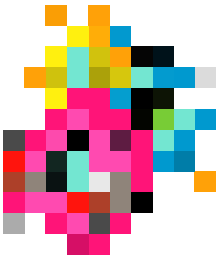
26 recommendations on content governance

a guide for lawmakers, regulators, and company policy makers

accessnow

Access Now **(https://www.accessnow.org)** defends and extends the digital rights of users at risk around the world. By combining direct technical support, comprehensive policy engagement, global advocacy, grassroots grantmaking, and convenings such as RightsCon, we fight for human rights in the digital age.

For more information about this report, please contact Javier Pallero (javier@accessnow.org)



This paper is a publication of Access Now and was written by Eliška Pírková and Javier Pallero with the collaboration of Access Now's policy team. The authors would like to especially thank Ben Wagner, Daphne Keller, Mathias Vermeulen, Kate Klonick, Aleksandra Kuczerawy, Brett Solomon, Fanny Hidvégi, Guillermo Beltrà, Juliana Castro, Donna Wentworth, and the Access Now policy team for their contributions.

Table of contents

I. Executive summary

The internet has given us an essential tool to exercise human rights, including access to information and freedom of opinion and expression, among others. Services that act as intermediaries for the flow of information, especially platforms such as social media services and search engines, play an important role in this. Digital platforms theoretically give everyone the opportunity to connect, but repressive governments can interfere with that capacity through legislation that endangers people's rights.

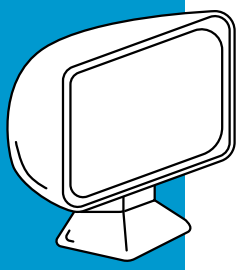
At the same time, the rules that platforms use to govern content and user activity – typically developed unilaterally – are often designed and applied in ways that are at odds with freedom of expression, privacy, and other fundamental rights. This, in turn, can enable new forms of exploitation, both by private and public actors.

The actions that platforms and governments take in this area, and those they fail to carry out, can harm societies and vulnerable populations in particular. Societal phenomena that are increasingly common, such as disinformation campaigns or illegal content that can incite hatred or violence, can manifest in a variety of forms with diverse characteristics that present a wide range of challenges. Their impact and the implications for human rights differ significantly according to the specific geographical, cultural, and political context. In some parts of the world, the negative consequences and collective harm to targeted groups can represent or lead to atrocities.

Therefore, this paper does not seek to establish a universal set of specific solutions for the complex and thorny issues that content governance raises. Instead, building on our experience in policy development across the globe, we offer basic human rights-centered guidelines that can serve as the minimum basis for governance policies that are fit-for-purpose, given that stakeholders must consider the specific actors and technologies in play in their region. We plan to elaborate further on regional and issue-based implementation of these recommendations in future publications.

This paper starts by defining the three main types of governance structures that are being used today: regulatory, self-regulatory, and co-regulatory. Then we explore the human rights risks associated with each type and offer recommendations to address those risks. Where it is possible, we recommend concrete baseline policies to address key issues such as intermediary liability, automated measures, and self-regulation decisions, among others.

We aim to reach decision-makers across the globe with the goal of putting human rights at the forefront of every debate about content governance. Doing so is the only pathway for creating a digital future that reinforces shared ideals of freedom, openness, and democratic values, with the potential for returning power to the users.



II. Introduction: how internet regulation has evolved

THE EARLY DAYS AND THE EMERGENCE OF REGULATION

When the first global social media services debuted in the late '90s and early 2000s, popular opinion – held by technology experts, analysts, and ordinary internet users alike – was that the internet would enable more people to access and share more information than ever before. For many, the potential for unfettered access and exchange of information worldwide suggested that the internet would empower users and serve as a force for unleashing collective action and democratization.[1] Indeed, many individuals and organizations globally have been able to harness the internet as a tool for achieving real impact, using online activism to defend human rights and strengthen democracy. Access Now and other digital rights organizations have worked to protect and preserve this capacity. However, as the commercial internet developed, governments around the world began to recognize and respond to the problems that arise when content, including content deemed illegal, is shared online.

Many of the first regulatory efforts focused on the internet involved copyright violations and pornographic material.[2] Less prominent were efforts to address the societal phenomena that dominate policy discussion about the internet today, such as hate speech, defamation, or terrorist content. The initial focus on copyright and pornography would lead to the voluntary adoption by online platforms of the first generation of upload filters for detecting illegal content.

While the first footprints for internet regulation were shaped by government and industry-led concern over copyright and pornographic materials, so-called intermediaries[3] developed self-regulatory mechanisms to deal with the issues arising from other forms of online expression, such as hate speech, harassment, bullying, nonconsensual pornography, and more. Today, we are still dealing with the early influence of governance undertaken with an intellectual property-driven mindset, versus a more holistic approach that considers the full spectrum of issues raised by online speech. The bifurcation of approaches has taken place at the decision-making level, in both the public and private sectors. The increased use of automated tools to regulate online speech, developed for copyright enforcement but threatening free expression, has served to highlight this splintering of approaches.

CHALLENGES TO USER RIGHTS ONLINE: SCALING UP THE PROBLEM

How users access and exchange information has changed significantly from the early days of the internet. Not only has access to the internet expanded, the emergence of social media platforms has facilitated and increased the number of interactions. This implies more access to information and an expansion of enjoyment of freedom of expression and other fundamental rights.[4]

It has also brought new difficulties, like the creation and dissemination of content governments have deemed illegal, at an unprecedented scale. There has been a

similar scaling up of content that, while it may not be illegal in a given country, is considered "harmful" or undesirable, whether by online platforms, platform users, or regulators in nation-states. Material used for cyberbullying, spreading disinformation, or violent content fall under this category.[5] It is in this context that online platforms have engaged in self-regulation, creating and enforcing rules about acceptable expression on their services.

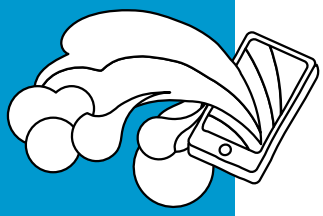
These internal rules, usually called "terms of service" or "community guidelines," among other names, typically include a list of prohibited behaviors on the platform. They define what is or is not acceptable content and behavior. They may also explain and justify platforms' principles and values, as embodied in the guidelines.

Due to the increasing pressure from platform users and regulators to do more to stop the spread of illegal and harmful user-generated content, online platforms have progressively tightened their rules with regard to all types of content, including hate speech and material implicated in incitement to violence.[6] Critically, however, the process platforms use to implement their self-imposed rules has remained for the most part opaque. In fact, of the actors involved in internet governance, private actors disclose the least amount of information about how their regulatory mechanisms are formulated or enforced. [7]

Nation states' concerns about the spread of illegal content across online platforms are largely justified. It is urgent to find adequate responses to complex societal phenomena such as hate speech or online radicalization. But this cannot be an excuse for overbroad censorship of users' speech. The approaches and tools that governments and platforms use must respect international human rights standards and the rule of law, placing the fundamental rights and freedoms of users at their core.

If the governance solutions developed by governments and platforms (through regulation and self-regulation respectively) are ill-advised, rushed, or do not incorporate international human rights principles and safeguards, they can increase, rather than decrease, the risks for users.

While governments and platforms struggle with pressure to respond to public relations crises, market influences, or complaints by constituents, the most vulnerable people in this dynamic are the users. At Access Now, our mission is to defend and extend the digital rights of users at risk across the globe, and this includes working to ensure that at-risk individuals and groups do not become victims of censorship or abuse online, whether through government policies or corporate practices. Our team works internationally, with offices and staff in regions around the world, and we wish to underscore the need for a heightened understanding of the social, cultural, and legal nuances of this debate. Through our Digital Security Helpline[8] and in our coalition work, we are in direct contact with activists, journalists, human rights defenders, and others negatively impacted by illegal or undesirable content. These same users have been placed in the cross hairs of government regulations and company self-regulatory practices that, however well-meaning, have hurt their capacity for free expression and empowerment. Our hope is that this guide can give decision-makers the building blocks for vindicating their rights and preserving their freedoms.



III. Content governance: how the rules are made and who enforces them

Actors in our increasingly complex online communications ecosystem have the duty to consider human rights. Governments are obligated to protect these rights, while companies are responsible for respecting them. It is in the context of these duties and responsibilities that we examine how content governance rules are made and who enforces them.

In this paper, "governance" refers to the complex processes and interactions that public and private stakeholders create and engage in to enforce rules for governing content online.[9] We divide content governance into three main categories: **state regulation**, enforced by governments; **self-regulation**, exercised by platforms; and **co–regulation**, undertaken by governments and platforms together through mandatory or voluntary agreements.

Each type presents unique challenges for the protection of human rights online, as we outline below.

STATE REGULATION

By state regulation, we mean any binding legal or regulatory instrument that local, national, or regional public institutions enact through their legislative or regulatory processes. States typically define the online content that is to be considered illegal through their criminal codes. In some jurisdictions, states also rely on issue-specific legislation to regulate content, as in the case of intellectual property rights.

SELF-REGULATION

Online platforms define what kind of content is acceptable using their services, often by creating their own terms of service.[10] These terms can contain definitions that encompass illegal content and content that, despite being legal, is considered undesirable.

This is an exercise of self-regulation that can be done unilaterally or after consulting users or other stakeholders. Platforms carry out self regulation primarily in two ways: through moderation or curation of content.

CONTENT MODERATION

Content moderation is the practice through which a company that owns and runs an online platform that hosts or provides access to user-generated content[11]– such as a search engine or a social media service – makes decisions about whether to host or continue hosting a specific piece of content under its terms of service.

A decision about whether to host content could entail taking the content down permanently or temporarily, either on the platform as a whole or in relation to certain groups of users in a specific geographical area.

CONTENT CURATION

Decisions regarding the reach, prominence, or amplification of certain content (often called "content curation"[12]) determine how many and which groups of users are exposed to the content and the way in which it is presented. These decisions are informed by different criteria and methods.[13] Online platforms often recommend content to their users via news feeds or personalized suggestions. Content curation decisions could entail boosting the reach and exposure of some forms of speech, or demoting or limiting that reach.

CO-REGULATION

Co-regulation is an increasingly popular approach that adds to the complexity of content governance. Co-regulatory regimes can be understood as self-regulation that is actively encouraged, supported, and sometimes monitored by public authorities. Typical examples of co-regulatory mechanisms are voluntary codes of conduct resulting from dialogue between private actors and national or regional authorities.

Co-regulatory mechanisms can have different levels of formality. Co-regulatory initiatives, in a strict sense, often include a formal regulatory element — such as a law or administrative decision — that acts as a framework and governs the activities of the actors involved, including rules and consequences of different kinds.[14] But in other cases, such cooperation between public authorities and private actors is governed by informal voluntary agreements that also set rules and objectives. The Christchurch Call to Eliminate Terrorist and Violent Extremist Content[15] and the Global Internet Forum to Counter Terrorism,[16] among others, are examples.

Since these initiatives raise many of the same human rights concerns, in this paper we use "co-regulation" broadly to refer to any initiative where both governments and private actors play a role, whether or not they are linked to underlying legal or regulatory frameworks.

IV. Risks: how content governance decisions affect human rights

Content governance that is incompatible with basic human rights rules and principles imperils free expression, access to information, freedom of opinion, association, privacy, and other human rights, and it impacts different populations in diverse ways. Below we highlight the challenges for different types of users and expand on the most prominent risks associated with common content governance mechanisms.

GOVERNANCE DECISIONS AFFECT DIFFERENT CATEGORIES OF USERS DIFFERENTLY

Platforms theoretically give a voice to everyone, but harmful speech can also be accelerated in a networked environment. Countless people have been harmed by the amplification, proliferation, and ubiquity of illegal or harmful speech. Those victimized do not comprise a homogeneous group, but have included marginalized populations such as women, religious or ethnic minority groups, people of color, and members of the LGBTQ community.[17] Somewhat ironically, these same individuals and groups are often decentralized and rely on online platforms to find people with common interests, get support, exchange information, express their opinions, participate in politics, organize, work, and access culture, among other exercises of fundamental rights. Losing access to content they have published, having their voices silenced, or being removed from spaces of public discourse and deliberation can have enormous negative impact on their lives and constitutes a terrible loss for public debate.

The same is true for content governance decisions that affect civil society organizations, activists, and journalists. When preparing regulatory responses for content governance, states should reconcile the protection of personal data with the right to freedom of expression and access to information, especially in connection to journalism, scientific research, or artistic or literary expression. To protect these activities, legal frameworks must provide for exemptions or derogations from general prohibitions regarding particular categories of online content.

Those making content governance decisions, whether governments or companies, must ensure extra care and attention to the design and implementation of human rights safeguards for vulnerable groups. Accordingly, any policies affecting those groups should guarantee their effective participation in the design and implementation stages.

Content moderation decisions are an increasing concern for users at risk

Access Now's Digital Security Helpline, which works with individuals and organizations to keep them safe online, has seen an increase in cases related to content moderation decisions that affect users most at risk.

Issues related to content moderation were part of approximately 20% of the Digital Security Helpline cases for 2019[18] (~311 cases). These cases can generally be classified into two categories: content recovery and content takedowns. Content recovery cases include cases where users considered their content was wrongfully taken down. This can be caused by errors in the application of the platforms' terms of service, often derived from issues with automated flagging systems but also from mistakes made by human moderators. Content takedown cases include requests from clients to assist with taking down content that often includes harassment, impersonation, doxxing, and calls to violence or death threats against human rights defenders, journalists, and other members of civil society in retaliation for their work. In many of these cases, users contact Access Now's Digital Security Helpline as a measure of last resort, after their requests through official social media platforms' channels have been either ignored or denied.

HUMAN RIGHTS RISKS LINKED TO STATE REGULATION

There are specific concerns associated with the exercise of state power. The risk of government abuse is always a possibility, and not only in countries governed by authoritarian regimes. Therefore human rights standards, the rule of law, functioning democratic institutions, evidence-based policies, transparency, and participation are essential for any regulation of expression. There is an absence of these safeguards in some regulatory proposals and enacted legislation, and this represents a concrete risk for the rights of users.

ISSUES WITH LEGALITY, NECESSITY, AND PROPORTIONALITY

State actors often have legitimate concerns related to the dissemination of illegal content online. Sadly, state regulators all too often do not abide by the fundamental legal principles to which they are bound. Laws that are incompatible with human rights often serve as a pretext for demanding that platforms swiftly remove content from their sites, which can suppress legitimate discourse and dissent.[19] Some states rely on disproportionately restrictive criminal laws that contain broad definitions of crimes such as extremism, defamation, blasphemy, and other speech-related acts. For example, in early 2019, Australia amended its criminal code in just 48 hours to add a prison sentence for business executives of online platforms that fail to take down "extremist content" within a tight time frame. [20]

In other cases, governments have proposed and adopted legislation that abuses self regulatory and voluntary measures deployed by private actors.[21] Such an approach is incompatible with basic human rights safeguards and ultimately leads to over-removal of content or outright censorship.

State regulatory models should focus specifically on expressly illegal content and avoid regulation regarding ever-evolving definitions of online societal phenomena, such as disinformation, hate speech, or terrorist content. Restrictions of the right to freedom of expression must be clearly prescribed by law, pursue a legitimate aim, be necessary in a democratic society, and be proportionate to the aim pursued.

The European Court of Human Rights determined in its jurisprudence[22] that any expression, including online content, has to be assessed in the context and circumstances under which it is disseminated, taking into account its possible impact, the means of its dissemination, the authors, and their public influence. Inherent in that prescription is the idea that overbroad legal definitions will ultimately lead to unnecessary and disproportionate interference with the right to freedom of expression.

DELEGATION OF JUDICIAL FUNCTIONS TO PLATFORMS

There is a trend in legislation focused on delegating some functions typically performed by states to platforms. The German Network Enforcement Act, also called NetzDG,[23] imposes an obligation on platforms to evaluate and remove content that is illegal under German criminal law, a function that is normally performed by an independent adjudicator.

Mechanisms of this kind are dangerous for human rights for a number of reasons. The objectives of a private company are very different from those of a public and democratic adjudicator. An independent judicial or administrative adjudicator should act within the confines of appropriate national legislation and human rights law. Such adjudicator should consider what is the best course of action to attain a legitimate aim[24] for a democratic state and consider necessity and proportionality principles.

Fundamental rights enshrined in political constitutions provide for the needed economic and political independence of adjudicating authorities. An expertise on the subject matter under their competence is also required. Finally, adjudication systems provide for appeal and redress.

These characteristics and safeguards are sometimes difficult to apply in practice. There is a need for swift responses and governments often struggle with resources, capacity building, and other practical challenges. Nevertheless, we need to strengthen and reformulate access to justice following legal and human rights-based guidelines. Most of the characteristics that ensure the public interest goes first are not present in the decision-making processes of private platforms.

LACK OF EVIDENCE AND INFORMATION WHEN REGULATING

Any state regulation addressing online societal phenomena such as disinformation, hate speech, or terrorist content must always be grounded in solid evidence. Numerous legislative proposals that have been recently proposed or adopted by national governments around the world seek to combat undesirable content without proper factual backing to show the impact the content may have on vulnerable groups or a link between offline and online

behavior. This is the case in legislation proposing the takedown of entire websites and services to fight disinformation and online hatred in Honduras[25] and Bangladesh.[26]

The risk of regulating in an attempt to reach a particular outcome without fully understanding the long-term principles that could be compromised is currently a problem at play in all governments. Without knowledge, cleverness in policy does little good, and can create even more harm. The lack of evidence-based policy making has led to the adoption of hasty regulatory responses that focus on swift content removal and on the quantity of removed content. Regulation that pushes for speed and quantity of content removal may in turn generate over-compliance by online platforms that results in illegitimate takedowns of user-generated content.[27]

LACK OF A TIME-SENSITIVE AND PARTICIPATORY APPROACH

Governments often react to societal phenomena by rushing to enact legislation, without enabling sufficient debate and without consulting relevant stakeholders. This opens the risk of regulatory responses that are not adequate to tackle the issues of concern and that may put innocent users at risk, including journalists and vulnerable communities.

HUMAN RIGHTS RISKS LINKED TO SELF-REGULATION

Decisions made by platforms about what types of content are allowed and how it is presented to users affect their capacity to express their ideas and access information online. Content moderation and curation have an impact at both the individual and collective levels, since individual decisions by platforms – especially those made by global dominant platforms – have a cumulative impact, shaping the space for discussion and potentially silencing the voices of entire communities. This can represent a significant risk for users in general, and for vulnerable and marginalized communities in particular. Following, we outline some of the most prominent risks stemming from content moderation and curation practices.

RISKS LINKED TO CONTENT MODERATION

LACK OF CLARITY AND TRANSPARENCY IN RULES

It is not always clear how online platforms make content moderation decisions. Recent research has shown that content takedowns are largely arbitrary and subject to the discretion of platforms.[28] This is particularly worrying in a context in which governments are exerting public pressure on platforms to increasingly moderate content faster and with more precision, vastly outstripping the real-world technical capacity to do so. Inconsistency is to be expected, since the human beings that make content moderation decisions work under tight deadlines with little training or support,[29] and traditionally have lacked geographic, linguistic, and cultural diversity. When moderation is machine-assisted, automated decisions can fail spectacularly to understand the contextual nuances of language.[30]

The lack of meaningful transparency jeopardizes the enjoyment of users' right to appeal and access to adequate remedy. Moreover, the lack of transparency and proper accountability mechanisms negatively impacts users' right to receive and impart information. This is a pressing issue, as the internet has become an essential tool for participation in activities and discussion concerning topics of public interest.[31] Thus, online platforms serving as "modern public squares"[32] are taking direct part in shaping public discourse.

LACK OF DEFENSE AND REMEDY IN DECISIONS

Users often have little or no opportunity to respond to content takedowns, assert the legitimacy of the content, or get remedy for improper removal. There are well-known cases in which content that has artistic or historical value has been taken down due to overly strict interpretation of a company's terms of service.[33] For these reasons, the U.N. Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has called for greater transparency and accountability in content moderation decisions,[34] as have a number of civil society organizations. Perhaps as a result, some platforms, including dominant players, have begun to share more information about their internal procedures in an attempt to be more open about their moderation decision-making.[35]

CONTENT MODERATION BY INFRASTRUCTURE PROVIDERS

Content moderation decisions by intermediaries acting at the infrastructure level (such as network and cloud security services) raise additional concerns. Their decisions, especially if internet infrastructure keeps consolidating, can result in rendering entire websites and services inaccessible. That is an extreme measure that should be carefully considered and evaluated, taking into account clear rules and principles of necessity and proportionality in a way similar to the considerations made by states when ordering the shutdown of entire sites or services.[36]

RISKS LINKED TO CONTENT CURATION

RISKS OF THE "ATTENTION ECONOMY"

The business model of some online platforms relies heavily on user attention and engagement, which are considered and treated as an economic resource. The time users spend on online platforms is one of the key factors that determine platforms' economic gain through the display of advertising. Most online platforms curate their news feeds and search results in order to increase relevance and engagement, and profit from targeted advertising based on heavily privacy-invasive forms of online tracking. This can create an incentive for the amplification of sensationalist, homogeneous, and low-quality information.

This "attention economy" and the competition in "attention markets" has caught regulators unprepared.[37] Current government regulatory models have been largely unable to effectively address the issues imposed by invasive, non-transparent, and non-consensual forms of information recommendation

– such as "timelines" and "news feeds" of social media services – that can affect the right to form and express opinions online.

RISKS FOR FREEDOM OF OPINION

The freedom to hold opinions without interference is an absolute right. That means it does not admit exceptions or restrictions and implies the freedom to develop beliefs, ideologies, reactions, and positions without coercion of any kind.[38]

The way in which online platforms design interfaces and curate information streams can have consequences for the right to freedom of opinion. Platforms decide what information is presented more prominently in user interfaces, combining organic and sponsored content, in ways that are not always transparent or enabling of user choice. The way in which platforms allow and often incentivize reactions through nudges can also influence the formation of opinions. Most of those mechanisms lack the scrutiny of users and the general public, including researchers and regulators.

LACK OF INFORMATION REGARDING CURATION MECHANISMS

Users are usually not aware or given enough information as to how recommendation algorithms and information feed the hierarchization of content. Some researchers have found that recommendation algorithms may lead users to increasingly radicalized or disturbing content.[39] In another very controversial case, Facebook was found to manipulate user timelines to experiment on the effect of content curation on human emotions.[40]

Lack of transparency deprives users, regulators, and the general public of discovering and addressing issues like the ones mentioned. Transparency is essential to enable independent auditing and avoid undesired outcomes in automated curation and ad-targeting, which in the context of platforms, can have discriminatory effects.[41] Additionally, transparency in curation mechanisms is a precondition for enabling true user agency over the tools that help shape their informational landscape.[42]

DATA PROTECTION CONCERNS

Content curation aimed at maximizing ad revenue is based on widespread data collection, typically carried out without proper transparency, user consent, or other basic data protection safeguards. Curation should be taking place in a context of strong enforcement of data protection laws, such as the General Data Protection Regulation (GDPR) in the European Union. Where countries do not yet have adequate privacy and data protection frameworks, they must prioritize their adoption to achieve the highest protective mechanisms for the right to privacy. [43]

THE HUMAN RIGHTS RISKS SPECIFIC TO CO-REGULATION

Co-regulatory models are those that combine elements of state and self-regulation and they can either be based in regulations or be established through voluntary agreements. Typically, co-regulation entails formal cooperation

between state and private actors in joint institutions. These models may prove to be a beneficial form of governance provided they are transparent, truly participatory, informed by research, based on evidence, subject to effective democratic oversight, and accompanied by appropriate redress mechanisms.

Recently adopted co-regulatory mechanisms such as the E.U. Code of Conduct on Countering Illegal Hate Speech do not meet such important conditions.[44]

LACK OF MEANINGFUL AND ROBUST TRANSPARENCY AND TIMELY PUBLIC PARTICIPATION

Some co-regulatory initiatives are introduced through direct negotiation between governments and platforms, in ways that are not transparent enough and that don't allow for the timely participation of interested parties or the general public. Moreover, the government entities in charge of designing the commitments in some of those agreements are usually not members of parliaments, and they don't always subject the agreements to legislators for approval afterward.

This can lead to issues of accountability since it is not possible for the general public and the subjects of the governments involved to see the negotiations and scrutinize the interests, proposals, and concessions of each party. This was the case for the initial implementation of the the aforementioned "ChristChurch Call"[45] and the first transparency report from the Global Internet Forum to Counter Terrorism (GIFCT).[46] At the date of publication of this report, the GIFCT and the ChristChurch Call are going through a reform process that aims at solving some of these issues, which constitutes a step in the right direction.

RUSHED INITIATIVES AND LACK OF LEGAL CERTAINTY

Similar to state regulation, co-regulatory efforts can also be rushed through as an attempt to provide easy, short-term fixes to highly complex phenomenon. Often, co-regulatory initiatives are designed in an informal manner, without adequate grounding in binding legal frameworks, creating a situation of legal uncertainty that is highly problematic for all parties, and especially for users.

Recent E.U. co-regulatory initiatives constitute examples of hasty policy-making that have resulted in shifting more pressure and responsibilities onto online platforms. They include the E.U. Code of Conduct on Countering Illegal Hate Speech, the Recommendation on measures to effectively tackle illegal speech online,[47] the Guidelines on Freedom of Expression Online and Offline,[48] as well as the E.U. Internet Forum.[49]

These soft law co-regulatory measures have led to gradually shifting more and more responsibility for illegal user-generated content to online platforms. While voluntary in nature, they serve as a mechanism to pressure private actors to do "something" about the spread of illegal content under the threat of hard law regulation if they do not deliver to policy makers' expectations. The lack of legal certainty leads to possible over-compliance which in practice results in illegitimate takedowns.[50]

V. Guidance: 26 recommendations for content governance that respects human rights

HUMAN RIGHTS RESPONSIBILITIES FOR ONLINE PLATFORMS AT SCALE

The decisions that platforms and states make to govern content impact the right to freedom of expression to a different extent depending on the features of the service in question and the broader context. For example, in a small or limited-access discussion forum that hosts a couple hundred people engaging in debate on specific topics, these decisions can interfere with an individual's ability to express an opinion and access information in that specific space. At scale, the capacity for interference is multiplied: a dominant social media platform can effectively shut entire populations out of a shared discussion, in a context where access to competing or alternative services may be limited. This implies additional obstacles for the exercise of rights, not only to the freedom of expression and access to information, but also rights such as freedom of association, negatively impacting participation in political life and other social and cultural rights.[51]

Decision-makers must consider the characteristics, functioning, and context of platforms to identify the challenges to user rights derived from their use. Those in the public and private sector should derive conclusions and make decisions based on careful, timely, and participatory study.

Following are some examples of impact to rights derived from the characteristics and functioning of platforms.

IMPACT DERIVED FROM THE FUNCTIONING OF PLATFORMS

Some platforms provide a space for two sides of a market to meet, or for users to upload and exchange creative content. When an online platform does nothing to influence the exchange of content, they have in many cases been considered and treated as a simple intermediary. These platforms have benefitted from legal protections that limit liability for content their users may create and upload.

However, the most popular online platforms today do not act as mere conduits of online communication. They are not acting exclusively as passive intermediaries.[52] Some have features that represent more than the hosting of content. These platforms organize and curate user-generated content based on users' behavioral data. When they take action to boost user engagement, such as by prioritizing or quantifying the popularity of certain types of sensational content, including content for harassment and disinformation campaigns, and use algorithms to personalize content, the act of content moderation can become a commodity from which the platforms benefit.[53] Increasing engagement means that users spend more time on the platform, share more of their data, and are exposed to more ads.

The way that user-generated content is organized and presented is relevant for human rights. Use of tools and strategies for curation, such as recommendation algorithms and organizing content in a hierarchy in a news feed, can have consequences for user rights, and how a company structures or modifies the hierarchy of information is therefore important to protecting these rights. Companies can enable users to take an informed, active role in the configuration of recommendation algorithms, or they can relegate users to a more passive role. When user agency is limited and there is no meaningful transparency, a service should have increased responsibility to provide safeguards in accordance with human rights principles.

IMPACT DERIVED FROM MARKET DOMINANCE / SCALE

However we choose to categorize online platforms, those making content governance decisions, particularly governments when regulating, must consider the current context, in which several large platforms that operate at a global level control an enormous market share and make highly sophisticated decisions regarding content. These platforms are often referred to as gatekeepers of the information society[54] because they exercise substantial power over public discourse. They are also in an extraordinarily powerful position to directly or indirectly affect the behavior of users of their services. Hence, these platforms should bear a higher degree of responsibility toward their users as well as to the general public.

IMPACT DERIVED FROM THE TECHNICAL UTILITY OF PLATFORMS AND SERVICES

In addition to social media platforms, there are other communications intermediaries whose decisions have enormous power over public discourse. Internet service providers, network security services, and hosting providers, among others, are examples. They control important parts of communications infrastructure and are often among the services that users and site owners may choose from. For example, the decision of network security provider Cloudflare to stop providing service to certain sites sparked concerns about the lack of clear rules for the case and the difficulties of transparency and accountability for such decisions.[55]

The impact that a platform or service can have on public discourse and user rights varies according to how they function (such as whether they actively curate and organize the display of information); their technical capabilities (such as whether they carry out important technical functions); and their position in the market (such as whether they are dominant players). Those making decisions regarding content governance — platforms when they engage in self regulation and governments when they regulate — must consider these differences in impact to provide solutions that are fit-for-purpose and appropriately address human rights concerns.

Defining market dominance when discussing platform regulation

Identifying which online platforms have become so dominant that they may require special regulatory approaches is a key public policy question of our time. Regulators around the world are grappling with this to decide whether and how to reform their competition policy frameworks or when considering new platform regulation.

Different authors outline different elements to consider when creating rules that define and interpret dominance online. Without being exhaustive, here are some that regulators could incorporate into the analysis:

- number of users impacted / potentially impacted by the decision
- degree of participation in editing or curation of content (in content curation activities)
- market dominance
- legal / economic status of the service (e.g. classification as a common carrier, operation under limited licensed frameworks, etc.)
- position of technical control over communications (infrastructure and connectivity operators
- cost of exclusion [56]

A. Recommendations for state regulation that respects human rights

1. ABIDE BY STRICT DEMOCRATIC PRINCIPLES

When setting up rules to govern states' obligations and intermediary responsibilities for protection of users' human rights, national authorities should adopt a formal legal framework to guarantee legal certainty, legitimacy, and harmonization of regimes. In order to secure the protection of the right to freedom of expression, a formal legal instrument must contain protective safeguards that are established through a democratic process that respects the principles of multistakeholderism and transparency and is subject to public debate. Whatever form the law will have, it must be foreseeable and accessible.[57]

The adoption of such a legal instrument does not necessarily exclude the use of co-regulatory models. However, any co-regulatory model should be grounded in the foundation of a binding legal framework adopted by state actors, in order for all the necessary accountability mechanisms to be present, as a way to prevent private actors making any non-transparent and possibly arbitrary decisions.

The formal legal framework should have a clearly defined scope, contain the definition of associated procedures – such as notice-and-action – and set high transparency standards for both states and online platforms. Most importantly, the legal framework must reinforce a clear distinction between the obligations of states and the responsibilities of private actors to protect users' human rights.

Even the voluntary codes of conduct of co-regulatory nature adopted by international organizations and negotiated directly with private actors, such as the E.U. Code of Conduct Against Illegal Hate Speech, must comply with an established legal framework.

Normally, under state regulation, independent and impartial judicial and other competent authorities are responsible for the implementation and oversight of the adopted legal or regulatory regimes. Any legal framework that seeks to regulate interference with content dissemination has to comply with the principle of legal certainty. Legislation applicable to internet intermediaries and online users must be accessible to all actors that fall into its scope, including all affected parties. The law has to be foreseeable, so everyone is fully aware of their obligations and rights.

What we advocate for

Any form of state regulation should be based on a formal legal framework adopted through a transparent, inclusive, democratic process. The regulation should have a precisely defined scope that is foreseeable and accessible to all actors, including users. It must establish carefully balanced definitions of individual procedures and high transparency standards. It must contain a clear distinction between the obligations of states and the responsibilities of private actors to protect users' human rights.

2. ENACT SAFE HARBORS AND LIABILITY EXEMPTIONS

Intermediary liability regimes are a fundamental piece of any state regulation related to content governance. Their objective has been to establish a balance between enabling and supporting innovation in the market for digital services and strengthening the protection of the right to freedom of expression and information of internet users.

While intermediary liability regimes exist in national jurisdictions across the globe, the U.S. and E.U. models have historically been decisive in shaping other national legal models. Section 230 of the Communications Decency Act (CDA) immunizes internet platforms from liability for most user-generated content in the U.S., including removal of lawful speech from and by online platforms. Thus, it is the strongest safe harbor provision in the world. The second significant U.S. law that served as a strong inspiration for the E.U. legislators when creating their own intermediary liability framework is the Digital Millennium Copyright Act (DMCA). Among other features, the DMCA contains limitations on copyright liability for internet service providers when they act as a mere conduit for infringing material. Finally, the European e-Commerce Directive provides for liability exemption for intermediaries, if they remain a mere conduit of information produced by users. The rationale behind these regimes is twofold: strengthen the protection of the right to freedom of expression online, while at the same time unleashing the innovative power of the internet ecosystem. These laws, and other similar ones around the globe, arguably paved the way for the internet we know today.

The protection to online platforms granted by Section 230 of the CDA is rather unique and was heavily influenced by the large protective scope of free speech granted by the First Amendment of the U.S. Constitution. According to the U.S. courts' interpretation of Section 230, the safe harbor clause precludes not only strict liability for platforms but also intermediary liability for distributors, such

as website operators.[58] The main purpose of the clause is to grant statutory immunity to online intermediaries regardless of the type of service they provide. In recent years, both regulatory models have come under increased scrutiny, with some arguing they weaken protections of online users. In the U.S., recent court interpretations have resulted in full immunity for platforms, while leaving victims of wrongdoing with no remedy.[59]

The European E-Commerce Directive, meanwhile, aims to harmonize minimum standards for intermediary liability across the E.U. member states. The liability exemption regime under Articles 12-15 differs from the U.S. safe harbor clause. Under the E.U. regime, once an intermediary has actual knowledge of illegal content on its platform, it must act expeditiously or be held liable. To enable this attribution of liability, the European conditional model of liability is directly linked to a notice-and-action procedure that is not provided by law. Its implementation is left in the full discretion of E.U. member states.

The main difference between the safe harbor provision in Section 230 and the European approach is that Section 230 ensures that intermediaries will not be held liable if they edit, filter, or remove a piece of content, even if the content enjoys the protection of the First Amendment. The current European framework lacks the safe harbor protection for online platforms that seek to address illegal and harmful content more proactively, even if they are often pushed by policy makers and state actors to do so via soft law co-regulatory measures. Such a lack of legal certainty incentivizes over-removals and over-compliance with policy makers' wishes in order to escape the threat of legal liability.

What we advocate for

We support safe harbor clauses that enable content moderation in ways that respect human rights. Such clauses should be established through clear rules by a formal legal framework. Any "voluntary" proactive measures imposed on intermediaries via state pressure are not acceptable. Under no circumstances should the delegation or "backsliding" of government functions on self-regulation be permitted in a democratic society.

- Intermediaries should be protected from liability for third-party content by a safe harbor regime. However, we oppose full immunity for intermediaries because it prevents them from holding any kind of responsibility, leaving victims of infringement with no support, access to justice, or appeal mechanisms.
- In order to strengthen the principle of legal certainty and predictability for intermediaries, as well as online users, rules that protect intermediaries must be clear and precise, while enabling ways to address illegal content when it is either manifestly illegal or when the intermediary is placed on notice that it is illegal.
- Strict liability regimes are always inappropriate to address illegal content online, since they can create incentives for platforms to over-police content.

3. DO NOT IMPOSE A GENERAL MONITORING OBLIGATION

A general monitoring obligation, which state actors impose on intermediaries, is a mandate to undertake active monitoring of the content and information that users share, usually via automated measures for content recognition, applied indiscriminately and for an unlimited period of time.[60] This type of monitoring violates the right to freedom of expression and therefore should never be imposed on online platforms.

According to the Manila Principles,[61] intermediaries should never be required to monitor content proactively as part of an intermediary liability regime. The Council of Europe Recommendation on the roles and responsibilities of internet intermediaries[62] establishes that state authorities "should not directly or indirectly impose a general obligation on platforms to monitor content which they merely give access to, or which they transmit or store, be it by automated means or not." In his 2018 report on the promotion and protection of the right to freedom of opinion and expression,[63] U.N. Special Rapporteur David Kaye clarifies that states should refrain from establishing laws or arrangements that would require the "proactive" monitoring or filtering of content, as it would be both inconsistent with the right to privacy and likely to amount to pre-publication censorship. However, voluntary general monitoring exercised by private actors in their own discretion is not currently prohibited by any legislation.

The current legal regime regulating intermediary liability for user-generated content in Europe prohibits states from imposing a general monitoring obligation of user-generated content on online platforms. This provision, together with a conditional model of liability stipulated in Article 14, seeks to strengthen the protection of the right to freedom of expression and information of individual online users.

THE DIFFERENCE BETWEEN SPECIFIC AND GENERAL MONITORING OBLIGATIONS

It is near impossible to provide a clear distinction between specific and general monitoring of online content in practice. Based on the Court of Justice of the European Union (CJEU) judgement in *L'Oreal and Others*,[64] specific monitoring is conduct exercised by a platform that must concern infringements of the same nature by the same recipient of the same right and only for a strictly determined and limited period of time.[65] Further, specific monitoring has to target a specific case, that is, it must be limited in terms of the subject and the duration of the monitoring.[66] Although the aforementioned case law concerns copyright infringement, the applied filtering technique remains the same, regardless of the different types of targeted content.

In practice, specific monitoring of online content would be performed using automated tools such as content recognition technologies. In order to be effective, these technologies would have to be applied to all user-generated content hosted by online platforms, regardless of the different context for the content in question. Therefore, specific monitoring may lead to imposing the obligation on online platforms to prevent the upload of illegal content and thus,

to actively monitor all content on their platforms in order to achieve that end.[67] This would enable the circumvention of the general monitoring prohibition that currently exists under European legislation.

Therefore, considering that specific monitoring can easily turn into general monitoring, specific monitoring should not be formulated as a legal obligation imposed on intermediaries by a formal legal framework.

What we advocate for

No general or specific monitoring obligations should be imposed by states on private actors.

Any use of content recognition technologies that results in the indirect exercise of general monitoring should not be mandated nor allowed by law.

No legal provision should ever mandate, incentivize, or give platforms any sort of indication that they should be proactively filtering content before it is uploaded.

4. DEFINE ADEQUATE RESPONSE MECHANISMS

Very few countries specify notice-and-action procedures in national laws. The lack of harmonized procedures has led to serious disparities in the implementation of intermediary liability regimes in Europe and beyond. To guarantee adequate responses that are tailored according to the specific category of user-generated content, a formal legal framework should establish specific procedures for notification mechanisms.

Notification mechanisms also need to be visible, easily accessible, user-friendly, and contextual. As an example of bad practice in this regard is Facebook's compliance with the German Network Enforcement Act (NetzDG). The special notification form required by German law, as deployed by Facebook, is hardly visible to its users and its interface design is very poorly done, especially when compared to the mechanism designed by Facebook to flag content that allegedly violates Facebook Community Standards, which is clearer and easier to use.

DIFFERENT TYPES OF CONTENT SHOULD REQUIRE DIFFERENT FORMS OF NOTIFICATIONS.

If the content is publicly accessible (that is, visible even to those who are not subscribers), then it should be possible for people who are not signed in to submit a notice. It is important that specifically tailored notice-and-action mechanisms for concrete categories of illegal online content or activities are accompanied with proper human rights safeguards that limit their potential intrusiveness. The system of specifically tailored notice-and-action procedures enables a better mitigation of human rights conflicts in cases concerning intermediary liability.

In order to make notification mechanisms more effective, they should be easy to use and maximize the information that is given to users. Therefore, platforms should provide a list of reasons for submitting the notice.

5. ESTABLISH CLEAR RULES FOR WHEN LIABILITY EXEMPTIONS DROP

As explained above, under certain legal frameworks, when platforms have "actual knowledge" of the infringing content, they get stripped of their safe harbor protections.The law needs to establish clear standards to determine when and how communication intermediaries obtain "actual knowledge" of illegal content on their platforms.

What we advocate for

In most cases, only orders issued by a court or an independent impartial administrative body can constitute actual knowledge of the illegality of third-party content. This requires the expeditious reaction from the communications intermediary to retain its legal immunity.

As the Council of Europe Recommendation on the roles and responsibilities of internet online platforms stipulates, a private notification can amount to actual knowledge only in the case of manifestly illegal content and provided that:

- the content is manifestly illegal
- the manifestly illegal content can be recognized and identified by the majority of intermediaries
- the type of content is clearly defined by law and a notice about its presence on the platform doesn't lead to any kind of proactive monitoring obligation.

However, under no circumstances should intermediaries be obliged or encouraged to actively search for manifestly illegal content.

6. EVALUATE MANIFESTLY ILLEGAL CONTENT CAREFULLY AND IN A LIMITED MANNER

A legal framework has to specify what type of content is considered manifestly illegal and which vulnerable groups should receive a special protection (for example, minors). Content is manifestly illegal when it is easily recognizable as such without any further legislative or factual analysis by platforms.[68] A typical example of such content is child exploitation materials on the internet.

The list of manifestly illegal content should be determined by law and has to be specific, referring to concrete definitions within national criminal codes, civil law provisions, and other relevant legal measures. Since most illegal content requires an analysis of context, actors, and behaviors, the cases of manifestly illegal content should be defined and interpreted in a restricted manner. Most importantly, all relevant legislation, and especially national criminal law provisions, should be revised to comply with the international human rights law standards stipulated in the International Covenant on Civil and Political Rights (ICCPR)[69] and European Convention on Human Rights (ECHR).[70]

In the case of manifestly illegal content, the law could require platforms to temporarily restrict access to the content upon obtaining knowledge through private party notification, before receiving a court or independent administrative

body's order requiring them to remove the content. In all cases, the review from an independent adjudicator should be performed without delay.

What we advocate for

Failure to remove adequately defined manifestly illegal content after a private notification by a third party should be the only situation when platforms might be held liable for not removing content without an order from an independent adjudicator.

Platforms should never be held liable for not removing content which is not manifestly illegal according to specific legal provisions.

The abuse of private notifications should be discouraged in regulation, by establishing appropriate and proportionate sanctions against the users or trusted flaggers involved in it.

7. BUILD RIGHTS-RESPECTING NOTICE-AND-ACTION PROCEDURES

Notice-and-action procedures are mechanisms online platforms follow for the purpose of combating illegal content upon receipt of notification.[71] Notice-and-action is a broad term that comprises several mechanisms with different types of responses to illegal content. They are all, however, initiated by a notice.[72] Formal legal frameworks need to establish the most suitable notice-and-action mechanisms.

Different types of illegal online content and activities may require different responses specifically tailored to the type of user-generated content that they are supposed to tackle. However, the law has to clearly define the procedures and provide appropriate safeguards for their application by states.

This report proposes the following notice-and-action procedures for addressing specific types of user-generated content. The proposal departs from extensive research conducted by various experts in the field of content governance.[73]

First, for copyright infringement, notice-and-notice mechanisms have been proven to provide the most balanced measure for tackling copyright material shared on platforms.[74] Under notice-and-notice, an intermediary receives a notification with a complaint, which they then forward to the content provider.[75] The involvement of the intermediary ends there. Once the content provider receives the information from the platform, the case rests in the hands of alleged primary wrongdoer, that is, the user. The notified party then has to choose whether it wants to remove the content or to respond to the notification within the prescribed period of time. [76]

Second, for more context-dependent user-generated content and specifically in case of alleged defamation, we suggest implementing a notice-wait-and-takedown mechanism. Under this procedure, intermediaries have to forward any notice concerning allegedly illegal defamatory content to that content provider and then wait a week before enforcing the blocking or removal. This "softer interpretation

of notice-and-takedown"[77] guarantees a possibility of response to the content provider before any action is taken. Therefore, the mechanism enables prevention of wrongful content takedowns and gives an opportunity to users to be heard out.

In all circumstances, notice-and-judicial takedown, meaning the order to block or to remove user-generated content has to be issued by a judicial authority, should always be available to all online users, including both content providers as well as victims of infringement.

Ideally, safeguards – like the ones listed in the box below – should be directly built within notice-and-action procedures. In order to limit undesired outcomes of procedures, a particular type of notice-and-action should apply based on its pursued aim and the potential risk to the protected rights.

What we advocate for

In order to ensure the protection of fundamental rights, the details of notice-and-action procedures need to be clearly defined by law and be fit-for-purpose.

States must regularly evaluate the possible unintended effects of any restrictions before and after applying particular notice-and-action procedures.

States are obliged to seek the least intrusive measures for human rights of users.

The following notice-and-action procedures should be considered as adequate, determined by the type of infringement at stake as well as the category of content:

- Notice-and-notice
- Notice-wait-and-takedown mechanism that enables a content provider to file a counterclaim
- Notice-and-judicial takedown, where courts review the legitimacy of content removals, should always be available to all users, regardless of the type of content
- Private notice-and-takedown should only be used in the limited cases of content that is legally defined as manifestly illegal

States should abstain from adopting notice-and-stay-down mechanisms that establish an obligation to prevent the content from ever being available in the future, usually through automated measures that imply general monitoring.

WHAT IS A VALID NOTICE?

In order for a notice to be valid, it has to contain sufficient information for platforms to act upon. It needs to be precise and adequately sustained. The conditions that a valid notice needs to meet should be specified in law. However, concrete requirements for a valid notice should be determined by the type of content in question.[78]

What we advocate for

Basic minimum requirements of a valid notice:

- Reason for complaint
- Location of the content
- Evidence for the claim
- Consideration of limitations, exceptions, and defense available to the content provider
- Declaration of good faith

Notices submitted by states should be based on their own assessment of the illegality of the notified content, in accordance with international standards. Language for content restrictions should provide for notice of such restriction being given to the content producer/issuer as early as possible, unless this interferes with ongoing law enforcement activities. Information should also be made available to users seeking access to the content, in accordance with applicable data protection laws. Users should not be forced to identify themselves when submitting the notice and they should provide their contact details only on a voluntary basis.

REQUIREMENTS FOR "TRUSTED FLAGGER" PROGRAMS

Trusted flaggers are entities with specific expertise and dedicated structures for detecting and identifying illegal online content.[79] Only independent bodies with specific expertise should be entitled to become a trusted flagger. The legal framework should clearly stipulate the criteria for becoming a trusted flagger. Furthermore, any legal framework regulating intermediary liability should contain proper legal safeguards for the impartiality and independence of those trusted flaggers, which would safeguard more balanced decisions. Flaggers should lose their status if any conditions for exercising their function are violated. Under no circumstances should the conditions for the institution of trusted flaggers be determined solely by private platforms.

What we advocate for

Trusted flagger programs are acceptable only if the requirements for becoming one as well as the legal safeguards for their independence are governed by law. Anything else would ultimately shift states' obligations to private platforms.

8. LIMIT TEMPORARY MEASURES AND INCLUDE SAFEGUARDS

Temporary blocking could be applied as a temporary measure in circumstances where the infringement is time sensitive but only for specific types of illegal content (emergency measures). In cases when potentially illegal content is being contested, temporary measures would allow the blocking of access to a specific piece of content pending the resolution of the conflict or a response from the content provider or an independent adjudicator.[80] Temporary emergency

measures must be strictly limited to avoid state abuse of this tool to make other kinds of content unavailable without an appropriate procedure for the determination of its illegality. Most illegal content, such as content that infringes copyright, for example, does not fit the urgency-related nature of temporary measures. A law should clarify the specific cases in which temporary measures can be applied.

Any time frames designated for steps to be taken as part of notice-and-action procedures – including for temporary measures – should be clearly determined by the legislative framework. Time frames for content removal cannot be too short, as this will incentivize illegitimate takedowns of lawful content. A number of recent legislative proposals are an example of this, such as the E.U. proposed Regulation for combating the dissemination online terrorist content, which pushes for a one-hour time frame for content removal.[81]

Such an extremely short time limit should not be considered as proportional to its legitimate aim, even if it is limited to exceptional cases. Due to the short time limit, there is no real possibility for a content creator to appeal a platform's decision and consequently, to seek the appropriate remedy in case of an illegitimate takedown. Therefore, a very short time limit imposes serious threats to the right to freedom of expression and information as well as the users' right to a fair trial and adequate remedy.

Even though in some jurisdictions there are discussions of specific time frames,[82] there is no research-based evidence to support a one-hour deadline for compliance with content removal. Appropriate time frames would depend on the regulatory and operational context of each jurisdiction.

What we advocate for

Here are some time-related elements that should be specified in law, as suggested by the Council of Europe Recommendation on the roles and responsibilities of internet intermediaries:

- the time limit for forwarding the notification to content provider
- the time limit for counter-notification submitted by the content provider
- the time limit for decision-making about content removal or its maintenance on the platform
- the exact time frame for delivering informed decisions to all involved parties
- the period of time to start a judicial review of an intermediary decision

9. MAKE SANCTIONS FOR NON-COMPLIANCE PROPORTIONATE

The only situation when intermediaries should be held legally liable is when they fail to comply with an order from an independent and impartial adjudicator to remove the content in question or when they fail to remove statutorily defined manifestly illegal content upon private notice. Any sanctions imposed on intermediaries for non-compliance must be proportional. If sanctions become disproportionate, it is very likely that they will lead to

over-compliance, which could harm free expression and access to information shared on online platforms.[83]

What we advocate for

Any sanctions imposed on intermediaries for non-compliance with removal requests must be proportionate. They have to directly correlate with their failure to comply with the content removal or restriction order.

10. USE AUTOMATED MEASURES ONLY IN LIMITED CASES

Automated measures such as upload filters or hash database scanning of different kinds can be useful for assessing the large amounts of information that are shared on online platforms. But they have a fundamental problem: they are unable to interpret context before making a blocking or takedown decision. This implies a serious risk for freedom of expression and access to information in the majority of cases, where an interpretation of context is needed.

In order to avoid false-positives or excessive blocking, and protect the legitimate use of portions of illegal content for public interest means, such as for news reporting, the use of automated measures should be accepted only in limited cases of manifestly illegal content that is not context-dependant, and should never be imposed as a legal obligation on platforms. A typical example of this kind of content is sexual abuse against minors. In any case, legislation should abide by the following parameters.

What we advocate for

- Any use of automated tools has to be based on clear and transparent policies, including transparency mechanisms for the independent assessment of their creation, functioning, and evaluation.
- Such use has to follow a legitimate purpose (for example, restricting access to specific non-context-dependant illegal content). Legitimate purpose should always be determined by an independent judicial authority or other independent administrative body whose decisions are subject to judicial review.
- Platforms can use automated tools in accordance with their own policies – including in automated flagging mechanisms – if they are transparent and in line with international human rights standards.
- States' supported application of proactive automated measures by online platforms for content recognition cannot result in actual knowledge and consequently, legal liability.
- The use of automated tools should not result in a general monitoring obligation for platforms or other communications services.
- Automated content takedown systems implemented by platforms should allow for human review, in order to avoid false positives.

11. LEGISLATE SAFEGUARDS FOR DUE PROCESS

The following safeguards are extremely important and must be present in any regulation in order to guarantee due process for content creators. Additionally, the safeguards are necessary to provide legal certainty, predictability, and proportionality in measures, for the benefit of content providers but also platforms and users.

PROVIDE NOTIFICATION TO CONTENT PROVIDERS

To notify a content provider before any action is taken is absolutely essential for securing users' right to a fair trial and proper remedy. The notification should state the reason for the removal or blocking, provide a precise explanation of what rights the content provider has and the possibilities to appeal the decision or opt for judicial review. The only exception for the obligation to send a notification to a content provider applies in situations where such notification could hamper law enforcement activities, such as the prevention, investigation, detection, and prosecution of criminal offenses.[84] However, such an exception has to be proportional to its legitimate aim and necessary to the goal pursued, and it has to be grounded in the rule of law.

What we advocate for

Users that act as content providers should be notified before any action against their content is taken.

The notification should contain, at least:

- the reasons for removal or blocking;
- a precise explanation of the content provider's rights;
- an explanation of possibilities to appeal the decision;
- the clearly stated option of judicial redress.

Exception: endangering a criminal investigation of a serious crime. The conditions for exemptions have to be defined by law in proportional manner and be in line with international human rights law standards.

ESTABLISH COUNTER-NOTIFICATION

Notice-and-action procedures, if adopted in regulation, have to allow for counter-notification. Counter-notification enables content providers to object to individual complaints targeting their content. It is a precondition of fairness in any decision-making process. Any exception to counter-notification has to be defined by the formal legal framework.

What we advocate for

- Counter notices are a necessary tool for the right to defense in the context of notice-and-action procedures.
- Users should be able to submit a counter-notice before any action is taken against their content.
- Conditions for the use of counter notifications specified in the law should not be too demanding because it could discourage content providers from using this mechanism.
- The law needs to specify what type of content and situation may lead to an exception to the use of counter-notices.

12. CREATE MEANINGFUL TRANSPARENCY AND ACCOUNTABILITY OBLIGATIONS

In order to establish meaningful transparency that focuses on the quality instead of the quantity of content governance decisions, a formal legal framework needs to provide for clear and robust transparency requirements for both states and private actors. An appropriate legal framework would also allow for an accurate assessment of existing policies, their functioning, and their effectiveness.

Transparency is a precondition for gathering evidence about the implementation and the impact of existing laws. It enables legislators and judiciaries to understand the regulatory field better and to learn from past mistakes. To avoid hastily drafted legislation that often misses its own purpose and becomes potentially human rights intrusive, regulators have to be able to monitor how the initial objectives are being fulfilled. Proper monitoring will then determine whether there is a need for more regulatory responses or the already applicable legal framework is satisfactory. This goal can be practically implemented only if complete and relevant data about content removals performed by states and private entities is made accessible.

WHAT ARE THE REQUIREMENTS FOR TRANSPARENCY REPORTS ISSUED BY STATES?

Transparency reports should be submitted by states as well as by private actors. There are numerous public authorities that issue notices for removal of illegal content or monitor platforms in order to prevent illegal activities. For instance, under the German Network Enforcement Act (NetzDG) that imposes conditional liability for illegal user-generated content on social networks, it is the Federal Office of Justice that oversees the enforcement of the law. In July 2019, the German Federal Office of Justice – Bundesamt für Justiz, or BfJ – decided to fine Facebook for violating transparency requirements stipulated in the law.[85]

While the decision may have a positive impact on transparency, it uncovers another pressing issue: the BfJ cannot be considered a politically independent regulator, because it reports directly to the Minister of Justice. While it enforces transparency requirements for private actors, it is not subject to the same transparency threshold. The independence of a regulator might not be a serious concern in some countries that uphold democratic principles and the rule of law.

However, in countries experiencing democratic backsliding or being governed by authoritarian regimes, the lack of a regulator that is independent from the state may lead to suppression of dissent and other gross human rights abuses.

What we advocate for

In order to protect democratic discourse and the public scrutiny of political leadership, states should make publicly available and on a regular basis, valid information on:

- the number and nature of content restrictions as well as the categories of personal data that they requested from intermediaries. States' transparency reports should include all content-related requests issued to intermediaries.
- the clearly defined legal basis that their request was based on, including those based on international mutual legal assistance treaties.
- the exact steps that were taken as a result of their requests.

WHAT ARE THE REQUIREMENTS FOR TRANSPARENCY REPORTS ISSUED BY PLATFORMS?

Platforms disclose very little information about how private rules and mechanisms for self- and co-regulation are formulated and carried out. In particular, disclosure concerning actions taken pursuant to private removal requests under terms of service is "incredibly low." [86] States should require that intermediaries disclose precise, simple, and machine readable information about all interferences with users' right to freedom of expression and information, right to privacy, protection of their personal data, and other fundamental rights. This information should be easily accessible to the general public.

Only through the combination of comprehensive transparency reports by states and intermediaries can regulators as well as individual users gain a realistic picture of how online content moderation works. Meeting such a transparency threshold would create the foundation of research for evidence-based policy making. To help achieve this, private actors should adopt a meaningful and comprehensive transparency approach to their operations, sharing information about how they develop and implement the rules for content moderation.

What we advocate for

Intermediaries should include in their transparency report, at least, the following information:[87]

- the number of all received notices
- type of entities that issued them, including private parties, administrative bodies, or courts
- reasons for determining the legality of content or how it infringes terms of service
- concrete time frames for notifying the content provider before any action is taken, for filing the counter-notice, the exact time that will pass before the content is restricted, and the time frame for an appeal procedure
- the number of appeals they received and how they were resolved

13. GUARANTEE USERS' RIGHTS TO APPEAL AND EFFECTIVE REMEDY

Decision-making about user-generated content requires a careful and often complicated balancing of rights. Ultimately, errors resulting from a wrongful assessment of cases are inevitable. Therefore, the establishment of clear and easily accessible appeal mechanisms is the main guarantee of procedural fairness. In this report, we distinguish between two mechanisms: dispute settlement directly provided by platforms and judicial redress guaranteed by states.

First, all appeal mechanisms provided by communications intermediaries should be accessible, affordable, and transparent. However, they should never replace a judicial remedy granted by courts. Appeal mechanisms should be available to online users in cases of content removal as well as when intermediaries refuse to comply or ignore users' removal requests. We will explore the review mechanisms that platforms should put together as part of their responsibilities to protect human rights below.

Second, independent and impartial judicial redress must always be available to online users, especially when dispute settlements at the intermediary level are considered to be insufficient by affected parties. Even though judicial redress is always an option in theory, it is of high importance that the option of judicial redress is granted by law. Similarly, the injunction has to be available also in cases of illegitimate content removals, so content providers can equally benefit from the possibility of injunctive relief.

B. Recommendations for self-regulation that respects human rights

Content moderation and curation decisions can have ramifications not only for free expression but also other fundamental rights, such as the right to freedom of association, as well as for the enjoyment of economic, social, and cultural rights.

To prevent, address, or mitigate human rights violations on the one hand, and to promote the enjoyment of human rights on the other, we have developed basic human rights guidelines[88] for content moderation, outlined below.

Any platform that makes decisions about the speech of third parties should follow these principles, regardless of other legal obligations. As platforms grow in size, geographical reach, and influence, serving as intermediaries of public discourse, straying from them will represent heightened risk for the human rights of users. For dominant online platforms, which can have a significant impact on public discourse,[89] it is critical to interpret and follow the principles more strictly.

Making decisions about online speech is a particularly complex exercise, and one that needs careful fine tuning so that moderation or curation goals do not have unintended consequences for fundamental rights and freedoms. Any content moderation mechanisms that online platforms design and deploy should adhere to these guidelines:

1. PREVENT HUMAN RIGHTS HARMS

Platforms must consider human rights from the design of their products through the development and implementation of content moderation and curation policies and practices. This must be aimed at achieving an online environment that furthers the free exchange of ideas, empowers users, and protects the rights of vulnerable communities.

What we advocate for

Among other actions, platforms should:

- ▶ Bake in human rights protections to any new policies and services, rather than relying on a model of scaling up first and addressing abuses later
- ▶ Consult third-party human rights experts and civil society organizations regularly, especially before launching new products, features, or services

2. EVALUATE IMPACT

Platforms should also perform participatory and periodic public evaluations to determine how content moderation and curation decisions are impacting the fundamental rights of users and take the necessary steps to mitigate any harm.

What we advocate for

Platforms should:

- ▶ Share information proactively with researchers and civil society to allow them to independently evaluate the human rights impacts of content moderation and curation decisions
- ▶ Contribute, including through economic support, to the work of researchers and civil society groups performing independent evaluations
- ▶ Openly elaborate and incorporate human rights impact-evaluation protocols into their operations to streamline the work of researchers, civil society, and regulators

3. BE TRANSPARENT

All content moderation and curation criteria, rules, sanctions, and exceptions should be clear, specific, predictable, and properly informed to users in advance.

What we advocate for

In particular, platforms ought to:

- Obtain valid and informed consent from users with regard to the rules, criteria, sanctions, and exceptions that are going to be applied to their activities in the service
- Ensure that consent can be revoked in an easy and streamlined manner
- Include guidelines to explain the company's internal process for interpreting and applying content moderation rules, to ensure that decisions on content are as predictable and understandable as possible
- Detail what constitutes a violation of the rules, what the corresponding sanctions are, how appeal processes work, and how the policy will be applied
- Make all information available in the official language of the country where the service is provided and have it written in simple terms, avoiding excessive technical terminology and references to other documents
- Notify users of any changes to these rules and ensure they are explicitly accepted by users before they can be applicable
- Inform users about the collection and use of their data and make all the rights granted through data protection principles and laws available, in a way consistent with the highest standards for such protection
- Abstain from any practices aimed at "nudging," influencing, or manipulating users without their knowledge or consent. Use of content curation technology, such as news feed hierarchization or recommendation algorithms, should be made as clear and transparent as possible

4. APPLY THE PRINCIPLES OF NECESSITY AND PROPORTIONALITY

The sanctions platforms impose on users for violating content moderation rules should be proportional to the harm being addressed and take effectiveness and the impact on user rights into consideration (necessity). Severe penalties, such as the banning of a user from an online service, or termination of service to entire web pages or applications, should be a measure of last resort and only take place if there is a serious infringement or after repeated offenses. In any case, a thorough evaluation of the impact of the measure must be performed.

5. CONSIDER CONTEXT

Platforms should not apply content moderation rules in a "one size fits all" fashion. In addition to using human rights principles as a universal baseline for making content moderation decisions, platforms should take social, cultural, and linguistic nuance into account, as much as possible.

What we advocate for

To achieve that, platforms should:

- ▶ Develop and review content moderation and curation rules — and any accompanying guidelines for interpretation — with the permanent input of local civil society, academics, and users
- ▶ Invest in human resources and the development of the necessary social and cultural expertise for content moderation decisions, in order to better consider the corresponding details and consequences
- ▶ Pay special attention in cases arising from conflict zones and in situations that can impact vulnerable populations
- ▶ Evaluate individual context when judging the behavior of particular users, whenever possible, taking previous behavior, compliance with sanctions, and other factors into account

6. DON'T ENGAGE IN ARBITRARINESS OR UNFAIR DISCRIMINATION

The application of context-based, nuanced content moderation decisions should be as coherent, systematic, and predictable as possible in order to avoid arbitrariness. Platforms should pay special attention to how their content moderation rules are implemented, whether by the company moderators using internal processes or by users via reporting mechanisms, to ensure that they do not unfairly target marginalized communities.

7. FOSTER HUMAN DECISION-MAKING

Online platforms should not solely rely on automated decision-making for content moderation.

What we advocate for

If the necessities of scale or the sheer volume of user-generated information make reliance on automated decision-making necessary, online platforms should:

- ▶ Clearly inform users about the use of automated decision-making technology
- ▶ Limit the use of automated decision-making technology to content that requires less interpretation in order to be considered in violation of terms of use

In terms of content moderation decisions, platforms should also

- ▶ Provide them with the right to request a human review of their case

In the case of content curation, online platforms should inform users clearly when and where in the service or application curation technology is being used. They should also:

- Inform users about the criteria used for prioritization or recommendation in a clear manner
- Allow users to modify those criteria or opt out of content curation, whenever possible

In both automated moderation and content curation, platforms should:

- Make automated systems as transparent as possible
- Publish information about how these systems are used and the procedures behind their application
- Make the systems available for independent auditing

8. CREATE NOTICE-AND-REVIEW MECHANISMS

Communications intermediaries should notify users when a moderation decision has been about their content or speech.

What we advocate for

This notification should contain:

- adequate information about what sparked the decision
- the specific rule that was broken
- how content moderation guidelines were interpreted
- the action that will be taken
- clear instructions for submitting appeal

A notification should also contain the necessary information to ask for a review of the decision. Review mechanisms must be directly and easily accessible and be addressed within a reasonable time frame, particularly if content is made unavailable in the interim. They may be provided by the company or through recourse to an external entity, such as an oversight board or an industry-wide council.

What we advocate for

Following the recommendations of the Human Rights Council Special Representative on business and human rights, non-judicial grievance mechanisms should also be:[90]

Legitimate: having a clear, transparent, and sufficiently independent governance structure to ensure that no party to a particular grievance process can interfere with the fair conduct of that process

Accessible: being publicized to those who may wish to access it and providing adequate assistance for aggrieved parties who may face barriers to access, including language, literacy, awareness, finances, distance, or fear of reprisal

Predictable: providing a clear and known procedure with a time frame for each stage and clarity on the types of process and outcome it can (and cannot) offer, as well as a means of monitoring the implementation of any outcome

Equitable: ensuring that aggrieved parties have reasonable access to sources of information, advice, and expertise necessary to engage in a grievance process on fair and equitable terms

Rights-compatible: ensuring that its outcomes and remedies accord with internationally recognized human rights standards

Transparent: providing sufficient transparency of process and outcome to meet the public-interest concerns at stake and presuming transparency wherever possible; non-state mechanisms in particular should be transparent about the receipt of complaints and the key elements of their outcomes

Based on dialogue and engagement: focusing on processes of direct and/or mediated dialogue to seek agreed solutions, and leaving adjudication to independent third-party mechanisms, whether judicial or non-judicial[91]

9. PROVIDE REMEDIES

Platforms should provide effective remediation to users affected by its policies, products, or practices. This includes content moderation decisions, in the cases in which they cause harm to users, such as when the rules have been applied erroneously or excessively.

What we advocate for

Among other ideas, platforms could provide remediation through various pathways, according to the case. For example:

- Restoring eliminated content in case of an illegitimate or erroneous removal
- Providing a right to reply, with the same reach of the content that originated the complaint
- Offering an explanation of the measure
- Making information temporarily unavailable
- Providing notice to third parties
- Issuing apologies or corrections
- Providing economic compensation[92]

10. ENGAGE IN OPEN GOVERNANCE

Online platforms, especially dominant players, should create mechanisms for the participation of their users and other interested parties in the governance of its applications and services. Taking the feedback of those affected by content moderation and curation decisions into account is vitally important for the correct assessment of human rights risks.

What we advocate for

To facilitate this, online platforms should:

- Enable the meaningful participation of users in a timely manner and at different stages of the creation, implementation, and evaluation of content governance rules and technological developments
- Engage different groups of users, particularly those most affected by certain rules, decisions, or technologies
- Designate local points of contact to receive feedback and respond to users and civil society, whenever possible. Those points of contact should speak the local language and be versed in the social and cultural reality of the region
- Take part in truly open, independent, transparent, and participatory initiatives aimed at increasing transparency and oversight over content moderation and curation decisions

C. Recommendations for co-regulation that respects human rights

Co-regulatory mechanisms play a significant role in content governance, whether they are formal or informal. This is evident in cases where cooperation between governments and platforms is necessary to address some of the most salient challenges in regulation: the need for swift action and global coordination to address illegal content, among others.

But in order to make co-regulation a tool for addressing societal needs and extending the rights of users, important considerations need to be made. The actors involved in co-regulatory endeavors should comply with the human rights obligations and responsibilities pertaining to their functions, as outlined in previous sections of this paper. Additionally, some special considerations should be addressed.

1. ADOPT PARTICIPATORY, CLEAR, AND TRANSPARENT LEGAL FRAMEWORKS

As we note in our recommendations for states above, any co-regulatory model should be grounded in a binding legal framework adopted by state actors. This would enable the necessary accountability mechanisms that prevent private actors from making any non-transparent and possibly arbitrary decisions.

An appropriate legal framework should determine safeguards for users and an independent oversight mechanism for any co-regulatory mechanism. It needs to clearly establish and distinguish states' obligations and intermediaries responsibilities in order to protect the human rights of online users.

All co-regulatory approaches should comply with international human rights, appropriate transparency requirements, and with the principle of meaningful participation.

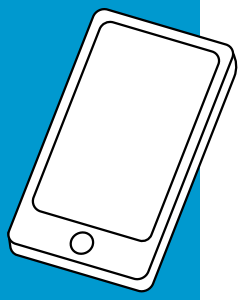
2. DON'T SHIFT OR BLUR THE RESPONSIBILITIES OF ACTORS

Co-regulatory initiatives should never be understood as a replacement for the duties or responsibilities of the individual actors involved. Governments should stop allowing or encouraging private actors to decide upon the legality of user-generated content and its restriction.

In all cases, actors should enable users' access to effective remedy, including judicial remedy and redress.

3. PREVENT ABUSE

States should avoid any action that may lead to the abuse of co-regulatory measures such as general content monitoring, intermediaries' over-compliance, and over-removal of user-generated content.



VI. Conclusion

The growth in use and ease of access to internet services has brought new challenges for public space online. The idea of an open borderless space where people can exercise their freedom to access information and express opinions has collided with threats to those ideals, arising from the actions of information gatekeepers and the exploitation of technology for violent and harmful means.

It is evident that all actors in the information space need to take action to address the negative effects of some uses of the internet. This includes governments and platforms that act as intermediaries of information online, but also other actors that contribute to the formation of public discourse such as traditional media outlets and political actors.

We welcome the predisposition of some governments and platforms to review their duties and responsibilities and to enable participatory dialogue to find solutions for illegal and so-called harmful and violent content online. But we are also worried about rushed approaches, non-transparent endeavors, and proposed solutions that are at odds with basic human rights.

There is no silver bullet to solve complex problems such as the proliferation of illegal, violent, or harmful content on the internet. Solutions solely based on censorship or criminalization of users' activity risk backfiring by affecting the rights of vulnerable groups, independent media, artists, activists, and human rights defenders, among others. Protecting and preserving their opinions is as important for strengthening our democracies as keeping people safe and fighting criminal behavior.

Governments need to listen and act in a democratic and transparent way. Regulatory proposals should be evidence-based and go beyond criminalization to consider economic concentration, business models, data protection, user education, and social inclusion as crucial issues for addressing illegality and violence online. Harmful activity online is often a reflection of social problems. Political actors may do more to address these problems by abandoning incendiary rhetoric that leads to division and has the potential to incite violence.

Platforms need to step up in their responsibilities, in accordance with the impact that they have on public discourse. This includes abiding by human rights standards in their content moderation and curation practices and baking transparency, privacy, data protection, and impact assessments into all their practices and products. Other platforms that participate and enable public discourse, such as traditional media outlets, also need to make a commitment to improve their practices.

Any actions undertaken by governments and platforms to govern content should have a very specific objective in mind: to enable a diverse, free, open, and safe online space. Our hope is that this guide can help ensure a healthy public discourse, strengthen our democracies, and preserve the enjoyment of human rights through free expression and access to information for everyone.

VII. Glossary

To assist decision-makers in discussion of approaches for human-rights-based content governance, following is a short glossary of the terms we use in this paper as we have defined them.

Actual knowledge: Refers to when an online platform becomes aware of the existence of specific illegal content on the platform. Actual knowledge is usually obtained via notifications coming from various sources, including judicial order or trusted flaggers. In this paper, we specify that only a valid notification can constitute actual knowledge.

Content curation: A form of self-regulation. By curating content, online platforms make decisions regarding the reach, prominence, or amplification of certain material. This determines which and how many users are exposed to select material and the way in which it is presented. Content curation requires the use of data about users' interactions on the platform and is often facilitated by machine learning systems.

Content governance: In this paper, content "governance" refers to the complex processes and interactions delineated by public and private stakeholders to create and enforce rules for governing content online. We use content governance as an umbrella term for three specific regulatory approaches to user-generated content: self-regulation, co-regulation, and state regulation.

Content provider: The entity that uploads content to a platform, whether an individual or a company or other legal entity.

Co-regulation: Co-regulatory regimes can be understood as self-regulation that is actively encouraged, supported, and sometimes monitored by public authorities. Typical examples of co-regulatory mechanisms are voluntary codes of conduct resulting from dialogue between private actors and national or regional state authorities.

Intermediary liability: Generally refers to the legal responsibility of online platforms for illegal or potentially harmful activity exercised by their users through their services. Intermediary liability regimes began to emerge in national legislation in the 1990s, and the definition of this term can differ across jurisdictions globally. This paper looks specifically at the liability models in the United States and European Union due to the global influence they have had in shaping legal responses to content governance.

Manifestly illegal user-generated content: Online content is manifestly illegal when it is easily recognizable as such without any further legislative or factual analysis by platforms. This includes, for example, material depicting child abuse.

Notice-and-action: An umbrella term for a variety of mechanisms aimed at eliminating illegal, infringing, or potentially harmful content online. Notice-and-action enables online platforms to make decisions about which content remains accessible and which should be removed without involving public oversight. Notice can be submitted directly by a user, for instance, to flag content that allegedly violates the terms of service, or by a public authority in the form of

judicial order, after which the platform takes action.

Online platforms: Enable individuals to use information and communication technologies to facilitate various types of interactions among their users. They exist in a variety of forms, from social media to platforms for economic collaboration. The interactions among users generate data that are consequently collected and used by platforms. While the use of such data significantly enhances users' experience, it also raises myriad human rights concerns that are discussed in this paper.

Principle of legal certainty: Guarantees that the application of the law to a specific situation must be predictable and therefore, that it must provide to its subjects clarity on how to regulate their own conduct. Legal certainty is a fundamental principle enshrined in national constitutions as well as in international law. It is an essential precondition for proper functioning of the rule of law in democratic society.

Principle of proportionality: Regulates how states exercise the powers invested in them. Under this rule, all actions of public authorities that may limit the exercise of a human right must meet the following requirements: they must be necessary in a democratic society and be the least restrictive towards users' human rights; serve a legitimate aim established in human rights law; and strike a balance between the means used and the objectives pursued.

Prohibition of general monitoring: States should not impose the obligation on online platforms to monitor all user-generated content that they transmit or store. The prohibition of general monitoring currently exists under the European legal framework.

Safe harbor: A provision in intermediary liability laws that protects online platforms from being held liable for illegal or potentially harmful activity exercised by their users through their services if those platforms meet certain criteria (which should be prescribed by law).

Self-regulation: A form of content governance exercised by online platforms based on their own terms of service or "community guidelines," usually via content moderation practices that leverage both automated tools and human moderators.

State regulation: Refers to any legally binding or regulatory instrument that local, national, or regional public institutions enact through their legislative or regulatory processes.

Terms of service: A set of rules in the form of legal agreement between the online platform and a user who wishes to use its services.

Valid notification: In order for a notification regarding content to be valid, it has to contain sufficient information for platforms to act upon. It needs to be precise and adequately sustained. The conditions that a valid notification needs to fulfil should be specified in law.

Endnotes

1. Barlow, J.P. (1996). A Declaration of the Independence of Cyberspace. Retrieved from https:// www.eff.org/cyberspace- independence
2. Gillespie, T. (2018). Custodians of the internet, p. 30-31. Retrieved from https://www. researchgate.net/ publication/327186182_Custodians_of_the_internet_Platforms_content_ moderation_and_the_hidden_decisions_that_ shape_social_media
3. This is a broad concept that encompasses different kinds of platforms acting at different levels of the online communications stack. Application providers (such as social media or search services), webpages, forums, internet connection providers, infrastructure and internet security service providers, and even device manufacturers can be considered intermediaries in user communications. It is important to remember that not all of them intervene actively in the content of information exchanges between their users. In the context of this paper, we use the term "online platform" to refer to platforms and services that act as intermediaries for user-generated speech and actively make decisions about the type or prominence of content they allow on their services.
4. Like economic, social, and cultural rights. See Lara, J.C. (2015). Internet access and economic, social, and cultural rights. Association for Progressive Communications (APC). Parts 3 and 4. Retrieved from https://www.apc.org/sites/default/files/ APC_ESCR_Access_Juan%20Carlos%20 Lara_September2015%20(1).pdf
5. The 2019 U.K. Online Harms White Paper uses the term "online harms" which includes illegal content, but also covers users' behaviors that are deemed harmful but not necessarily illegal. See United Kingdom Secretary of State for Digital, Culture, Media & Sport et. al. (2019). Online Harms White Paper. Retrieved from https://assets.publishing.service.gov.uk/ government/ uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf
6. Facebook. (n.d.). Community Standards. Part III, Objectionable Content. Retrieved from https:// www.facebook.com/communitystandards/objectionable_content
7. Ranking Digital Rights. (2019). Corporate Accountability Index. Chapter 4 Freedom of Expression. Retrieved from https:// rankingdigitalrights.org/index2019/report/freedom-of-expression/
8. The Digital Security Helpline is a service provided by Access Now that offers real-time, direct technical assistance and advice to civil society groups and activists, media organizations, journalists and bloggers, and human rights defenders. The Helpline helps organizations assess digital security risks, resolve existing problems and adopt best practices, helping them get into a secure mindset for the future. The service is available 24 hours a day in nine languages: English, Spanish, French, German, Portuguese, Russian, Tagalog, Arabic, and Italian. For more information on Access Now's Digital Security Helpline visit https://www.accessnow.org/help/
9. "Governance can refer abstractly to all processes of governing. It supplements a focus on the formal institutions of government with recognition of more diverse activities that blur the boundary between state and society. It draws attention to the complex processes and interactions involved in governing. Governance can also refer, more concretely, to the rise of new processes of governing that are hybrid and multi-jurisdictional with plural stakeholders working together in networks. It describes recent changes in the world." Bevir, Mark (2012). Governance: A Very Short Introduction. OUP Oxford.
10. These rules might have different names or refer to other documents like "community guidelines" that set the rules for acceptable expression on the platform.
11. The concept of speech as used in this paper includes user expression in all the forms that are technically possible on a platform, such as via text, images, videos, etc.
12. This type of decision is also referred to as "content ranking" or "content distribution," among other names.
13. Platforms can make decisions on exposure/reach on the basis of increasing "relevance," responding to user interests, policy decisions, etc. With regard to methods for content curation, platforms could determine what to prioritize independently or implement an automated

decision-making system. They could prioritize certain content for a whole category of users or tailor it to the preferences of individuals. In all cases, the lack of awareness of users to algorithmic bias is a concern. See Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethics and Information Technology, 15(3), 209–227.

14. Mardsen, C. (2011). Internet co-regulation: European law, regulatory governance, and legitimacy in cyberspace, p. 44. Cambridge University Press.
15. Christchurch Call to eliminate terrorist and violent extremist content online (2019). Retrieved from https://www. christchurchcall.com/
16. Global Internet Forum to Counter Terrorism (n.d.). Retrieved from https://www.gifct.org/
17. Human Rights Council (2018). Report of the Special Rapporteur on violence against women, its causes and consequences of online violence against women and girls from a human rights perspective. A/HRC/38/47. Retrieved from https://www. ohchr.org/EN/HRBodies/HRC/RegularSessions/Session38/Documents/A_HRC_38_47_EN.docx
18. Counting censorship and harassment cases involving YouTube, Facebook, and Instagram.
19. Human Rights Council (2018). Op.cit.
20. Karp, P. (2019). Australia passes social media law penalizing platforms for violent content. The Guardian. Retrieved from https://www.theguardian.com/media/2019/apr/04/australia-passes-social-media-law-penalising-platforms-for-violent- content
21. See for instance the institute of referrals encompassed in the E.U. proposed Regulation combating the dissemination of online terrorist content.
22. European Court of Human Rights (2018). Stomakhin v.Russia, No. 52273/07, paras. 93-131; (2008). Leroy v. France, No. 36109/03; (2014). M'Bala M'Bala v. France, No. 25239/13, paras. 37-39.
23. German Parliament (2017). Netzdurchsetzunggesetz -Network Enforcement Act-. Retrieved from German Law Archive at https://germanlawarchive.iuscomp.org/?p=1245
24. Bychawska-Siniarska, D. (2017). Protecting the Right to Freedom of Expression under the European Convention on Human Rights. 4.5. Retrieved from https://rm.coe.int/handbook-freedom-of-expression-eng/1680732814
25. Pallero, J. (2018). Honduras: new bill threatens to curb online speech. Access Now. Retrieved from https://www.accessnow.org/honduras-new-bill-threatens-curb-online-speech/
26. Taye, B. & Jit Singh Chima, R. (2018). Bangladesh wants to fight "fake news" during elections with internet shutdowns. We can't let that happen. Access Now. Retrieved from https://www.accessnow.org/bangladesh-wants-to-fight-fake-news- during-elections-with-internet-shutdowns-we-cant-let-that-happen/
27. Douek, E. (2017). Germany's Bold Gambit to Prevent Online Hate Crimes and Fake News Takes Effect. Lawfare. Retrieved from https://www.lawfareblog.com/germanys-bold-gambit-prevent-online-hate-crimes-and-fake-news-takes-effect
28. Kretschmer, M. & Erickson, K. (2018). How much do we know about notice-and-takedown? New study tracks YouTube removals. Kluwer Copyright Blog. Retrieved from http://copyrightblog.kluweriplaw.com/2018/06/12/much-know-notice-takedown-new-study-tracks-youtube-removals/
29. Newton, C. (2019). The Trauma Floor. The Verge. Retrieved from https://www.theverge.com/2019/2/25/18229714/ cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona
30. This is the case, for instance, in automated takedowns of political speech and marginalized voices based on copyright upload filters. See Reda, J. (2017). When filters fail: These cases show we can't trust algorithms to clean up the internet. Retrieved from https://juliareda.eu/2017/09/when-filters-fail/
31. European Court of Human Rights (2016). Cengiz and Others v. Turkey, para. 49. Retrieved from

https://hudoc.echr.coe.int/eng?i=001-159188

32. Supreme Court of the United States (2017). Packingham v. North Carolina. Retrieved from https://www.supremecourt.gov/opinions/16pdf/15-1194_08l1.pdf
33. See Roger Denson, G. (2017). Courbet's Origin Of The World Still Too Scandalous For Media-Savvy Facebook. Huffington Post. Retrieved from https://www.huffpost.com/entry/courbets-1866-the-origin_b_1087604 and Levin, S. et al. (2016). Facebook backs down from 'napalm girl' censorship and reinstates photo. The Guardian. Retrieved from https://www.theguardian. com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo
34. Human Rights Council (2018). Op.cit.
35. Wong, C. Solon, O. (2018). Facebook releases content moderation guidelines – rules long kept secret. The Guardian. Retreived from https://www.theguardian.com/technology/2018/apr/24/facebook-releases-content-moderation-guidelines-secret-rules
36. Organization of American States (2019). Press Release R116/19. The Office of the Special Rapporteur condemns closure of Radio Caracas Radio 750 AM, the censorship of television channels, restrictions on the internet, and the arrest of journalists in Venezuela. Retrieved from https://www.oas.org/en/iachr/expression/showarticle.asp?artID=1140&lID=1
37. Wu, T. (2017). The Attention Economy and the Law. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_ id=2941094
38. United Nations General Assembly (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. A/73/348. Page 11. Retrieved from https://freedex.org/wp-content/blogs.dir/2015/ files/2018/10/AI-and-FOE-GA.pdf
39. Fisher, M. Taub, A. (2019). How YouTube Radicalized Brazil. The New York Times. Citing work by Jonas Kaiser and Yasodara Córdova, with Adrian Rauchfleisch of National Taiwan University. Retrieved from https://www.nytimes. com/2019/08/11/world/americas/youtube-brazil.html Similarly, see "On YouTube's Digital Playground, an Open Gate for Pedophiles" by the same authors, retrieved from https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html
40. Goel, V. (2014). Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry. The New York Times. Citing research by Adam D. I. Kramer et. al. Retrieved from https://www.nytimes.com/2014/06/30/technology/facebook- tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html
41. Goodman, R. (2017). Facebook's ad-targeting problems prove how easy it is to discriminate online. NBC Think. Retrieved from https://www.nbcnews.com/think/opinion/facebook-s-ad-targeting-problems-prove-how-easy-it- discriminate-ncna825196
42. Matsa, K. Shearer, E. (2018). News Use Across Social Media Platforms 2018. Pew Research Center. Retrieved from https:// www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/
43. Access Now, Civil Liberties Union for Europe, European Digital Rights (2018). Informing the "Disinformation" Debate. Retrieved from https://edri.org/files/online_disinformation.pdf
44. European Commission (2016). Code of Conduct on countering illegal hate speech online. Retrieved from https://ec.europa. eu/newsroom/just/document.cfm?doc_id=42985.The E.U. Code of Conduct requires online platforms to review all submitted valid notices reporting illegal online hate speech in less than 24 hours and to remove or disable access to such content.
45. Pallero, J. (2019). Access Now on the ChristChurch Call: rights, wrongs, and what's next. Retrieved from https://www.accessnow.org/access-now-on-the-christchurch-call-rights-wrongs-and-whats-next/
46. Diaz, A. (2019). Global Internet Forum to Counter Terrorism's 'Transparency Report' Raises More Questions Than Answers. Just Security. Retrieved from https://www.justsecurity.org/66298/gifct-transparency-report-raises-more-questions-than- answers/

47. European Commission (2018). Commission Recommendation on measures to effectively tackle illegal content online. Retrieved from https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively- tackle-illegal-content-online
48. European Commission (2019). Fighting Terrorism Online: EU Internet Forum committed to an EU-wide Crisis Protocol. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/ip_19_6009
49. European Commission (2019). Fighting Terrorism Online: E.U. Internet Forum committed to an E.U.-wide Crisis Protocol. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/ip_19_6009
50. The rush to comply with political pressure and other needs associated with illegal or harmful content has led companies to adopt practices that lack transparency and proportionality. The removal of evidence of human rights violations as part of the effort to curb terrorist content is an example of this. See York, J., Al Jaloud, A., Al Kathib, H., Kayyali, D. (2019). Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content. Retrieved from https://www.eff.org/wp/ caught-net-impact-extremist-speech-regulations-human-rights-content#Blunt
51. Feld, H. (2018). Part III: Cost of Exclusion as a Proxy for Dominance in Digital Platform Regulation. Public Knowledge. Retrieved from https://www.publicknowledge.org/news-blog/blogs/part-iii-cost-of-exclusion-as-a-proxy-for-dominance- in-digital-platform-reg
52. For instance, the E-Commerce Directive stipulates in Recital 42 that activity of intermediary has to be of a mere technical, automatic, and passive nature. The intermediary can have neither knowledge of nor control over the information which is transmitted or stored. The distinction between active and passive intermediary was further clarified by the Court of Justice of the European Union (CJEU) according to which a passive intermediary has neither knowledge nor control over the information which is transmitted or stored. See joined Cases C-236/08 and C-237/08 Google France v Louis Vuitton et al (2010) ECR I-241; Case C-324/09 L'Oréal v eBay International (2011) ECR I-6011; Case C-291/13 Papasavvas (2014) EUECJ C-291/13
53. Gillespie, T. Op.cit.
54. Barzilai-Nahon, K. (2008). Towards a Theory of Network Gatekeeping: A Framework for Exploring Information Control. Journal of the American Society for Information Science and Technology, Vol. 59(9), p. 1493–1512. Laidlaw, E. (2015). A Framework for Identifying Internet Information Gatekeeper. International Review of Law, Computers & Technology, Vol. 24, No. 3, 2010. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2667902
55. Prince, M. (2019). Terminating Service for 8Chan. Cloudflare. Retrieved from https://blog.cloudflare.com/terminating- service-for-8chan
56. "That is to say, the greater the cost to individuals and firms (whether as consumers or producers or any of the other roles they may play simultaneously on online platforms), the greater the need for regulations to protect platform users from harm. If a firm is 'too big to lose access to,' then we should treat that firm as dominant." Feld, H. op. cit.
57. "Legislation applicable to internet intermediaries and to their relations with states and users must be accessible and foreseeable." Council of Europe (2018), op. cit.; "'Governments must publish all legislation, policy decisions, and other forms of regulation relevant to intermediary liability online in a timely fashion and in accessible formats." Manila principles on intermediary liability (2015). Retrieved from https://www.manilaprinciples.org/
58. United States District Court, S.D. New York (1991). Cubby, Inc. v. CompuServe Inc; Supreme Court, Nassau County, New York (1995). Stratton Oakmont, Inc. v. Prodigy Servs. Co. Retrieved from https://h2o.law.harvard.edu/cases/4540
59. United States Court of Appeals, First Circuit (2016). Jane Doe No. 1 v. Backpage.com, LLC. Retrieved from https:// caselaw.findlaw.com/us-1st-circuit/1728752.html. In this case, the First Circuit held that Section 230 protects the choices of websites as speakers and publishers,

stating the following: "Congress did not sound an uncertain trumpet when it enacted the CDA, and it chose to grant broad protections to internet publishers. Showing that a website operates through a meretricious business model is not enough to strip away those protections."

60. Court of Justice of the European Union (2012). SABAM v. Netlog. Case C-360/10. Retrieved from http:// curia.europa.eu/ juris/document/document.jsf?docid=119512&doclang=EN
61. Manila Principles (2015), op.cit.
62. Council of Europe (2018). Recommendation CM/Rec(2018)2 of the Committee of Ministers to member states on the roles and responsibilities of Internet intermediaries. Retrieved from https://rm.coe. int/1680790e14 intermediaries. Retrieved from https://rm.coe.int/1680790e14
63. Human Rights Council (2018). Op.cit.
64. Court of Justice of the European Union (2011). L'Oreal SA v. eBay International AG, C-324/09. Retrieved from http://curia. europa.eu/juris/liste.jsf?num=C-324/09
65. Court of Justice of the European Union, op. cit.
66. Court of Justice of the European Union (2016). McFadden v. Sony Music, Case C-484/14. Retrieved from http://curia. europa.eu/juris/document/document.jsf?docid=183363&doclang=EN&mode=lst&occ=first
67. Kuczerawy, A. (2019). Intermediary Liability and Freedom of Expression in the EU: Concepts and Safeguards, pp. 296-297. Available from https://intersentia.com/en/intermediary-liability-and-freedom-of-expression-in-the-eu-from-concepts-to-safeguards.html
68. Kuczerawy, A. (2019), op. cit., p. 305.
69. International Covenant on Civil and Political Rights (December 16, 1966). Retrieved from https://www. ohchr.org/en/ professionalinterest/pages/ccpr.aspx
70. European Convention on Human Rights (November 4, 1950). Retrieved from http://www.hri.org/docs/ ECHR50.html
71. European Commission. (2012). A coherent framework for building trust in the Digital Single Market for e-commerce and online services, p.13, ft. 49. Retrieved from https://eur-lex.europa.eu/legal-content/en/ TXT/?uri=CELEX:52011DC0942
72. Kuczerawy, A. (2019). Intermediary Liability and Freedom of Expression in the EU: Concepts and Safeguards, p. 38. Volume 3 of Ku Leuven Centre for It & IP Law. Intersentia.
73. Aleksandra Kuczerawy, Christina Angelopoulos, Stijn Smets, and Joris van Hoboken, among others.
74. This mechanism for copyright infringements has been adopted by Canada as the core provision in the 2012 digital copyright reform. Canadian Copyright Modernization Act (2012). Retrieved from https://laws-lois.justice.gc.ca/eng/ annualstatutes/2012_20/fulltext.html
75. Organization for Economic Cooperation and Development. (2011). The Role of Internet Intermediaries in Advancing Public Policy Objectives – Forging partnerships for advancing policy objectives for the Internet economy, p. 57. Retrieved from https://www.oecd.org/sti/ieconomy/ theroleofinternetintermediariesinadvancingpublicpolicyobjectives.htm
76. For the overview of national implementations of notice-and-notice procedure, please consult Kuczerawy, A., op.cit.
77. Angelopoulos, C., Smet, S. (2016). Notice-and-Fair-Balance: How to Reach a Compromise between Fundamental Rights in European Intermediary Liability, p. 23. Retrieved from https://papers.ssrn.com/ sol3/papers.cfm?abstract_ id=2944917
78 Manila Principles. (2015), op.cit., European Commission. (2018). Commission Recommendation on Measures to Effectively Tackle Illegal Content Online. Retrieved from https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online
79. European Commission. (2017). Communication on Tackling Illegal Content Online - Towards an enhanced responsibility of online platforms, p. 8. Retrieved from https://ec.europa.eu/digital-single-market/en/ news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms
80. Kuczerawy, A. (2019), op. cit., p. 307.

81. Council of the European Union. (2018). Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online

82.Council of Europe. (2018). Op. cit.

83. Council of Europe. (2016). Comparative study on blocking, filtering, and take-down of illegal internet content, p. 16. Retrieved from https://edoc.coe.int/en/internet/7289-pdf-comparative-study-on-blocking-filtering-and-take-down-of-illegal-internet-content-.html

84. Council of Europe. (2018). Op. cit. .

85. Wagner, B. et al. (2020). Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act. FAT* '20, January 27–30, 2020, Barcelona, Spain.

86. Ranking Digital Rights. (2017). Submission to UN Special Rapporteur for Freedom of Expression and Opinion David Kaye: Content Regulation in the Digital Age. Retrieved from https://rankingdigitalrights.org/wp-content/uploads/2018/01/RDR- 2018-David-Kaye-Submission.pdf

87. European Commission. (2017). Communication on "Tackling Illegal Content Online – Towards an enhanced responsibility of online platforms." Retreived from https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms

88. These principles are largely consistent with those recommended by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression and those contained in the Santa Clara Principles on Transparency and Accountability in Content Moderation (n.d.). Retrieved from https://santaclaraprinciples.org/

89. Mirani, L. (2015). Millions of Facebook users have no idea they're using the internet. Quartz. Retrieved from https:// qz.com/333313/milliions-of-facebook-users-have-no-idea-theyre-using-the-internet/

90. Human Rights Council. (2008). Protect, Respect and Remedy: a Framework for Business and Human Rights. Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, John Ruggie. A/HRC/8/5. Retrieved from https://www2.ohchr.org/english/bodies/ hrcouncil/docs/11session/A.HRC.11.13.pdfReport of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, John Ruggie. A/HRC/8/5. Retrieved from https://www2.ohchr.org/english/bodies/hrcouncil/docs/11session/A.HRC.11.13.pdf

91. Human Rights Council. (2009). Business and human rights: Towards operationalizing the "protect, respect and remedy" framework - Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises. A/HRC/11/13. Retrieved from https://www2.ohchr.org/english/ bodies/ hrcouncil/docs/11session/A.HRC.11.13.pdf

92. Council of Europe (2018). Op. cit.

Access Now **(https://www.accessnow.org)** defends and extends the digital rights of users at risk around the world. By combining direct technical support, comprehensive policy engagement, global advocacy, grassroots grantmaking, and convenings such as RightsCon, we fight for human rights in the digital age.