example, comparable messaging services should be subject to equivalent reporting and redress outcomes, even if their technical designs differ. In a conventional messaging service, this may involve the platform's ability to access reported content for moderation and enforcement. In an end-to-end encrypted service, where routine access to message content is not available, the same safety objective may be met through alternative mechanisms, such as user-initiated reporting tools that allow flagged messages, metadata, or relevant contextual signals to be shared for review.²⁴⁹

The obligation is therefore anchored in the safety outcome, not the technical architecture. Different systems may rely on different mechanisms, but comparable services must meet equivalent baseline standards of protection.

Based on these principles, the table that follows covers an illustrative list of baseline, proactive and advanced obligations, which may apply across different tiers of digital services.

²⁴⁹ Seny Kamara, Mallory Knodel, Emma Llansó, Greg Nojeim, Lucy Qin, Dhanaraj Thakur, and Caitlin Vogus, *Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems* (Washington, DC: Center for Democracy & Technology, August 2021), <https://cdt.org/insights/report-outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/>.

Table: Illustrative list of obligations

Baseline Obligations <i>Designed to ensure service provider accountability</i>	
Services	Obligation
All DSPs	DSPs must publish clear and accessible terms and conditions governing the use of the service, including the rights and responsibilities of users, and notify users of any material changes.
	DSPs must put in place grievance redressal mechanisms and respond to users within a reasonable period . ²⁵⁰
	DSPs must restrict access to content upon obtaining actual knowledge or being notified by the regulator. <i>[This represents the existing design of lawful content blocking under Section 69A r/w 79, IT Act 2000]</i>
	DSPs must prominently display the contact details of designated PoCs responsible for coordination with regulators on their website, app, and other relevant platforms.
Proactive Obligations <i>Designed for prevention of user harm (applied in addition to baseline obligations)</i>	
Services	Obligation
Class 2: Public UGC Services	DSPs must publish and maintain clear safety policies explaining how they identify, assess, and respond to prohibited content, harmful conduct, and other safety risks on the service.
	DSPs must implement accessible and effective mechanisms that enable users to report content or conduct, and to control their public exposure - including blocking or muting other users, managing who may interact with or engage with their content, and limiting the discoverability of their content or profile, where applicable. Where a user is reasonably identifiable as a minor ²⁵¹ , these protections, including restrictions on unsolicited contact from unconnected users and reduced discoverability, must apply by default.
	DSPs must take reasonable steps to prevent the spread of online harms, which could include moderation and content-ranking practices.

²⁵⁰ What constitutes a reasonable period is beyond the scope of this report.

²⁵¹ The framework specifically includes obligations relating to minors because this category can be identified through a legally recognised status (age) or observable in-context signals, without requiring DSPs to infer other sensitive or protected characteristics of users. By contrast, other harms arising from user conduct are addressed through the general reporting, moderation, and enforcement obligations set out in the table.

Table: Illustrative list of obligations

Proactive Obligations <i>Designed for prevention of user harm (applied in addition to baseline obligations)</i>	
Services	Obligation
Class 2: Private UGC Services	DSPs must publish and maintain clear safety policies explaining how they identify, assess, and respond to prohibited content, harmful conduct, and other safety risks on the service.
	DSPs must implement accessible and effective mechanisms that enable users to report content or conduct, and to control who may initiate or continue communication with them - including blocking or restricting specific users, limiting unsolicited messages or contact requests, and managing participation in group communications, where applicable. Where a user is reasonably identifiable as a minor , protections against unsolicited contact and unconsented group additions must apply by default.
	DSPs must take reasonable steps to prevent the spread of online harms, calibrated to the reduced visibility DSPs have into private communications; this may rely on user-initiated reporting rather than on proactive scanning.
Class 2: Other Services	DSPs must publish clear policies on permitted and prohibited uses of the service. Where the service enables users to share, publish, or expose content to others, DSPs must additionally explain how misuse will be identified and addressed, provide accessible mechanisms for users to report it, and take reasonable steps to address the harms it enables, proportionate to the extent of such sharing or exposure.
Class 3: Generative Interaction Services	DSPs must publish clear interaction and output policies , specifically outlining the categories of prohibited or restricted AI-generated output. DSPs must also clearly and conspicuously disclose that a user is interacting with an AI system.
	DSPs must implement reasonable output safeguards , such as filtering or refusal mechanisms for clearly high-risk output categories (e.g. facilitating self-harm, sexual content involving minors, or instructions for inflicting serious harm). Where a user is reasonably identifiable as a minor , DSPs must take reasonable measures to limit design features known to drive emotional dependency or prolonged engagement, while providing session-limit and guardian-accessible controls.
	DSPs must implement accessible and effective mechanisms for users to report or flag a specific AI-generated response as harmful, unsafe, inaccurate, or inappropriate. Where user reports reveal a pattern of recurring harmful outputs despite proactive safety measures, DSPs must take reasonable steps to investigate and remediate the issue.

Table: Illustrative list of obligations

Advanced Obligations <i>Designed for mitigation of systemic risks (applied in addition to baseline and proactive obligations)</i>			
Services	Obligation		
Significant Public UGC & Significant Private UGC	Safety by Design Mandates <i>Ensuring that platform design and practices inherently mitigate identified risks and incorporate appropriate measures.</i>	Transparency Mandates <i>Releasing public and regulatory disclosures on specific risk occurrences and efficacy of countermeasures.</i>	User Empowerment Mandates <i>Providing users greater control over their experiences to avoid and/or report harms arising out of systemic risks.</i>

Rationale Underlying the Advanced Obligations

With respect to the advanced obligations, the table below elaborates the rationale underlying the three families of obligations and illustrates how each addresses a distinct regulatory gap, together forming a comprehensive framework to mitigate varied systemic risks. *An indicative application of this idea to disinformation as a systemic risk is set out in Appendix III.*

<i>Safety By Design</i>	<i>Transparency</i>	<i>User Empowerment</i>
Safety-by-design measures address structural aspects of service architecture that contribute to or amplify systemic risks, whether through unintended consequences or misaligned incentives. These obligations focus on modifying product features, default settings, amplification mechanisms, or interaction pathways that materially influence risk propagation. Such measures may require DSPs to redesign or constrain specific functionalities, introduce friction where appropriate, implement	Transparency obligations about risk manifestation and countermeasures reduce information asymmetry, helping build a shared understanding among different stakeholders on how risks manifest across services. By mandating disclosure of risk parameters and mitigation efforts, these obligations can strengthen accountability and create incentives for early, voluntary risk management before harms escalate.	User empowerment measures strengthen user agency in reducing exposure to systemic risks. These may include tools that allow users greater control over their experience , enabling them to limit or avoid certain categories of content, conduct, or contact. They may also include structured user reporting mechanisms that equip individuals to flag harmful speech or behaviour. Together, such measures function as distributed safeguards, allowing users to protect themselves while

built-in safeguards, or conduct risk assessments prior to deploying high-risk features. The objective is to mitigate risks at the architectural level rather than relying solely on downstream moderation.		contributing to broader risk detection and mitigation.
--	--	--

What would these look like? <i>Examples: Disinformation, Grooming, Non-Consensual Intimate Imagery (NCII)</i>		
Safety By Design	Transparency	User Empowerment
For example, to address the proliferation of disinformation, messaging services that enable forwarding may be required to introduce friction-based safeguards, such as labelling forwarded messages or limiting bulk, one-click forwarding to multiple recipients. The objective is not to prescribe fixed numerical thresholds, but to signal the types of design interventions that can meaningfully reduce virality risks. Similarly, where grooming is identified as a Systemic Risk, social media platforms may be required to avoid high-risk default settings, such as making children's accounts publicly accessible by default. Safety-by-design interventions could include restricting unsolicited adult-minor interactions, limiting discoverability of minor accounts, and reducing visibility of sensitive content to vulnerable users.²⁵²	For example, periodic transparency reporting on disinformation could move beyond aggregate removal statistics to provide granular insights into risk trends on a service. Such reports might identify high-risk categories, coordinated amplification patterns, or temporal spikes, for instance, surges in electoral disinformation during campaign periods. Analysing these patterns can guide targeted countermeasures by both service providers and regulators. Reports could also detail the measures platforms have already taken to address emerging problems and assess their effectiveness, providing insights for future interventions.	Users frequently encounter confusing reporting interfaces, unclear complaint categories, or multi-step flows that make it difficult to flag harmful content effectively. Such design friction can delay intervention and obscure the true prevalence of harm. Where a harm such as non-consensual intimate imagery (NCII) is identified as a systemic risk, DSPs could be required to provide clear, intuitive, and easily accessible reporting pathways tailored specifically to that category of harm. This would avoid forcing users to rely on vague options, such as "this made me uncomfortable", and instead ensure that NCII reports are distinctly classified, accurately recorded, appropriately prioritized, and routed for timely review and action.

²⁵² Social media platforms have a financial incentive to maximise user engagement and reach. However, applying stricter design choices to younger users' accounts, such as limiting visibility or interactions - can help counteract adverse incentives that may otherwise expose young users to harmful content and conduct. Instagram Teen Accounts provide a useful example. See Meta, "Instagram Teen Accounts," accessed March 11, 2026 <https://about.instagram.com/blog/announcements/instagram-teen-accounts>.

6.4. How should this be implemented?

The following sections explain the procedural aspects of the implementation of the proposed framework and the consequences of non-compliance.

6.4.1. Periodic Assessments and Consultative Implementation Process

One of the primary challenges in India's technology regulatory landscape is the **temporal unpredictability of regulatory action**. As newer technologies emerge and online risks become more visible, regulatory responses often originate from dispersed sources and lack a coherent, predictable framework.²⁵³

To address the negative consequences of regulatory unpredictability - a core objective of this framework - we propose embedding **recurring periodicity** into the assessment of systemic risks and the calibration of associated countermeasures, including advanced obligations. Our recommendations propose that the framework stipulate a cyclical and recurrent review/re-assessment process, that assures all stakeholders that the regulator will necessarily be open for engagement on online safety issues at predictable windows. This implementation approach is **participatory and consultative**, and creates structured avenues for input from industry, experts, civil society, and the public at each stage of operationalization.

Over time, these review cycles would strengthen both regulatory capacity and the broader policy ecosystem, enabling more effective implementation and compliance. Regular consultations, supported by reports, hearings, and submissions, would also help develop a coherent body of guiding principles, gradually establishing shared standards and best practices while reducing the risk of regulatory unpreparedness in a fast-evolving technological landscape.

India has, in other domains, benefited from cyclical and institutionalised policymaking processes. A prominent example is the GST Council, which meets regularly to deliberate on tax reforms and rate adjustments. While not every meeting produces major changes, the routine convening of the Council creates a structured forum for centre-state coordination, institutional learning, and capacity building, contributing to the smoother evolution of the GST framework.

Such periodicity and consultative approaches, however, **require a clearly designated implementing authority (ideally, an independent "regulator")**. This authority must be responsible for identifying systemic risks and conducting regular assessments to determine the appropriate advanced obligations applicable to relevant classes of DSPs.

²⁵³ Konark Bhandari, Ajay Kumar, Amlan Mohanty, Shatakratu Sahu, Arun K. Singh, Shruti Sharma, Tejas Bharadwaj, and Raj Shukla, "A Tech Policy Planning Guide for India—Beyond the First 100 Days," Carnegie Endowment for International Peace, November 2024, <https://carnegieendowment.org/research/2024/11/tech-policy-planning-guide-india>; see also Kahaan Mehta, "The Shein Comeback: A Masterclass in India's Ad hoc Tech Governance," *The Indian Express*, May 9, 2025, <https://indianexpress.com/article/opinion/columns/shein-come-back-digital-come-back-tiktok-ban-chinese-app-9992566/lite/>.

6.4.2. Who is the “Ideal” Regulator?

For a forward-looking framework such as the one being proposed here, the “ideal” regulator is one that is **permanent, independent**, and possesses **technical expertise** and **adjudicatory ability** required for effective implementation. This is critically important in the domain of online safety, which is uniquely complex from a regulatory perspective.

The rapidly evolving nature of the digital ecosystem requires that the regulator:

- **Maintain a continuous understanding of the risk landscape**, ensuring that its priorities remain relevant and evidence-based. This demands the capacity to monitor developments in the digital economy and a robust grasp of evolving technology architectures and platform design.
- **Design interventions that are both effective and proportionate** (as elaborated above). This, in turn, requires the ability to conduct regular and inclusive consultations, synthesize stakeholder inputs, and formulate strategies that reflect diverse perspectives.
- **Enforce the framework credibly and consistently**, ensuring that relevant DSPs comply with both the letter and spirit of the law. While enforcement capacity in any sector needs to be fairly robust, in the case of online safety, regulators have to be able to implement non-prescriptive rules on a uniquely diverse set of regulated entities. Effective enforcement therefore requires significant technical expertise, institutional capacity, and operational support to uphold the rule of law while advancing safety objectives.

These considerations reinforce our recommendation that implementation of an online safety law be entrusted to a permanent, well-resourced, and specialized body. A dedicated regulator would be better positioned to adapt regulatory strategy as risks evolve and technologies change.

An additional consideration is the depth and sensitivity of the dialogue such a regulator would inevitably conduct with stakeholders. Engagement would span early-stage policy development through to investigation and enforcement. Much of this dialogue would touch upon speech, politically salient online activity, and potentially confidential or proprietary information relating to DSP practices. The framework must therefore ensure that such exchanges are conducted rigorously, candidly, and constructively.

For this to function effectively, **the regulator must be institutionally independent from political and other external pressures**, fostering trust among all stakeholders. This implies that the regulator’s leadership should consist of full-time, independent experts who are not concurrently serving as functionaries within any branch of the State.

India has benefited immensely from investing in independent regulatory capabilities - such as with *SEBI (securities market regulator)* and *TRAI (telecom regulator)*, which represent some of the most evolved legal and regulatory infrastructure amongst the breadth of Indian legal landscape. We

believe similar benefits await to be reaped in the realm of online safety, when implemented by a specialized body.

6.4.3. Periodic Risk Review and Regulatory Update Process

In our vision, the regulator kicks off implementation of this framework by identifying systemic risks by applying the three-part test detailed earlier in this chapter. As a part of the identification exercise, the regulator may rely on DSPs transparency reports, multi-stakeholder consultations and in-house research and investigations. DSPs may also be encouraged to undertake **voluntary risk assessments** (including mitigation plans) to inform the risk identification process.

Each time this exercise is cyclically taken up, the regulator may:

- Identify **new systemic risks** that have emerged (with a brief statement on its satisfaction of the three-part test),
- **De-notify** previously identified systemic risks (where the risk vectors have ceased to satisfy the three-part test), or
- Decide that no modifications to the list of systemic risks in force is warranted.

Just as systemic risks are reviewed periodically, **the countermeasures that certain DSP classes may adopt (i.e., list of advanced obligations) to address these risks must also be subject to periodic review by the regulator.**²⁵⁴ This review should similarly be consultative and evidence-based, incorporating industry submissions, civil society feedback, transparency reporting, and investigative findings. Voluntary risk assessments and mitigation plans submitted by service providers can serve as structured inputs into this process, helping the regulator evaluate practical effectiveness and proportionality.

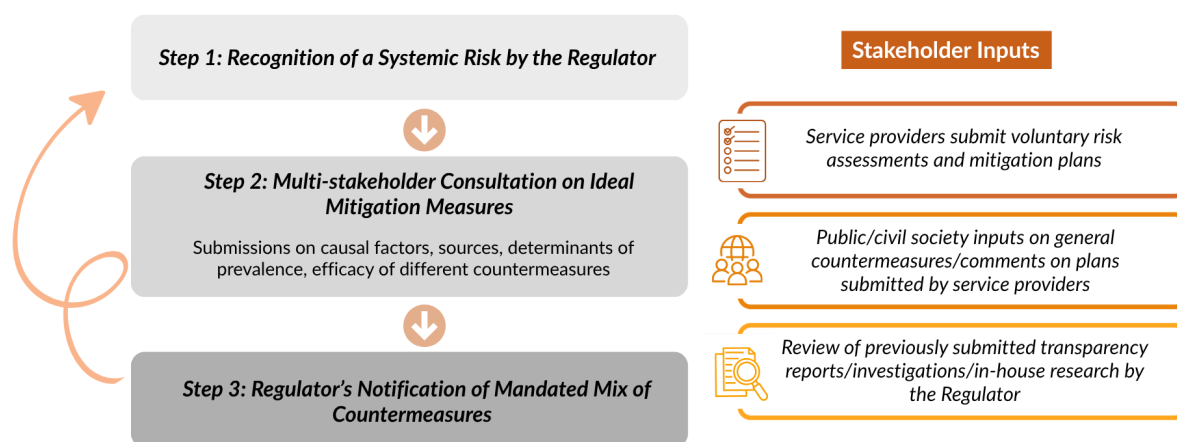


Figure: Systemic risk identification process

²⁵⁴ Either as part of the same review cycle as systemic risks, or an independent cycle with a phased gap.

Crucially, predictable and cyclical review exercises allow the regulator to incorporate new evidence, recalibrate obligations, and course-correct as risks evolve. This avoids locking the framework into static solutions that may become outdated in a rapidly changing technological environment.

6.4.4. Enforcement and Penalties in Case of Non-Compliance

1. Enforcement process

In a shift from the existing architecture of the IT Act, we propose a **penalty mechanism** for DSPs as the consequence of **failure to comply with obligations related to duty of care**. The regulator acting upon third party complaints, or on its own motion, may open a formal investigation by issuing show cause notices to DSPs for alleged non-compliance, resulting in penalties where complaints are borne out.

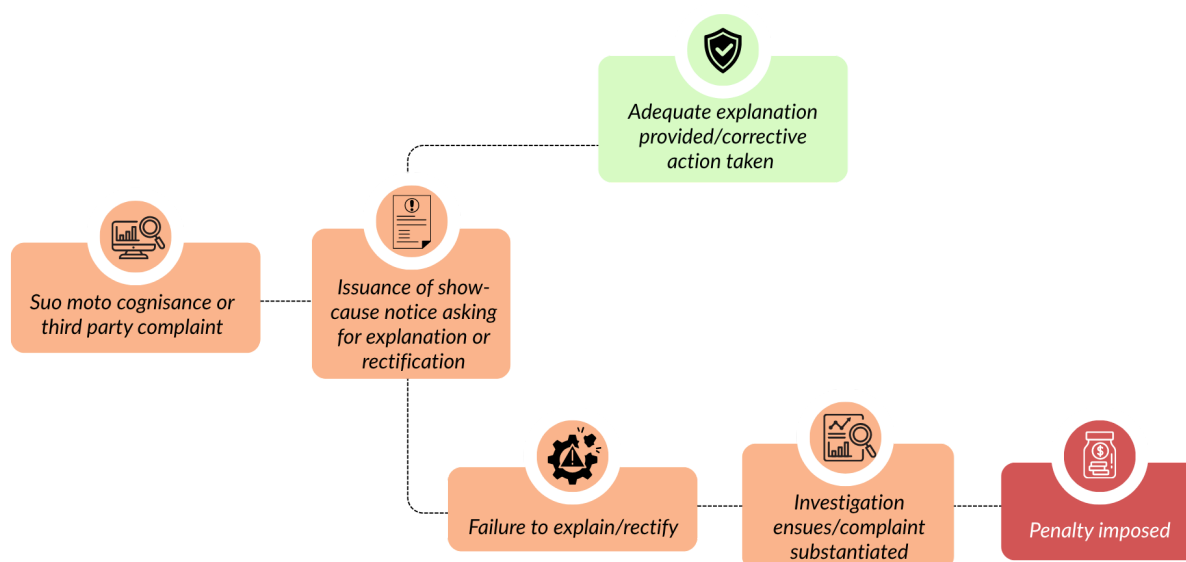


Figure: Enforcement process

We believe that complaints or suo motu action should be grounded in material that reasonably indicates both a DSP's non-compliance and the tangible impact of such practices in leaving a systemic risk insufficiently mitigated. This approach creates stronger incentives for meaningful and proportionate compliance that is conceptually aligned with the underlying lapse, **rather than resorting to withdrawal of safe harbour immunity for violations** that may be unrelated to the specific failure.

2. Enforcement philosophy

The enforcement approach is intended to be constructive and facilitative. Rather than relying primarily on punitive sanctions, the framework adopts a graduated response model. This creates

Explainer | How are countermeasures identified, operationalized and monitored?

01

Regulator opens consultations to identify systemic risks



Industry actors may submit voluntary risk assessments to assist in identifying which online risks rise to the level of systemic risks. Such assessments shall be publicly available to help reduce information asymmetry and provide greater insight into the feasibility and implementation of potential technical safeguards.



Civil society organizations, advocacy groups, academics, experts, and members of the public may review industry's voluntary risk assessments and submit additional evidence or material to inform the regulator's prioritisation of risks.

On examination of material submitted, the regulator notifies a list of systemic risks.

02

Regulator opens consultations to identify what advanced obligations should be specified to guide countermeasures for Significant Public and Private UGC Services



Actionable for relevant DSPs: At this stage, covered DSPs have the opportunity to recommend countermeasures that are best suited to their service classification. *For instance, providers of encrypted messaging services may highlight that their most effective interventions to address CSAM are user reporting mechanisms and assessment of unencrypted information such as group description and profile pictures.*



This is also an opportunity for civil society and trusted flaggers to highlight missing countermeasures in existing interventions deployed by DSPs.

03

Regulator notifies advanced obligations, comprising a list of safety measures deemed to be the expected standard of compliance for covered DSPs



At this stage, the regulator notifies the advanced obligations applicable to covered DSPs, setting out the expected standard of compliance. These obligations would be articulated within the three pre-defined families identified before (safety-by-design, transparency, and user empowerment). **The objective is neither to prescribe very rigid, feature-specific mandates nor to defer entirely to industry self-regulation.** The goal is to ensure regulatory clarity and enforceability while allowing flexibility in implementation.

The recent amendments to the IT Rules, 2021 relating to synthetically generated content (SGI) illustrate how such a balance may be achieved. Earlier drafts proposed highly prescriptive requirements (such as labels covering 10% of the surface area of the visual display). The final amendments instead require "prominent" labelling of synthetic content while avoiding rigid, feature-specific mandates. This approach preserves regulatory accountability while allowing service providers the flexibility to design and deploy appropriate technical solutions.

For both existing and newly identified systemic risks, the regulator should also specify a reasonable compliance timeline, calibrated to the complexity and resource implications of the obligations imposed.

This will lead covered DSPs to assess their countermeasures against the notified compliance standard, resulting in the following pathways and potential outcomes (in decreasing order of compliance).

DSP's practices align fully with the notified advanced obligations

A DSP's existing measures may already meet both the substance and intent of the notified obligation. This is a plausible outcome under the proposed model, given that regulatory standards would have been shaped through structured consultations and informed by prevailing industry practices. Where DSPs have implemented effective measures consistent with the notified obligation, they would be considered fully compliant.

Substantive alignment with the notified obligations (including outcome-equivalent alternatives)

A DSP's measures may either broadly align with the notified advanced obligations, or differ in form while achieving comparable safety outcomes. Given that advanced obligations are intended to provide principled guidance rather than rigid prescriptions, both situations should ordinarily be treated as compliant, provided the measures adopted are reasonably effective in mitigating the identified systemic risk.

Regulatory action should not be triggered merely because a DSP has implemented a different solution. However, where a credible complaint leads to an investigation, the DSP would have the opportunity to demonstrate that its safeguards are aligned in spirit with the notified obligation and achieve equivalent or superior risk mitigation outcomes. This would constitute a valid defence to an allegation of non-compliance.

This approach preserves flexibility for innovation and context-specific design, while maintaining accountability through outcome-based evaluation. It also accommodates cases where a DSP, though falling within a "significant" service class, can demonstrate that its design materially insulates it from contributing to a particular systemic risk. In such instances, proportionate differentiation is permitted, subject to evidentiary justification. *This can cover situations such as Wikipedia's challenge to its Category 1 designation under the UK Online Safety Act, arguing that requirements such as user verification were disproportionate given its existing safeguards against illegal content.* Such a model then allows DSPs that are confident in their posture to refrain from adopting specific measures, provided they can substantiate their position with evidence if subject to review or investigation.

DSP's practices are stated to be in alignment with the notified obligations but not effectively operationalized

Where a DSP represents that it has implemented a notified advanced obligation, it must ensure that the measure is meaningfully operationalized in practice. If significant deficiencies are observed in implementation, such that the effectiveness of the measure is materially undermined, the regulator may initiate a review to assess compliance. Isolated or minor errors, however, are not intended to trigger enforcement action.

For instance, where a listed advanced obligation for Significant Private UGCs is to "introduce meaningful friction in one-click message forwards" (to counter disinformation), and a DSP reduces reach limit for one-click forwards from 200,000 to 100,000, it is possible that third parties may submit material to the regulator demonstrating how this reduction has not contributed materially to mitigating systemic risk on the platform. When this complaint is investigated, the regulator may find the measures inadequate, even if the DSP has technically complied.

DSP's practices are misaligned with notified countermeasures AND are ineffective

If practices are found to be misaligned and ineffective, the DSP would be exposed to regulatory investigation and enforcement action.

space for DSPs subject to review to demonstrate, through credible material, that their services meet the objectives of the notified obligation.

It also encourages timely remedial action during the review process. Where a DSP takes appropriate corrective steps to address identified gaps, this may reduce the need for further enforcement measures. In this way, the framework promotes compliance through course-correction rather than escalation.

The enforcement framework also **recognizes clear defences** where a DSP can demonstrate that:

- It **does not materially contribute** to the identified systemic risk and should therefore not be subject to further action; or
- The safeguards it has adopted, though different in form from the notified advanced obligations, are **at least equally effective** in mitigating the systemic risk. DSPs may substantiate these positions through structured transparency reporting and risk assessment disclosures - as seen in online safety assessment reports in Singapore²⁵⁵ and the UK²⁵⁶.

We believe these **'off-ramps' are important to build** in any online safety enforcement framework since it is practically impossible to design legal mandates that effectively address a given risk across all services without imposing unnecessary obligations on services that are already safe. As highlighted earlier, an example that squarely covers this scenario is the *Wikimedia v. Secretary of State litigation*.²⁵⁷ Wikipedia primarily contested that though it was already "safe" with regards to disseminating illegal content, its categorisation as a Category 1 service in the UK was prejudicial in as much as it would be forced to comply with drastic measures such as user verification that was not necessary in its case.

In fact, in our model, such DSPs would not need to contest their categorisation as a Significant Public UGC. Even if a DSP in Wikipedia's proverbial shoes were issued a show cause notice for non-compliance with, say, advanced obligations prescribed to counter the systemic risk of disinformation, the DSP would have ample opportunity to argue that:

- Its service design meets the objectives of the online safety framework, and through the adoption of different safety-by-design practices rather than ones prescribed by the framework, it *effectively* discharges its duty of care.

²⁵⁵ Infocomm Media Development Authority, "Online Safety Assessment Report 2024," February 17, 2025, <https://www.imda.gov.sg/-/media/imda/files/regulations-and-licensing/regulations/online-safety/online-safety-assessment-report-2024-designated-social-media-services.pdf>.

²⁵⁶ Ofcom, "Year 1 Online Safety Risk Assessments," December 4, 2025, <https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/research-statistics-and-data/os-standards/online-safety-risk-assessments-report-year-one.pdf?v=408834>.

²⁵⁷ *Wikimedia Foundation and another v. Secretary of State for Science, Innovation and Technology*, [2025] (High Court of Justice, UK), <https://www.judiciary.uk/wp-content/uploads/2025/08/Wikimedia-Foundation-and-another-v-Secretary-of-State-for-Science-Innovation-and-Technology.pdf>.

- Further, that despite the alleged non-compliance, it has not materially contributed to the disinformation systemic risk, therefore not meriting further investigation (though, admittedly, this may be hard to argue absent objective research findings proving this).

In sum, we believe that the enforcement philosophy **should build room for DSPs to innovate in the way they apply safety standards**. Therefore, DSPs are free to adopt measures different from what the regulator may prescribe, provided they are able to demonstrate that these are at least as effective as the prescribed measures (if called upon to do so).

This helps maintain fairness, encourages voluntary compliance, and ensures that regulatory interventions are effective. **Importantly, it signals that the enforcement regime is not solely about punishment**. It also incentivizes proactive risk management and remediation, while encouraging a culture of continuous improvement and collaboration between the industry and regulators.

07

Open Questions

Through the course of the report, and in particular with the preceding two chapters, our effort has been to comprehensively outline the structure and working of an online safety framework for India. Nonetheless, since we have deliberately taken a bird's-eye view to outline the elements that should exist and work together, we have not gone into detail on what each element should look like. This allows us to keep our recommendations concept-first.

Below, we outline the details we have deliberately left out. While there is broad consensus on the *concept*, working out the finer details requires separate research and deliberation. A concept-focused study like this cannot do justice to that task and such questions may merit follow-up studies of their own. Therefore, as an exercise in completeness, we set out the following open questions:

Open Questions / Issues	Reference
Precise nomenclature of different DSP classes	Chapter 6 (6.2)
Precise metrics of determining the 'Significant' classes of DSPs*	Chapter 5 (5.2) Chapter 6 (6.2)
The duration of "reasonable" time periods to resolve user grievances under different Tiers of obligations.	Chapter 6 (6.3)
Guidance on what constitutes "reasonable" steps to prevent harm. ²⁵⁹	Chapter 6 (6.3)
Contents and format of transparency reporting	Chapter 6 (6.3)
Critical review and reform of content take-down/blocking practices under Indian law	Chapter 6 (6.3)
Structure and form of the nodal authority ("Regulator") tasked with implementing this online safety framework *	Chapter 6 (6.4)
Contents and format of risk assessments and mitigation plans	Chapter 6 (6.4)
Standards or guidelines on what "simple, effective and accessible user reporting and grievance redressal mechanisms" entail	Chapter 6 (6.4)
Duration of Regulator's review of systemic risks and countermeasures	Chapter 6 (6.4)
Procedural features and details of the enforcement process *	Chapter 6 (6.4)

²⁵⁸ *Also addressed in the text of the report, please refer to the relevant sections for further detail and reasoning

²⁵⁹ While giving regulatory guidance on what constitutes "reasonable" steps to prevent harm or "accessible" user reporting mechanisms, we recommend avoiding overly prescriptive requirements that fail to acknowledge differences in service functionalities.

Appendices



Appendix I

This appendix compares our proposed framework with the EU Digital Services Act (DSA), highlighting areas of convergence and divergence in classification approaches. It explains the rationale behind key design choices in our framework, including the greater granularity of classification, the treatment of different service functionalities, and the approach to regulatory obligations and exemptions.

Explainer: Proposed Framework vis-a-vis EU-DSA

Convergence with EU-DSA

Our proposed framework relies on some principles similar to the EU-DSA :

- Like the EU-DSA, we too feel the need for a functionality-based classification that **avoids a 'one-size-fits-all' approach**.
- Similar to the VLOPs and VLOSEs classes, our framework too accords more oversight and focus on **large scale DSPs** as a key focus of safety regulation by creating a class of "Significant" Public and Private UGC services at Tier III.
- At the first tier of classification, our framework achieves the objective of exempting some services from more demanding mandates.

Divergence from EU-DSA

We deviate from the EU-DSA's classification in the following key ways:

- We find that the prime focus category under the EU-DSA - the VLOPs (and VLOSEs) - encompasses an extremely broad set of digital services. The resultant set of services covered by the VLOP designation contains a set of services so disparate that it is infeasible to have any meaningful common measures that work well for the incoherent group. This feature limits the framework's ability to target *specific* safety mandates to the covered DSPs beyond generic process based mandates such as risk assessments. This model may be workable in the context of the EU-DSA system, which distributes regulatory responsibility for online safety between the European Commission and national authorities. However, the effectiveness of the EU-DSA in addressing online safety risks has itself been subject to scrutiny and debate.²⁶⁰

²⁶⁰ Johann Laux, Sandra Wachter, and Brent Mittelstadt, "Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA," *Computer Law & Security Review* 43, 105613, (October 28, 2021), <https://doi.org/10.1016/j.clsr.2021.105613>.

However, we feel that an omnibus and standalone framework - like the one being proposed in India - needs greater ability to set more detailed and differentiated guidance for different functionalities. **As a result, rather than a size-heavy classification, our focus classes (Significant Public and Private UGC) build in more granularity in classification.**

- At Tier I, the EU DSA retains the distinction between conduit, caching and hosting services established under the earlier E-Commerce Directive. However, the hosting category was conceived when the range of services performing this function was considerably narrower than it is today. Recognizing this limitation, our framework takes a different approach, grouping services according to their underlying functionality and the nature of user interaction they enable, rather than relying on these legacy categories. Caching and Conduit services are however treated similarly under our model, and hence collapsed into a single Class 1 category. **This approach provides a clearer basis for classifying diverse digital services and determining the obligations appropriate to each class.**

Some other deviations:

- We have **chosen not to create a blanket exemption for all small business DSPs**, given the potential for such services to be the site of high-severity harms in Indian contexts. However, compliance is made fair for these DSPs because their baseline obligations remain light while still focusing on safety.
- In our stakeholder interviews, including with EU-DSA experts, we found the distinction between search engines and other online platforms to be superfluous, and therefore, we have avoided the same design choice.

Appendix II

The table below applies the three-part test to a set of illustrative online harms to demonstrate how the proposed “Systemic Risk” framework may operate in practice. The examples are indicative rather than exhaustive and illustrate the types of arguments that may support the inclusion - or exclusion - of a risk within an online safety framework. Even if one does not agree with the specific reasoning set out in the table, the illustration demonstrates how such a framework can operate in practice.

Legend: Red: Fails the test; Green: Passes the test; Yellow: Unclear (partially fails or passes)

Assessing Different Risks Under the Three-Part Test			
	Test 1 <i>Widespread or systematic in its manifestation</i>	Test 2 <i>Caused or primarily exacerbated by digital services</i>	Test 3 <i>Vertical regulation is not adequate or effective</i>
CSAM (Child Sexual Abuse Material)	CSAM is a high-severity risk that is: • Widespread - with consistently high volumes reported globally and in India;²⁶¹ and, • Systematic, with actively involved organised networks and recurrent patterns in its production and circulation.²⁶²	CSAM circulation is significantly exacerbated by digital services that make its creation and dissemination easier, particularly through private messaging and online storage services. Features such as anonymity, end-to-end encryption, and the absence of proactive detection tools can further facilitate its spread. ²⁶³ This can enable persistent, large-scale circulation that would be far more difficult to sustain offline.	Traditional criminal laws²⁶⁴ are not designed to oversee DSP-level measures that could help prevent the spread of CSAM within digital platforms. This gap makes online safety regulation critical for effectively addressing the spread of CSAM.
Disinformation	Disinformation in India manifests as a persistent and	Service features can be exploited to amplify and	Disinformation often intersects with sectoral

²⁶¹ Yadav, Nandini. 2025. “2024 saw record high in web pages hosting child sexual abuse content, nearly 3 lakh CSAM reports filed.” *India Today*, April 25, 2025. <https://www.indiatoday.in/technology/news/story/2024-saw-record-high-in-web-pages-hosting-child-sexual-abuse-content-nearly-3-lakh-csam-reports-filed-2713888-2025-04-24>.

²⁶² Dixon, Sophie. 2023. “Child sexual exploitation (CSE) networks: reassembling structure and activity.” *Journal of Sexual Aggression* Volume 31, no. Issue 3 (May): 396-412. <https://www.tandfonline.com/doi/full/10.1080/13552600.2024.2374076#abstract>.

²⁶³ Quayle, Ethel. 2021. “Affordances, Social Media and the Criminogenic Nature of the Internet: Technology-Mediated Child Sexual Abuse.” In *Online Child Sexual Exploitation*, pp 33–48. N.p.: Springer. https://link.springer.com/chapter/10.1007/978-3-030-66654-5_4

²⁶⁴ Sections 75, 77, 294 and 295, Bhartiya Nyaya Sanhita 2023; Sections 13 and 15, Protection of Children from Sexual Offences (POCSO) Act 2012.

Assessing Different Risks Under the Three-Part Test

	Test 1 <i>Widespread or systematic in its manifestation</i>	Test 2 <i>Caused or primarily exacerbated by digital services</i>	Test 3 <i>Vertical regulation is not adequate or effective</i>
	recurring phenomenon across digital platforms , with documented links to real-world harm - mob violence, electoral manipulation, ²⁶⁵ and panic during health crises. In many cases, the narratives pushed by disinformation actors are linked to real-world objectives in shaping public discourse, creating observable patterns of narrative manipulation. This demonstrates both its widespread reach and systematic recurrence. ²⁶⁶	accelerate disinformation, through engagement-based algorithmic recommendations, frictionless sharing, absence of fact-checking mechanisms, etc. , ²⁶⁷ making the speed, scale and reach of risk significantly higher than in offline settings.	domains such as elections, public health, and public order. However, existing sectoral laws primarily regulate actors and conduct within those domains and are not designed to address service-level amplification dynamics . Because the scale and velocity of disinformation are closely tied to platform design and distribution mechanisms, ²⁶⁸ service-level governance measures become necessary to effectively mitigate the risk.
Defamation	Online defamation can cause significant reputational harm and may occur at scale across digital platforms. In some contexts, particularly political discourse, defamatory narratives may be amplified through coordinated campaigns or organised smear efforts. Journalists and media organisations are sometimes also targeted by smear campaigns which seek	Digital platforms can significantly increase the reach and persistence of defamatory statements through engagement-driven recommendation systems, frictionless sharing, and the networked nature of online communication. These features may allow reputational attacks to spread rapidly and reach large audiences before correction or	Defamation is fundamentally an adjudicatory harm that requires legal assessment of evidence, context and intent. Historically, the Indian judiciary has defined what constitutes defamation on a case-by-case basis, assessing subjective elements such as falsity, reputational harm, and

²⁶⁵ NDTV. 2024. "Nearly Half Of The Fake News Stories In India Are Political: Study." NDTV, December 20, 2024. <https://www.ndtv.com/india-news/nearly-half-of-the-fake-news-stories-in-india-are-political-study-7291481>.

²⁶⁶ World Economic Forum. 2024. *Global Risks Report 2024*. <https://www.weforum.org/publications/global-risks-report-2024/>.

²⁶⁷ Akers, John, and Gagan Bansal. 2018. "Technology-Enabled Disinformation: Summary, Lessons, and Recommendations." Cornell University. <https://arxiv.org/abs/1812.09383>; Ceylan, Gizem, Ian A. Anderson, and Wendy Wood. 2022. "Sharing of misinformation is habitual, not just lazy or biased." *Proceedings of the National Academy of Sciences*, (December). <https://doi.org/10.1073/pnas.221661412>.

²⁶⁸ Starbird, Kate, and Tom Wilson. 2020. "Cross-platform disinformation campaigns: Lessons learned and next steps." Harvard Kennedy School Misinformation Review. <https://misinforeview.hks.harvard.edu/article/cross-platform-disinformation-campaigns/>.

Assessing Different Risks Under the Three-Part Test

	Test 1 <i>Widespread or systematic in its manifestation</i>	Test 2 <i>Caused or primarily exacerbated by digital services</i>	Test 3 <i>Vertical regulation is not adequate or effective</i>
	to undermine the credibility of news reporting and erode trust in the media. ²⁶⁹ However, the extent to which defamation manifests as a systematic or ecosystem-wide phenomenon across digital platforms remains debated.	adjudication. In addition, the persistence and searchability of online content can prolong reputational harm, allowing defamatory allegations to remain accessible and repeatedly resurface over time. However, the underlying harm still originates in specific allegations about identifiable individuals rather than platform dynamics, making the relationship between platform architecture and the harm more indirect than in cases such as disinformation.	applicable defences. ²⁷⁰ Digital service providers may therefore find it difficult to address defamation proactively through service-level safeguards without risking the over-removal of lawful speech. As a result, requiring platforms to proactively manage defamation within an online safety framework is likely to offer marginal benefits , while introducing a significant risk of over-censorship and chilling legitimate expression.
Unfair trade practices on social commerce	India's social commerce market is rapidly growing and estimates suggest that it will grow ten-fold to USD 55 billion by 2030. ²⁷¹ As this ecosystem has grown, concerns have emerged regarding misleading influencer endorsements, undisclosed commercial relationships, and inadequate verification of sellers. While	Social commerce relies on digital platforms that combine social interaction with commercial promotion, including influencer marketing and algorithmically surfaced content. These features may amplify the visibility and reach of misleading endorsements or deceptive commercial claims, particularly where advertising disclosures are unclear or	India already has a comprehensive consumer rights framework that defines "endorsement" and "unfair trade practices" and introduces penalties for violations. The scope of this framework can easily be adapted to address harms associated with social commerce. Platforms themselves have limited

²⁶⁹ "Digital Safety: Protecting against online smear campaigns." 2022. Committee to Protect Journalists. <https://cpj.org/2022/09/digital-safety-online-smear-campaigns/>.

²⁷⁰ Law Commission of India. 2024. "Report No. 285." The Law on Criminal Defamation. <https://cdnbbsr.s3waas.gov.in/s3ca0daec69b5adc880fb464895726dbdf/uploads/2024/02/202402021139959703.pdf>; Internet Democracy Project. 2018. "Unshackling expression: A study on laws criminalising expression online in Asia." Internet Democracy Project. <https://internetdemocracy.in/reports/unshackling-expression-a-study-on-laws-criminalising-expression-online-in-asia/#defamation>.

²⁷¹ Press Trust of India. 2024. "Social commerce set for growth, concern of authenticity looms large." April 7, 2024. https://www.business-standard.com/industry/news/social-commerce-set-for-growth-concern-of-authenticity-looms-large-124040700339_1.html.

Assessing Different Risks Under the Three-Part Test			
	Test 1 <i>Widespread or systematic in its manifestation</i>	Test 2 <i>Caused or primarily exacerbated by digital services</i>	Test 3 <i>Vertical regulation is not adequate or effective</i>
	such practices have caused consumer harm and could increase in scale as the market matures, it is less clear whether they currently constitute a systematic risk across digital platforms.	verification mechanisms are weak. However, the underlying harm primarily arises from traditional unfair trade practices rather than from risks generated by platform architecture itself.	ability to address the core regulatory aspects of such harms, which typically involve seller verification, commercial disclosures, and consumer redress mechanisms that fall within the mandate of specialised regulators. <i>For example, India's securities regulator SEBI has extended its oversight to investment advice disseminated through social media, requiring individuals providing such advice to register with the regulator, and prominently display their name and SEBI registration number to signal credibility.²⁷² Digital service providers are required to help enforce the regulation by providing reporting channels and conducting advertiser verification.²⁷³ This approach allows specialised regulation of investment-related risks by the securities regulator while avoiding unnecessary expansion of an online safety framework.</i>

²⁷² MoneyControl. 2026. "SEBI issues social media platform usage norms for regulated entities." *MoneyControl*, February 26, 2026. <https://www.moneycontrol.com/news/business/markets/sebi-issues-social-media-platform-usage-norms-for-regulated-entities-13844702.html>.

²⁷³ Gupta, Ishika. 2025. "SEBI Tightens Social Media Ad Rules for Registered Intermediaries." *Medianama*, March 24, 2025. <https://www.medianama.com/2025/03/223-sebi-tightens-social-media-ad-rules-for-registered-intermediaries/>; Tiwari, Khushboo. 2025. "Sebi to deploy AI tool to track misleading ads, finfluencer content." *Business Standard*, October 7, 2025. https://www.business-standard.com/markets/news/sebi-to-deploy-ai-tool-to-track-misleading-ads-finfluencer-content-125100701245_1.html.

Assessing Different Risks Under the Three-Part Test

	Test 1 <i>Widespread or systematic in its manifestation</i>	Test 2 <i>Caused or primarily exacerbated by digital services</i>	Test 3 <i>Vertical regulation is not adequate or effective</i>
Obscene content	The concept of “obscenity” is inherently subjective and context-dependent, shaped by evolving legal standards and social norms. In India, concerns around obscene or sexually suggestive and explicit material have periodically gained prominence in public and policy debates, including judicial scrutiny of such content on online platforms.²⁷⁴ However, obscenity as a category of speech does not necessarily manifest as a widespread or systematic pattern of harm across digital services in the same way as risks such as disinformation or child sexual abuse material (CSAM). Rather, it typically concerns disputes over the permissibility of specific	Obscene or sexually explicit content is not inherently produced by digital platforms. In the case of OTT services, such material typically reflects editorial decisions made by publishers, and risks are addressed typically through classification systems and, in the case of children, through parental controls. For platforms that host user-generated content, digital architecture can significantly expand the accessibility and circulation of sexually explicit or suggestive material through features such as frictionless sharing, searchability, and large-scale distribution networks. Engagement-driven recommendation systems and monetisation features may also indirectly incentivise creators²⁷⁵ to produce	India’s criminal law framework, including provisions under the Bharatiya Nyaya Sanhita and related statutes, already addresses obscenity as an offence.²⁷⁷ Jurisprudence has evolved to delineate its contours, including in digital contexts.²⁷⁸ The IT Act, 2000 also provides mechanisms for government-directed takedowns. These regimes are primarily content-focused and punitive in nature. They do not directly regulate platform architecture, recommender systems, or age-verification design. However, because obscenity is a speech-sensitive and normatively contested category, expanding online safety regulation into this domain raises significant

²⁷⁴ Express News Service. 2025. “‘Do something... something legislative’: Supreme Court says obscenity on OTT platforms and social media serious issue.” *Indian Express*, April 29, 2025. <https://indianexpress.com/article/india/supreme-court-regulation-obscene-ott-social-media-9970238/>.

²⁷⁵ ACHARYA, DISHA, and AJAY RAG. 2026. “ETtech in-depth: Monetisation wrestles with moderation in livestream grey zones.” *The Economic Times*, February 23, 2026. <https://economictimes.indiatimes.com/tech/technology/monetisation-wrestles-with-moderation-in-livestream-grey-zones/articleshow/128685231.cms>.

²⁷⁷ Nizar, Senu, and Manan Katyal. 2025. “No New ‘Obscenity Law’ Can Solve the Problem of Poor Taste,” TQH: Quick Policy Review. The Quantum Hub. <https://thequantumhub.com/tqh-quick-policy-review-laws-on-obscenity/>.

²⁷⁸ In *Apoorva Arora v. State (Govt. of NCT of Delhi)*, which examined the web series *College Romance* for its alleged use of profanities and sexually suggestive language, the Supreme Court clearly articulated how the standard of obscenity applies to digital media. The Court held that although the series used profanities and sexual language, it did not provoke sexual or lustful thoughts in the mind of the average viewer and therefore did not qualify as obscene under the law. This importantly underscores that **determinations of obscenity require legal adjudication**. This also applies to other digital services such as social media platforms. Providers of such services, acting on their own, cannot proactively determine what constitutes obscenity and therefore should not take action beyond complying with government-directed takedown obligations. This argument, however, does not extend to nudity-related content, where issues of consent and unrestricted access to minors raise distinct concerns and justify proactive platform intervention.

Assessing Different Risks Under the Three-Part Test

	Test 1 <i>Widespread or systematic in its manifestation</i>	Test 2 <i>Caused or primarily exacerbated by digital services</i>	Test 3 <i>Vertical regulation is not adequate or effective</i>
	forms of expression. A more systemic concern may arise in relation to the repeated or large-scale exposure of minors to obscene material where age-gating, classification, or parental control mechanisms are inadequate. In such cases, the risk may be framed less as obscenity and more as a failure of child-protection safeguards.	increasingly sensational or sexually suggestive content in order to attract and retain audiences, even where obscenity is not an explicit design objective. Studies of youth media use have also found that algorithmic feeds and peer sharing often expose minors to sexual imagery even where users do not actively seek it out.²⁷⁶ At the same time, these dynamics arise primarily from engagement optimization and user behaviour. The systemic risk therefore lies in the amplification and incidental exposure created by platform architecture.	free expression concerns and risks regulatory overreach. Accordingly, while vertical regulation may not address every architectural dimension of dissemination, it substantially governs the core legal question of what constitutes unlawful obscenity.
Gaslighting ²⁷⁹	Gaslighting has been recognised as an emerging harm in public discourse, particularly among young users. In some contexts, it has also been linked to financial fraud and coercive online relationships.²⁸⁰ An initial draft of the proposed Digital India Act, released in 2023,	Digital service functionalities such as private messaging, anonymity, and persistent communication channels may facilitate gaslighting in particular contexts. While these functionalities can be exploited to harass, abuse or defraud others online, their design does not incentivise or	Indian legal frameworks, including the Protection of Women from Domestic Violence Act, the Bharatiya Nyaya Sanhita, POSH, and matrimonial law, recognise and penalise forms of mental cruelty, harassment, coercion, and fraud that may encompass gaslighting

²⁷⁶ Livingstone, Sonia, and Leslie Haddon. 2009. "EU Kids Online: Final Report." EU Kids Online. https://eucpn.org/sites/default/files/document/files/5_eu_kids_online_-_final_report.pdf.

²⁷⁹ Gaslighting is a form of emotional/psychological abuse where the victim is manipulated to question their own feelings, instincts, and sanity. It can involve denial, twisting facts, or making you feel like you're overreacting.

²⁸⁰ NT, Sarasvati. 2023. "Expect 'gaslighting' and 'catfishing' to be offences under upcoming Digital India Bill." Medianama. <https://www.medianama.com/2023/01/223-expect-gaslighting-catfishing-offences-draft-digital-india-bill-2/>.

Assessing Different Risks Under the Three-Part Test

	Test 1 <i>Widespread or systematic in its manifestation</i>	Test 2 <i>Caused or primarily exacerbated by digital services</i>	Test 3 <i>Vertical regulation is not adequate or effective</i>
	included gaslighting as an emerging online harm.²⁸¹ However, gaslighting is fundamentally a relational and interpersonal harm. While digital platforms may serve as the medium through which such behaviour occurs, it does not typically manifest as a large-scale phenomenon in the same way as disinformation.	structurally promote such behaviour. Gaslighting remains a psychological harm rooted primarily in interpersonal dynamics rather than in platform architecture. While digital affordances may enable repeated connection or concealment, they do not uniquely create or amplify the behaviour in the way that engagement-driven systems may amplify other harms.	behaviour.^{282,283,284,285} These regimes are capable of addressing the underlying wrongful conduct without requiring a dedicated online safety intervention. Expanding platform-level regulation specifically for gaslighting risks duplicating existing legal protections and creating definitional uncertainty around a concept that is psychologically nuanced and legally indeterminate.
Copyright infringement	Copyright infringement can be both widespread and systematic. In India, pirated content has moved beyond torrent and fringe websites, and is prevalent across mainstream social media and even encrypted services.²⁸⁶	Copyright infringement can be significantly exacerbated by the technological affordances of digital services. Features such as user-upload systems, embedding tools, remixing functionalities, and frictionless sharing mechanisms enable users to copy, reproduce, integrate, and redistribute copyrighted works at scale. The searchability and	India has extensive jurisprudence interpreting how the Copyright Act, 1957 applies to digital services and delineates their liability and responsibility in cases of infringement. It clarifies when DSPs may be held liable for secondary infringement and when they may avail safe harbour protections under both the

²⁸¹ Ministry of Electronics & Information Technology, Government of India. 2023. "Proposed Digital India Act, 2023." Digital India. <https://www.tcs.com/content/dam/tcs/pdf/discover-tcs/investor-relations/faq/proposed-digital-india-act.pdf>.

²⁸² Kumar, Sujeet. 2025. "Divorce on Ground of Cruelty in Hindu Marriage Act," A Presentation by Sujeet Kumar, Judge, Family Court, Kotdwar. NA. <https://cdnbbsr.s3waas.gov.in/s37a68443f5c80d181c42967cd71612af1/uploads/2025/03/20250321682441929.pdf>.

²⁸³ India. 2005. The Protection of Women from Domestic Violence Act, 2005 (Act No. 43 of 2005). [https://www.indiacode.nic.in/bitstream/123456789/15436/1/protection of women from domestic violence act%2C 2005.pdf](https://www.indiacode.nic.in/bitstream/123456789/15436/1/protection%20of%20women%20from%20domestic%20violence%20act%202005.pdf).

²⁸⁴ India. 2013. The Sexual Harassment of Women at Workplace (Prevention, Prohibition and Redressal) Act, 2013 (Act No. 14 of 2013). https://doe.gov.in/files/inline-documents/DoE_Prevention_sexual_harassment.pdf.

²⁸⁵ India. 2023. The Bharatiya Nyaya Sanhita, 2023 (Act No. 45 of 2023). https://www.mha.gov.in/sites/default/files/250883_english_01042024.pdf

²⁸⁶ Maheshwari & Co. 2025. "Fighting Digital Piracy and Copyright Infringement in India." Lexology. https://www.lexology.com/library/detail.aspx?g=610f3619-d8cb-4995-bde7-e168aa75072c&utm_

Assessing Different Risks Under the Three-Part Test			
	Test 1 <i>Widespread or systematic in its manifestation</i>	Test 2 <i>Caused or primarily exacerbated by digital services</i>	Test 3 <i>Vertical regulation is not adequate or effective</i>
		persistent accessibility of digital content can further accelerate the spread of infringing material. Unlike harms such as defamation or obscenity, where digital services primarily amplify the visibility of user-generated speech, copyright infringement is closely linked to platform functionalities that facilitate the reproduction and reuse of protected works.	IT Act, 2000 and the Copyright Act.²⁸⁷ For instance, jurisprudence has established the “actual and specific” notice standard of knowledge for safe harbour under copyright law, requiring DSPs to take down content upon receiving private legal notice.²⁸⁸ The Copyright Act also specifies takedown process for infringing content, and the Copyright Rules, 2013 provide the SoP for DSPs once they receive a written notice from a copyright holder.²⁸⁹ Although enforcement challenges persist, copyright infringement is governed by a specialised and comprehensive regulatory regime designed to balance rights protection, innovation, and intermediary liability. Emerging questions around AI training and the use of copyrighted material may require further debate within copyright law itself. However, introducing

²⁸⁷ Devadasan, Vasudev. 2022. “Report on Intermediary Liability in India.” Centre for Communication Governance. <https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/reportonintermediaryliabilityinindia-web-180123-344.pdf>.

²⁸⁸ Devadasan, Vasudev. 2022. “Report on Intermediary Liability in India.” Centre for Communication Governance. <https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/reportonintermediaryliabilityinindia-web-180123-344.pdf>.

²⁸⁹ India. 2013. The Copyright Rules, 2013. https://upload.indiacode.nic.in/showfile?actid=AC_CEN_9_30_00006_195714_1517807321712&type=rule&filename=copyright%202013searchable.pdf

Assessing Different Risks Under the Three-Part Test

	Test 1 <i>Widespread or systematic in its manifestation</i>	Test 2 <i>Caused or primarily exacerbated by digital services</i>	Test 3 <i>Vertical regulation is not adequate or effective</i>
			additional online safety regulation in this domain risks duplication, doctrinal inconsistency, and disruption of carefully calibrated legal standards.

Appendix III

The table below provides an illustrative list of safeguards that providers of Significant Public UGC and Significant Private UGC services may adopt to address the systemic risk of **disinformation**, in compliance with the different families of advanced obligations – *safety-by-design, transparency and user empowerment*.

Systemic Risk: Disinformation		
Type of Advanced Obligation	Illustrative examples of measures that may be adopted to comply	
	Significant Public UGC	Significant Private UGC
Safety by Design	• Nudge users to fact-check and opt for verified information. • Apply prominent contextual labels. (e.g., community notes, AI use detected) • Algorithmic deprioritization of identified content. • If federated: Moderators or admins to block or restrict users, content, or servers.	• Introduce additional friction in content dissemination, such as limiting single-click forwards, group sizes. • Apply prominent contextual labels (e.g., forwarded texts, AI use detected)
Transparency	Publish transparency reports with a designated section on dis/misinformation, including: • Overview (mitigation measures adopted, action taken on user reports, law enforcement cooperation) • Trends (top subject areas (e.g., health, elections) and top instances of viral content) • Performance metrics (impact of mitigation measures assessed through evaluations and comparative estimates over time)	
User Empowerment	• Content filtering mechanisms that allow users to avoid topics/feeds. • Put in place simple and accessible user reporting mechanisms to flag mis/disinformation.	• Put in place simple and accessible user reporting mechanisms to flag mis/disinformation.

