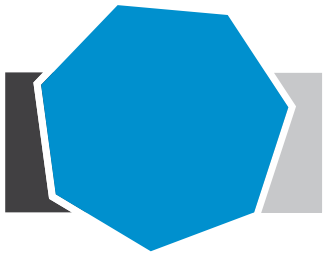


INTERNET & JURISDICTION POLICY NETWORK

CONTENT & JURISDICTION PROGRAM

OPERATIONAL APPROACHES
NORMS, CRITERIA, MECHANISMS

APRIL 2019
www.internetjurisdiction.net

FOREWORD

When the Internet & Jurisdiction Policy Network was founded in 2012, the importance of addressing jurisdictional issues online was hardly recognized by most stakeholders. The dominant view was simply that the anticipated mass increase in internet penetration would allow people around the world to better connect and share their ideas, contribute to greater freedom and create new economic opportunities. To a large extent, many aspects of this vision have materialized in the past seven years and we now take for granted the many benefits this unprecedented collective creation of mankind has brought.

In spite of that - or maybe because of it - attention in recent years has significantly shifted: hardly a day passes without major newspapers headlines about abuses online and the difficulty to address them, given the transnational nature of the network. We may rationally recognize that such abuses remain limited in proportion of the overall online activity, but the tremendous volume of the latter legitimately make the former an increasing concern for all actors. Addressing harmful content, criminal activities and other regulatory challenges in a rights-respecting and economically sustainable manner has emerged as a crucial question for the digital 21st century.

It may have been naive to think that the dark side of human nature would not express itself also in the digital space, but it behooves all of us now to avoid letting the pendulum swing too far in the other direction. We need to find collective solutions that not only protect the precious acquis of a global network but enable our digital society to develop further in a balanced manner. This can only be achieved through cooperation similar to that which enabled the emergence of the internet itself. Unfortunately, the existing international system of separate territorial sovereignties often represents an obstacle to such cooperation.

In the absence of clear international arrangements, after a long period of inaction, the last few years have witnessed a number of separate proposals and regulations to address abuses online. However well intentioned some of them may be, unilateral decisions adopted in an uncoordinated manner under the pressure of urgency may have detrimental unintended consequences. Yet, the very proliferation of initiatives demonstrates a shared concern to address these issues. This convergence in the willingness to act must be accompanied by increased communication, coordination and cooperation between actors. It is more than ever crucial to reiterate our firm belief in the necessity to tackle common problems in a collective manner.

Given the evolution in actors' mentalities, discourse and actions that we have witnessed in the past seven years, in particular in the context of the Internet & Jurisdiction Policy Network, we should be optimistic that we can develop common frameworks benefiting all stakeholders. By working intensely and in a constructive spirit in relentless pursuit of scalable, interoperable and resilient solutions, we can together address the most pressing issues of the digital society. The following *Operational Approaches* document represents an encouraging step in this direction, concretely illustrating what can be produced when actors commit to working together in pursuit of the common public interest.

Bertrand de La Chapelle

Executive Director

Secretariat of the Internet & Jurisdiction Policy Network

TOWARDS LEGAL INTEROPERABILITY

The Internet increasingly underpins political, economic and social interactions. However, as Internet penetration grows, so do cross-border legal problems. The transnational nature of the network challenges the territorial foundation of national legal systems. The number of internet users more than doubled in the last decade, and more than half the world's population is now online. How to jointly address pressing legal challenges at the intersection of the global digital economy, human rights and security has become one of the greatest challenges of the 21st century that will define the future of the cross-border internet and the digital society.

Since 2012, stakeholders from around the world work together in the Internet & Jurisdiction Policy Network to address the tension between the cross-border nature of the Internet and national jurisdictions. Its Secretariat enables multistakeholder cooperation and facilitates a global policy process engaging over 200 key entities from more than 40 countries and all stakeholder groups: governments, the world's largest internet companies, technical operators, civil society groups, academia and international organizations.

Stakeholders in the Internet & Jurisdiction Policy Network work together in currently three thematic Programs (Data & Jurisdiction, Content & Jurisdiction and Domains & Jurisdiction) to jointly develop policy standards and operational solutions through regular virtual and physical meetings, including regional sessions and Global Conferences. The Secretariat also maintains the I&J Retrospect Database tracking global trends, and launches in 2019 the world's first Internet & Jurisdiction Global Status Report.

The regular Global Conferences of the Internet & Jurisdiction Policy Network are institutionally supported by six international organizations: Council of Europe, European Commission, ICANN, OECD, United Nations ECLAC, and UNESCO. They were organized in the past in partnership with France (2016) and Canada (2018). The work of stakeholders in the Internet & Jurisdiction Policy Network has been presented to and recognized by key international processes, including the United Nations Internet Governance Forum, G7, G20 or the Paris Peace Forum, and covered in media outlets such as The Economist, New York Times, Washington Post, Financial Times, Politico or Fortune. The work of the Policy Network is financially supported by a unique coalition of over 20 governments, companies and organizations.

FROM ISSUES FRAMING TO AREAS OF COOPERATION

After four years of international consultations and meetings in the Internet & Jurisdiction Policy Network, stakeholders gathered for the first time on a global level in Paris on November 14-16, 2016 to address the future of jurisdiction on the cross-border Internet. On this occasion, over 200 senior representatives from all stakeholder groups stressed the urgency of finding mechanisms for communication, coordination and cooperation in order to establish legal interoperability and ensure due process across borders. At this 1st Global Conference, they recognized that no actor or stakeholder group can solve these new challenges on their own: collective action was needed to prevent the escalation of a legal arms race and the proliferation of legal uncertainty. On the basis of *Framing Papers*[1] for each of the three thematic I&J Programs, they accordingly identified key *Areas for Cooperation*[2] to proceed together.

FROM POLICY OPTIONS TO THE OTTAWA ROADMAP

These *Areas for Cooperation* served as mandate for the three thematic Programs Contact Groups formed as a result of the 1st Global Conference. Composed of Members from a diverse range of entities

[1] https://www.internetjurisdiction.net/news/framing-papers-released-for-data-content-and-domains

[2] https://www.internetjurisdiction.net/uploads/pdfs/GIJC-Secretariat-Summary.pdf

and experts most engaged in the issues, they were tasked to propose what can realistically and pragmatically be achieved within each of the I&J Programs. Members, with the support of the Secretariat, mapped their respective perspectives, compared approaches, fostered policy coherence, and identified possible steps for coordinated actions. The results of these focused discussions were synthesized in *Policy Options* documents[3] released for stakeholder consultations in November 2017.

They served as official input to structure discussions at the 2nd Global Conference of the Internet & Jurisdiction Policy Network in Ottawa, on February 26-28, 2018. Over 200 stakeholders from more than 40 countries decided there on concrete focus and priorities, agreeing for the first time on Common Objectives and Structuring Questions for each of the three Programs of the Policy Network. These Work Plans were consolidated in the *Ottawa Roadmap*[4].

OPERATIONAL APPROACHES

Building on the methodology of the work in the I&J Programs between the 1st and 2nd Global Conferences, over 120 Members from all continents and stakeholder groups officially begun their work in August 2018 in new Contact Groups to implement the Work Plans of the *Ottawa Roadmap*. Three neutral Coordinators were appointed to facilitate discussions. They were respectively:

- DATA & Jurisdiction: Robert Young, Legal Counsel, Global Affairs Canada.
- CONTENT & Jurisdiction: Wolfgang Schulz, Director, Humboldt Institute for Internet and Society.
- DOMAINS & Jurisdiction: Maarten Botterman, Director, GNKS Consult.

The Members of the three Programs' Contact Groups were committed to working together and develop operational policy approaches in preparation for the 3rd Global Conference of the Internet & Jurisdiction Policy Network. The mandate for the three Programs' Contact Groups was defined on the basis of the Structuring Questions of the *Ottawa Roadmap*'s Work Plans. Topic-specific Working Groups were established in each Program to conduct focused work and allow for more intense interactions on specific issues.

The *Operational Approaches* documents present the result of this process. They are a best effort by the Members of each Program's Contact Group to address the important cross-border issues pertaining to access to electronic evidence, content restrictions and moderation online, and requests for domain suspensions, in a manner consistent with due process and the protection of human rights.

THE 3rd GLOBAL CONFERENCE AND BEYOND

The 3rd Global Conference of the Internet & Jurisdiction Policy Network will be held on June 3-5, 2019, in Berlin, Germany. When they convene in Berlin stakeholders will discuss, on the basis of the *Operational Approaches*, how to advance the development of concrete policy standards and operational solutions. The *Berlin Roadmap* that will come out of this 3rd Global Conference will guide the next phase of work of stakeholders in the Programs of the Internet & Jurisdiction Policy Network, in particular:

- How proposals in the *Operational Approaches* documents (Norms, Criteria and Mechanisms) can be used to enhance legal interoperability;
- How to structure further work on issues already identified that require or warrant more in-depth discussions;
- How to address new issues identified at the 3rd Global Conference in a solutions-oriented manner;

[3] https://www.internetjurisdiction.net/news/policy-options-documents-released-for-the-2nd-global-internet-and-jurisdiction-conference

[4] https://www.internetjurisdiction.net/news/outcomes-of-the-2nd-global-conference-of-the-internet-jurisdiction-policy-network

CONTEXT

ONLINE CONTENT - THE CHALLENGES

Every day, several hundreds of millions of posts and pictures and hundreds of thousands of hours of videos are uploaded just on the major internet platforms and are available worldwide by default, by virtue of the internet's technical borderlessness. A broad diversity of private online services hosting user-generated content have thus become key instruments for the exercise of freedom of expression and public debate by billions of users thanks to permission-less and frictionless services.

However, content legal in one country can be illegal in another. In addition, as the number of Internet users grows, so does the diversity of their social, cultural, political or religious references, and thus their sensitivities towards various content items. Online services can also be misused and there is a growing awareness of the presence of illegal or harmful content online. Moreover, behaviors and opinions that were previously ephemeral and confined to the private sphere can now get significant visibility, broad geographic reach, temporal permanence, and even viral replication.

The common challenge for all actors is how to address abuses in a way that is timely and efficient, yet fully respects international human rights principles and enables the further development of the digital economy. The challenge is compounded because several jurisdictions are often involved. Particular efforts are necessary to enable the coexistence of different norms in online spaces, and ensure that content restrictions are necessary and proportionate, with appropriate due process safeguards.

No actor or category of actors can solve this conundrum on its own but we may have collectively waited too long to address these issues. As a result, public and private actors, under the pressure of urgency, now develop numerous initiatives in an uncoordinated manner, introducing significant changes in two regards.

EVOLUTION OF THE NORMATIVE LANDSCAPE

National legislations present very diverse levels of normative consistency with respect to the different types of content. Indeed, a significant consensus exists on the global unacceptability of some content (such as child sexual abuse material); yet, there is great variation regarding criteria for legitimate restrictions of many other types of content[1], including incitement to violence, hate speech, harassment, defamation, or misinformation. Legislations can legitimately reflect the specific cultural, historical, political and religious sensitivities of local communities regarding what content is acceptable or not. Yet, they sometimes might not fully respect international human rights standards and due process guarantees.

In recent years, public authorities around the world have increasingly strengthened the enforcement of their legislations regarding content online and new regulations have been developed or proposed. Ensuring compatibility of these different rules and determining the proper geographic extension of their application remain unsolved and a potential cause for tensions.

In parallel, providers have developed increasingly detailed - and frequently updated - Terms of Service and Community Guidelines setting rules applicable to their online spaces. Some of these rules are specific to the particular community the service aims to serve, but some are more generic. Given the prominent role played by the major operators in the ecosystem, the global applicability of these norms directly influences what content is deemed legitimate or not in cyberspace as a whole. Community Guidelines therefore increasingly represent an additional source that must be taken into account in this complex, hybrid normative landscape.

[1] *The Content & Jurisdiction Program does not primarily address issues related to intellectual property and copyright.*

Until recently, the key questions regarding online restrictions were: the applicability of territorially-bounded national laws on cross-border online spaces, the proper procedures for content restriction orders by public authorities, and how providers should respond to them. However, in light of the evolution described above, the mechanisms through which content is restricted by private operators in application of their own rules become an additional and important topic.

EVOLUTION OF THE ROLE OF INTERMEDIARIES

Section 230 of the 1996 Communications Decency Act (CDA) in the United States, the E-Commerce Directive in the European Union adopted in 2000, and some similar regulations in other countries[2] have historically granted broad protection to intermediaries for user-generated content, provided they acted expeditiously when notified of its illegality. However, in the context of increased awareness about abuses, some actors have called into question these intermediary liability regimes, advocating for a shift from the existing notice-and takedown frameworks towards platforms assuming a more proactive monitoring and moderation role.

As a result, public-private codes of conduct have been developed and new legislations increasingly impose more responsibilities on private providers, including short response times for certain types of content (in particular violent extremism), under the penalty of significant fines. In parallel with the growing level of detail of their Community Guidelines, major companies in response increasingly develop algorithmic tools, including some relying on artificial intelligence, for the detection of content that justifies restrictions, and prevention of its re-upload once identified. Large numbers of moderators are being hired, internal escalation paths established and recourse mechanisms envisaged, as private actors become the key decision-makers on content restrictions online.

These evolutions significantly alter the distribution of responsibilities between public and private actors, as well as the level of procedural guarantees implemented. The consequences may not yet be fully understood, given the different sizes, capacities and types of services of providers. Additional barriers to entry could hamper the emergence of new actors, preventing competition.

COOPERATION FRAMEWORKS

The different actors recognize the acute complexity of these challenges. They expressed interest in working jointly to develop procedures and standards allowing to reconcile the complementary aims of: maximizing the necessary prevention and remediation of harm, minimizing restrictions to freedom of expression and enabling the continued development of the digital economy. Clearer common guidelines and mechanisms are necessary to properly deal with abusive content: existing instruments based on strict territorial jurisdictions are challenged and some institutional innovation might be needed.

The work of the dedicated Contact Group of the Internet & Jurisdiction Policy Network, as presented in this *Operational Approaches* document, aims to contribute to this discussion by addressing the key elements of a general framework regarding responsible content moderation and restrictions.

The Internet & Jurisdiction Secretariat

[1] *The Stanford CIS "World Intermediary Liability Map" available at https://wilmap.law.stanford.edu/map lists such regulations on a national basis .*

TABLE OF CONTENTS

COORDINATOR'S MESSAGE

Debates about content and jurisdiction cannot be separated from the issue of online content moderation and restriction in general, which has a strong human rights dimension by nature. Under the international human rights framework, freedom of speech and access to information can only be restricted by states if duly justified by law and proportionate. However, under the well-established Ruggie Principles, companies also have a responsibility to respect human rights, and therefore abide by the standard of proportionality. Proportionality is also relevant for the cross-border effects of content restrictions.

Both national laws and the community standards set by companies now form a complex normative environment governing what kind of content is taken down or stays up on online platforms. States issuing content restriction requests on the basis of companies' community standards instead of their national laws is an example of potential hybridization between state and non-state governance. This raises delicate questions which can only be addressed with more legal clarity, as highlighted in the attached *Operational Approaches*.

The same is true for procedural safeguards and transparency. Beyond the question of the legal basis upon which content restrictions are made, who makes such decisions also matters. In that context, different forms of external bodies are suggested to provide avenues for recourse, be it for one platform or several, at the initiative of a particular company (e.g. Facebook), or on the basis of a national law. The Contact Group explored inter alia the key questions that the creation of such structures raises, if they are to be established in an inclusive and transparent way.

The present *Operational Approaches* document can hopefully help decision makers (e.g. public authorities, politicians, judges, company leaders) develop policies and make decisions, when necessary, in full respect of the principle of proportionality, including in what regards:

1. The geographic scope of restrictions,
2. The most appropriate action according to the type of content at stake, the corresponding potential harm and the context;
3. The use of technology (algorithmic filtering, including AI-based systems)

The proposed Operational Norms, Criteria and Mechanism will surely contribute to a global human rights-respecting policy framework pertaining to content moderation and restrictions. The discussion also highlights the need for a fundamental reflection on our current internet governance systems and their limitation. New institutional arrangements might be needed to at least ensure compatibility between very different governance regimes. In particular, debates about the respective roles and responsibilities of actors are particularly important. These challenges should be addressed and this effort hopefully points in the right direction.

Wolfgang Schulz,

Coordinator,

Content & Jurisdiction Program's Contact Group

MEMBERS OF THE CONTENT & JURISDICTION PROGRAM'S CONTACT GROUP

The Secretariat appointed a neutral Coordinator to facilitate the work of the Contact Group:

- **WOLFGANG SCHULZ**, Director, Humboldt Institute for Internet and Society

The discussions in Working Groups, which helped conduct focused work on specific topics, were moderated by neutral Facilitators:

- **HAWLEY JOHNSON**, Associate Director, Columbia Univeristy, Global Freedom of Expression Project
- **JUAN CARLOS LARA**, Content Director, Derechos Digitales
- **JASON PIELEMEIER**, Policy Director, Global Network Initiative (GNI)

MEMBERS OF THE CONTACT GROUP

CHINMAYI ARUN	Assistant Professor of Law and Executive Director, National Law University Delhi, Centre for Communication Governance,
SUSAN BENESCH	Project Director, Dangerous Speech Project
GUY BERGER	Director, Freedom of Expression and Media Development, UNESCO
ELLEN BLACKLER	Vice President Global Public Policy, The Walt Disney Company
AGNES CALLAMARD	Director, Columbia University, Global Freedom of Expression Project
MARIA PAZ CANALES	Executive Director, Derechos Digitales
MARK CARVELL	Head of International Online Policy, United Kingdom, Department for Culture Media and Sport
ALEXANDER CORBEIL	Senior Research Analyst, Canada, Department of Public Safety
JACQUES DE WERRA	Professor and Vice Rector, University of Geneva
AGUSTINA DEL CAMPO	Director, University of Palermo, Centre for Studies on Freedom of Expression (CELE)
HARLEM DESIR	Representative on Freedom of the Media, OSCE
ELENA DODONOVA	Administrator, Media and Internet Division, Council of Europe
ANRIETTE ESTERHUYSEN	Senior Advisor on Internet Governance, Association for Progressive Communications
MIRIAM ESTRIN	Public Policy Manager, Google
RAQUEL GATTO	Regional Policy Advisor, Internet Society
DANIEL HOLZNAGEL	Legal Officer, Germany, Federal Ministry of Justice and Consumer Protection
RAMAN JIT SINGH CHIMA	Asia Policy Director and Senior International Counsel, Access Now
DAVID KAYE	Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, United Nations
EDISON LANZA	Special Rapporteur for Freedom of Expression, Organization of American States (OAS)
EMMA LLANSO	Director, Free Expression Project, Center for Democracy & Technology

REBECCA MACKINNON	Director, Ranking Digital Rights, New America Foundation
KATHERINE MAHER	Executive Director, Wikimedia Foundation
GIACOMO MAZZONE	Head of Institutional Relations and Member Relations, European Broadcasting Union (EBU)
CORYNNE MCSHERRY	Legal Director, Electronic Frontier Foundation (EFF)
PAUL NEMITZ	Principal Advisor, European Commission, DG JUST
JUAN ORTIZ-FREULER	Policy Fellow, World Wide Web Foundation
ELENA PEROTTI	Executive Director, Public Affairs and Media Policy, WAN-IFRA
NICK PICKLES	Senior Strategist Public Policy, Twitter
JASON PIELEMEIER	Policy Director, Global Network Initiative (GNI)
THOMAS SCHNEIDER	Vice-Director, Switzerland, Federal Office of Communications
WOLFGANG SCHULZ	Director, Alexander von Humboldt Institute for Internet and Society
BERNARD SHEN	Assistant General Counsel, Microsoft
CARLOS AFFONSO SOUZA	Director, Institute for Technology and Society (ITS Rio)
CHRISTOPH STECK	Director, Public Policy and Internet, Telefonica
PETER STERN	Policy Manager, Stakeholder Engagement, Facebook

In addition to the Members of the Contact Group, the Secretariat wishes to thank the following actors for their engagement in discussions held in the Contact Group and its Working Groups.

ANA ANDRIJEVIC	Research and Teaching Assistant, University of Geneva
BACH AVEZDJANOV	Program Officer, Columbia University, Global Freedom of Expression Project
NIKKI BOURASSA	Program & Policy Officer, Global Network Initiative (GNI)
AMY BROUILLETTE	Senior Research and Editorial Manager, Ranking Digital Rights
JORGE CANCIO	Deputy Director of International Affairs, Switzerland, Federal Office of Communications
GIOVANNI DE GREGORIO	Legal Researcher, WAN-IFRA
JAN GERLACH	Senior Public Policy Manager, Wikimedia Foundation
TONEI GLAVINIC	Director of Operations, Dangerous Speech Project
XIANHONG HU	Program Specialist, Communication and Information, UNESCO
HAWLEY JOHNSON	Associate Director, Columbia University, Global Freedom of Expression Project
GAYATRI KHANDHADAI	Project Coordinator, Association for Progressive Communications
JUAN CARLOS LARA	Content Director, Derechos Digitales
NATHALIE MARECHAL	Senior Research Analyst, Ranking Digital Rights
DREW MITNICK	Policy Counsel, Access Now
PAULA REAL	Policy Fellow, Internet Society
AMOS TOH	Legal Advisor to the UN Special Rapporteur on Freedom of Expression
SARVJEET SINGH	Executive Director, Centre for Communication Governance, National Law University Delhi
PALOMA VILLA MATEOS	Manager, Public Policy and Internet, Telefonica
LIZ WOOLERY	Deputy Director, Free Expression Project, Center for Democracy & Technology

SYNTHESIS OF THE OPERATIONAL APPROACHES

The following *Operational Approaches* document is the result of a best effort by the Members of the Content & Jurisdiction Program's Contact Group to address the important issues identified in the *Ottawa Roadmap* of the 2nd Global Conference of the Internet & Jurisdiction Policy Network on February 26-28, 2018. The Work Plan that was refined there identified 13 important Structuring Questions to further guide interactions within the Content & Jurisdiction Program. These *Operational Approaches* are a joint contribution by some of the most engaged experts in this field to advance the ongoing debate on the complex issues of cross-border online content restrictions. **They should however not be understood as the result of a formal negotiation validated by these Members' organizations.**

On this basis, the Members of the Program's Contact Group, with the help of the Secretariat, produced the attached set of proposed Operational Norms, Criteria and Mechanism to provide a common frame of reference for the various actors. These *Operational Approaches* intend to help public and private decision-makers take into account the full range of relevant parameters when developing and implementing responsible frameworks, rules and practices to address abuses in full respect of international human rights principles.

Taking into account the limited time available to address these complex issues, the work of the Members of the Program's Contact Group was distributed among four thematic Working Groups, to propose, draft and refine elements that are documented according to the three-part structure presented on page 16.

These *Operational Approaches* will feed into the 3rd Global Conference of the Internet & Jurisdiction Policy Network on June 3-5, 2019 in Berlin, which is organized in partnership with the Government of the Federal Republic of Germany, and institutionally supported by the Council of Europe, European Commission, ICANN, OECD, United Nations ECLAC, and UNESCO.

STRUCTURE OF THE OPERATIONAL APPROACHES

The *Operational Approaches* document is organized according to the following three-part structure.

OPERATIONAL NORMS

This section identifies a set of norms that can help organize actors' behavior in their own actions and their mutual interactions. They focus on the operational level within the context of existing high-level principles.

The Content & Jurisdiction Operational Norms specifically identify elements pertaining to conceptual framework clarity, standards of proportionality, procedural guarantees and accountability.

OPERATIONAL CRITERIA

This section contains lists of elements or criteria that can be used by all categories of decision-makers when developing, evaluating, and implementing solutions. The purpose is for all actors to be able to discuss ideas, evaluate initiatives and debate proposals using common frames of reference and structuring questions.

The Content & Jurisdiction Operational Criteria address five important themes in the debate on content restriction online: **(I) Framework Clarity**, including types of content for which restriction requests are issued and the normative basis for said requests; **(II) Detection**, including the distinction between notices by third-parties, and detection by providers; **(III) Proportionate action**, including elements regarding timeliness, evaluation, geographically proportionate restrictions and a typology of actions; and **(IV) Scalability**, to call into attention the diversity of capacity of providers and countries.

OPERATIONAL MECHANISM

This third section presents a proposal for which operationalization efforts can be initiated in the period following the 3rd Global Conference of the Internet & Jurisdiction Policy Network, in Berlin.

The concept note details how to structure discussions regarding the new mechanisms for recourse after content restriction, and how to best organize the next steps during the 3rd Global Conference and in the follow-up work.

OPERATIONAL NORMS

A general approach regarding content restrictions[1] can be based upon the following elements.

FRAMEWORK CLARITY

Definitions - A shared vocabulary regarding the different types of illegal or harmful content and restricting actions informs the development and implementation of legal regimes and companies' practices.

Normative basis - Unambiguous and comprehensible wording of national laws and private Community Guidelines ensures normative predictability for everyone.

Responsibilities - The respective rights and responsibilities of public and private actors are clearly determined, taking into account, as appropriate, the nature and size of the services.

PROPORTIONALITY

Rights - Restriction decisions take into account and aim to reconcile, or at least balance, the potentially competing rights of all relevant actors.

Granularity - Restrictions are applied to the smallest content item possible that allows to effectively address the issue.

Geographically proportionate restrictions - Decisions by public authorities and private actors preserve the broadest availability of legitimate content.

Choice of action - A diversity of graduated technical solutions provides alternatives to content removal to ensure optimal respect of proportionality.

PROCEDURAL GUARANTEES

Formats - Requests for content restrictions provide sufficient supporting information for decision-making, according to clear submission formats.

Awareness - Users have access to information about content inaccessibility and the rationale thereof.

Flagging - Easy-to-use channels are available for users to flag content they believe violates the service's community standards.

Detection - A careful combination of automated detection and human review enables timely action while fully taking context into account to reduce the risks of over-restriction.

Notification[2] - Users are notified ahead of the enforcement of restriction decisions regarding their content. If justifiably demonstrable according to clear pre-agreed criteria that advance notification is not practical, advisable, or permissible, users are notified expeditiously after the enforcement of a restriction decision. Some situations may justify an exception to the general principle of user notification.

Emergency - Specific provisions establish conditions applicable in justifiable situations of emergency.

Recourse/remediation - Accessible, speedy, clearly documented and publicly available appeal mechanisms are available with content staying up whenever possible during the appeal.

---

[1] *"Content restrictions" cover actions by providers following requests pertaining to applicable national laws, and moderation on the basis of Community Guidelines.*

[2] *The question of the specific issues pertaining to media, and procedures and principles for notification to media were raised, and deserve a dedicated discussion.*

ACCOUNTABILITY

Decision-making chain - The provider's criteria, procedures and escalation paths pertaining to content restriction are sufficiently documented and available to the public.

Consistency - Providers apply consistent criteria when implementing their Community Guidelines and addressing legal requests, dedicating appropriate resources for that purpose.

Transparency - Detailed regular reporting in accessible and exportable formats from both public authorities and private actors provides legitimacy and accountability to content restriction mechanisms and decisions.

Oversight - Ongoing monitoring enables appropriate oversight of content restrictions to increase trust in due process and accountability.

OPERATIONAL CRITERIA

The following criteria represent the best efforts by the members of the Content & Jurisdiction Program's Contact Group and its Working Groups, as compiled by the I&J Secretariat, in identifying concise lists of elements that can be used by all categories of decision-makers when developing, evaluating, and implementing solutions. The purpose is for all actors to be able to discuss ideas, evaluate initiatives and debate proposals using common frames of reference and structuring questions.

The following documents should be understood as basis for future reference and work in the Internet & Jurisdiction Policy Network, following its 3rd Global Conference. Below is the list of Operational Criteria for the Content & Jurisdiction Program:

PART I - FRAMEWORK CLARITY

- CRITERIA A - Content Typology
- CRITERIA B - Normative Basis

PART II - DETECTION

- CRITERIA C - Third-Party Notices
- CRITERIA D - Provider Detection

PART III - PROPORTIONATE ACTION

- CRITERIA E - Timeliness
- CRITERIA F - Evaluation
- CRITERIA G - Geographic Proportionality
- CRITERIA H - Choice of Action

PART IV - NOTIFICATION / RECOURSE

- CRITERIA I - User Notification
- CRITERIA J - Recourse

PART V - SCALABILITY

- CRITERIA K - Capacity of Small Providers / Countries

PART I - FRAMEWORK CLARITY

CRITERIA A - CONTENT TYPOLOGY

A broad diversity of types of content can potentially be illegal in certain countries or represent a risk of harm to users. The ease of public access and viral propagation of expression that was previously kept in private also raise new challenges. In a context of lack of sufficiently clear and agreed international definitions, the table below is a non-exhaustive attempt (that might be subject to further refinement) at describing the main issues at stake, to help all actors develop diversified and nuanced approaches to each type of challenge, in the respect of international human rights. It should *not* be understood as a normative index of content that *should* be restricted.

TYPES OF CONTENT	DESCRIPTION
RIGHTS OF THE CHILD: ART. 24 ICCPR	
Article 24 states that every child shall have, without any discrimination as to race, color, sex, language, religion, national or social origin, property or birth, the right to such measures of protection as are required by his status as a minor, on the part of his family, society and the State. The Declaration of the Rights of the Child states, "the child, by reason of his physical and mental immaturity, needs special safeguards and care..."The Convention on the Rights of the Child defines children as individuals under the age of 18 and Art. 17 requires States Parties to "[e]ncourage the development of appropriate guidelines for the protection of the child from information and material injurious to his or her well-being, bearing in mind the provisions of articles 13 and 18." The Universal Declaration of Human Rights proclaims that childhood is entitled to special care and assistance.	
Child Abuse Material or anything objectionable involving minors	Content which includes sexual or sexually suggestive content involving minors, child abuse imagery or other content posted with the intent to do harm and take advantage of their youth. This may include image privacy rights for children under 13 and up to 18 depending on jurisdiction and context.
Grooming or predation	Online grooming is when a person uses social media to deliberately cultivate an emotional connection with a child in order to sexually abuse or exploit that child.
RIGHT TO PRIVACY: ART. 17 ICCPR	
Article 17 protects the right to respect of privacy, family, home and correspondence, and the protection of honor and reputation. It states that "[n]o one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks." Any restrictions need to be lawful, necessary and proportionate. See Also CCPR/C/GC/16[1].	
Breaches of personal information	Personally Identifiable Information (PII), Sensitive Personal Information (SPI) or confidential information which is disclosed without the person's consent. Definitions vary among jurisdictions but generally include any information which reveals a person's identity. This can also include content that facilitates identity theft by posting or soliciting personally identifiable information, sharing PII via external link, sharing private financial info of business, self, or others, and sharing private contact info. "Phishing" for instance is the use of fake email, text messages, or copycat websites to steal PII.

Defamatory personal content	Content which can cause injury to dignity, reputation or personality rights. Platforms generally do not restrict this type of content unless it crosses the threshold to hate speech or incitement. The platforms surveyed restrict allegedly defamatory content only when it becomes coordinated with an intent to harm. The rights of individuals to be protected from "unlawful attacks" on his/her "honor and reputation" must be balanced with the rights of speakers to hold opinions without interference and right to access information. However, political speech or opinion, criticisms of public officials acting in their official capacity or speech which is in the public interest, even if it is considered defamatory, will receive the highest level of protection under international law. Depending upon context and intent, defamatory content may be subject to legitimate restriction under art. 19(a) for respect of the rights or reputations of others.
Coordinated/organized attempts at defamation	*Defamatory content becomes coordinated when an individual or organized group spreads the impugned content simultaneously on various platforms.*
Defamatory autocomplete suggestions or search results pointing to defamatory content ("Google Bombing," "Googlewashing")	*These terms refer to a practice of artificially elevating a particular website in search results by linking it to a search term which may be derogatory or defamatory. Motivations may be personal, political, or just as a prank.*
"Right to be Forgotten"	*These claims surround requests to delist information that is no longer valid due to passage of time and changing of circumstances, and the continued accessibility of the information constitutes a violation of reputational rights.*
Impersonation (fake accounts/profiles/pages)	Copying a user's layout, using a similar username, or posing as another person in profiles, pages, comments, emails, or videos. This is often done with the intent to harm an individual or mislead viewers. It also includes bot or other applications for propaganda. This includes impersonation of a channel (i.e. YouTube) and impersonation of an individual (i.e. Facebook or Twitter) or of a company/corporation. Impersonation for satirical or artistic purposes would be protected.
"Deep Fakes"	*This has been described by the UK Government as "audio and videos that look and sound like a real person, saying something that that person has never said." Depending upon the context and intent, as well as the jurisdiction (i.e. US First Amendment), this content may be protected.*
Sexual objectification	Content that objectifies their targets, including through manipulated photographs and sexually explicit descriptions of their bodies. Photographs are often used without their consent and manipulated so that they appear in pornographic scenes or used in memes.
Unauthorized Dissemination of Intimate Images ("Revenge porn")	Distribution of sexually graphic images without the consent of the subject of the images. The abuser obtains images or videos in the course of a prior relationship, or hacks into the victim's computer, social media accounts or phone. This is usually done with the intent to harass, humiliate injure the person.

RIGHT TO FREEDOM OF EXPRESSION: ART. 19 ICCPR	
Art. 19 (1): Right to hold opinions without interference Art.19(2): Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice. General Comment 34 (CCPR/C/GC/34[2]) also highlights how States parties have to be proactive in putting in place "effective measures to protect against attacks aimed at silencing those exercising their right to freedom of expression" (para 23).	
Disinformation	Disinformation includes the distribution of false or inaccurate information (i.e. "fake news") for political, ideological or economic gain, either by individuals or bots. Extreme examples aim to influence elections and disrupt democratic processes. It can take the form of "news" articles based partially on fact and misinformation, tweeting, posting or commenting. This type of content often does not cross legal boundaries and must be carefully distinguished from opinion and satire which are protected forms of expression.
Medical misinformation	*Content which disseminates false or misleading information that could have detrimental effects on individual or public health and safety. Examples include promotion of false cures for illnesses and anti-vaccination advice. This may be done for financial gain or simply out of ignorance rather than with an intent to harm. This type of content has been subject to restriction where it has been deemed a threat to public health.*
Sexually explicit content; nudity or porn	Images of explicit sexual activity or fetishes, and nude or partially nude people in sexually suggestive poses. This type of content may be protected for adults depending on the jurisdiction or Community Guidelines but restricted for children or youth. Artistic, scientific, documentary or educational nudity is protected. Images which identify an individual and are posted without consent, may be subject to restriction.
Content critical of religion (Blasphemy/Apostasy)	Content critical of religion, including opinions or artistic works (i.e. satirical cartoons) is protected under ICCPR Art. 19 but it must be balanced with legitimate restrictions under art. 18 (3) (Freedom to manifest one's religion or beliefs may be subject only to such limitations as are prescribed by law and are necessary to protect public safety, order, health, or morals or the fundamental rights and freedoms of others) and art. 20 if it advocates religious hatred that constitutes incitement to discrimination, hostility or violence. Some States restrict this type of content under blasphemy/apostasy laws. Blasphemy laws, which restrict speech deemed offensive to prophets or religious leaders, are often archaic, overbroad and abused. In particular, they are used to punish religious minorities or women (i.e. honor killings) in conservative religious countries. Despite international efforts to repeal such laws they remain on the books in dozens of states, and some 13 States still hand down death sentences for the offense.

Art. 19(a): Legitimate Restriction for respect of the rights or reputations of others. General Comment 34 (CCPR/C/GC/34 para 35)[2] highlights that "When a State party invokes a legitimate ground for restriction of freedom of expression, it must demonstrate in specific and individualized fashion the precise nature of the threat, and the necessity and proportionality of the specific action taken, in particular by establishing a direct and immediate connection between the expression and the threat".	
Defamatory personal content	See Defamatory personal content under Right to Privacy (p. 21). General Comment 34[2] calls on State parties to "consider the decriminalization of defamation" and recommends that, in any case of defamation, "imprisonment is never an appropriate penalty" (para 47).
Bullying	Content that purposefully targets private individuals with the intention of degrading or shaming them. An extreme form of this is called "flaming." Does not apply to public figures who are expected to tolerate higher levels or criticism within reason, in so far as the content does not include hate speech or credible threats.
Harassment	Content which is disseminated on multiple occasions to cause an individual stress, humiliation, anxiety or fear of violence. Content may contain targeted swearing, grossly offensive comments, or threats of physical harm or even of death. Depending on intent and context, only speech which is considered hate speech or a credible threat may be restricted. Many states have harassment statutes which are applicable for online violations.
Coordinated/organized Harm	*Deliberately sabotaging or invading multiple online spaces for the purposes of harassing a target. Users are currently unable to report this scope and context of the harassment, as each platform will only consider the harassment happening on their own sites.*
Cyberstalking	*No legal definition but examples include repeated threatening or obscene emails or text messages, spamming, 'flaming' (targeted online verbal abuse), "Baiting" through taunts, or sending menacing unsolicited messages.*
Deadnaming	*Using/disclosing a transgender person's birth name to harass them, invalidate their identity, and/or inflict emotional stress.*
Doxing	*Searching for and publishing private or identifying information about a particular individual, often by hacking and with malicious intent. "Dox" is a slang version of "documents." Causing fear, stress and panic is the objective of doxing, even when perpetrators think or say it is "harmless."*
Art. 19 (3)(b): Legitimate restriction for the protection of national security or of public order (*ordre public*), or of public health or morals, if it conforms to the strict tests of necessity and proportionality. See CCPR/C/GC/34[2]	
Violent/graphic content	Content that is that is sensational or gratuitously violent or glorifies violence. This must be distinguished from graphic content which may have educational, scientific or public interest purposes, such as information on historical or current events, which would be protected under international human rights law. Depending on the type of content, it may require age verification or be age restricted.
Promoting or publicizing crime	Content which incites or abets criminal activity and is believed to be a credible threat to personal or public safety or property.

Content to organize violence or support violent organizations	Content that makes credible threats of serious physical harm (organized violence, murder, human trafficking) against a specific individual or defined group of individuals or expresses support or praise for groups, leaders or individuals involved in these activities.
Sexual violence and exploitation	Content that depicts, threatens or promotes sexual violence, assault or exploitation.
Abetting self-harm or suicide	Content that promotes harmful behavior, or anything that encourages or suggests self-harm, like mutilation, eating disorders or drug abuse. Content that identifies and negatively targets victims or survivors of self-injury or suicide.
Leaked confidential or secret information	Some states may justify restricting leaked sensitive information on national security grounds. However, information in the public interest which is leaked by "whistleblowers" may receive protection under the UN Convention Against Corruption. Art. 33 recommends states provide protections to "reporting persons" who report "in good faith and on reasonable grounds to the competent authorities any facts concerning offenses" covered by the convention. Art. 32 provides protection of witnesses, experts, and victims.
Lèse-majesté or comments critical of historical personages	A few jurisdictions restrict speech considered offensive to the monarchy or historical personages on the grounds that it is either treasonous or a threat to public order. However, open and critical discussion of political leaders and public figures is protected under international law. Only credible threats to incitement may warrant restrictions.
PROHIBITION OF PROPAGANDA FOR WAR AND INCITING NATIONAL, RACIAL OR RELIGIOUS HATRED: ART. 20	
Art. 20(2) Advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law	
Hate speech	Hate speech includes serious attacks on people based on their race, ethnicity, national origin, caste, religion, gender identity, sexual orientation, disability, veteran status or medical condition. May also include targeting people based on age, weight, immigration or veteran status. Examples are violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. This also may include imagery, such as of lynching, or coordinated conduct to discriminate or dehumanize. Live-streaming or posting of archived videos of live events which amplify or incite hate crimes may require immediate restrictions (i.e. Christchurch Massacre). ICCPR articles 18, 19, 20 and 26. ICERD art. 4. Although there is no internationally agreed upon definition of or threshold for hate speech, states have an obligation under the ICCPR to prohibit advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. It is tied to a guarantee of equal protection before and of the law. Evaluating hate speech can be very subjective and therefore great care must be taken as overly broad restrictions can infringe on other fundamental rights. See also CCPR/C/GC/11[3] and Rabat Plan of Action.

Violent extremist content	Most platforms prohibit content which they define as "extremist". Definitions include content which "incites" violence, "celebrates" terrorist acts, "instructs," solicits or advocates persons or a group of person to participate in the activities of a terrorist group or provides instruction on the making or use of weapons. Some platforms deny accounts to organizations labelled as terrorist or which engage in premeditated violent activities against persons or property as an act of intimidation with a political, religious or ideological purpose. This includes terrorist hoaxes. ICCPR Art. 20 prohibits any propaganda for war. However, any restrictions should expressly exclude content which is disseminated for "educational, journalistic, artistic or research purposes or awareness raising activities against terrorism." Content flagged as "terrorist" may in some instances constitute documentation of war crimes or atrocities, and therefore should not be deleted but shared with the appropriate law enforcement authorities. Although art. 19 (3)(b) provides a legitimate restriction for the protection of national security, there is no agreed upon definition for terrorist content under international law and to that end definitions for terrorist content should be "clear, foreseeable and narrow to prevent unlawful interferences with fundamental rights." See also Art. 6 ICCPR (Right to Life); International Instruments (UN)[4]
INTELLECTUAL PROPERTY RIGHTS	
Copyright	Copyright is a legal right that protects original creative works such as music, movies, works of art or books. It does not protect facts or ideas.
Trademark	A trademark is a word, phrase, symbol, and/or design (i.e. logos, brand names) that identifies and distinguishes the source of the product/services of one party from those of others. It does not expire as some copyrights do.
REGULATED GOODS AND SERVICES	
Content pertaining to regulated or illegal goods and services will vary depending upon the jurisdiction.	
Regulated goods and services	States as well as platforms restrict content which facilitates the purchasing, trading or selling of illegal drugs, illegal services (online gambling, counterfeit documents), stolen goods, firearms or other weapons, based on the laws in the jurisdiction. Some states or platforms may also restrict content which promotes the use of illegal drugs or weapons or provides instructions for manufacturing weapons (i.e. 3D printing).
Sexual solicitation	Content disseminated with the intent to engage in sexual activity for a fee or the functional equivalent of a fee.

FRAUD	
Fraudulent content seeks to deliberately deceive individuals for unlawful gain or to deprive them of their rights. Most jurisdictions cover fraudulent activities in their civil or criminal statutes.	
Misleading metadata (title, description, tags, annotations, and thumbnail)	Includes titles, descriptions, tags, annotations, and thumbnails being used to game or trick the search algorithms for online video, rather than being representative of the actual content in the video.
Blackmail/extortion	Content or messages which threaten to reveal embarrassing information or photos of the victim (which may have been collected illegally or with the consent of the victim) unless the victim provides favors, property or money. Sextortion is a form of blackmail where sexual information or images are used to extort sexual favors, money or other demands from the victim.
Scams: trick others for their own financial gain	Content that deliberately tries to mislead users for financial gain or to access PII. Examples are purchasing views, deceptive layouts, artificial subscriptions, serving pop-unders and re-directs, vote manipulation (manipulating content votes up/down), or "spoofing" brand name logos.
Spam	Content which includes targeted, unwanted, or repetitive content in videos, comments, private messages, or off-domain redirects. It may have the intent to artificially boost views or drive down scores and other metrics through coordinated campaigns (i.e. troll teams). It can also include misleading content or behavior, like deceptive design elements or suspicious pop-ups.
Artificial traffic spam: artificially incentivize viewers for engagement	*These are targeted messages to incentivize views, called "view count gaming," which tries to make a non-view into a view for financial gain.*
Fraudulent accounts	Accounts which are run by bots rather than humans with the intent to spread misinformation and distort debate. There is some overlap with "Disinformation" and "Impersonation".

[1] *CCPR/C/GC/16: https://www.refworld.org/docid/453883f922.html*

[2] *CCPR/C/GC/11: https://www.ohchr.org/Documents/Issues/Opinion/CCPRGeneralCommentNo11.pdf*

[3] *CCPR/C/GC/34: https://bangkok.ohchr.org/programme/documents/general-comment-34.aspx*

[4] *Art 6 - International Legal Instruments: https://www.un.org/counterterrorism/ctitf/en/international-legal-instruments*

CRITERIA B - NORMATIVE BASIS

A major challenge in the drafting and implementation of domestic laws, as well as companies' Community Guidelines[1], is the need to reconcile competing rights, namely freedom of expression and the prevention of harm.

1. **Reconciling diverse normative bases**
   A plurality of sources form an increasing elaborate normative landscape, combining:
   a. Overarching international or regional human rights principles, in particular:
      i. The Universal Declaration of Human Rights (UDHR);

[1] *It was highlighted that the core and foundational purpose of companies' ToS and Community Guidelines is to maintain the nature and benefits of the service for the user community.*

ii. The International Covenant for Civil and Political Rights (ICCPR), in particular articles 6, 17, 18, 19, 20, 24, 26;

iii. The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)

iv. The Declaration of the Rights of the Child

b. The diversity of applicable national and regional laws, either existing[2] or newly drafted for the digital context because of growing concerns regarding abusive online content.

c. The increasing importance of companies' Terms of Service (ToS) and Community Guidelines.

2. International normative consistency

The following categories can be used to ascertain the degree of international normative convergence on the various issues:

a. There IS universal agreement that the content/behavior is illegal AND there is strong substantive convergence around the world on the corresponding threshold criteria (example: child sexual abuse material);

b. There IS universal agreement that the content/behavior is illegal, BUT significant national variations exist in the criteria determining illegality (example: defamation);

c. The content/behavior IS NOT universally considered as illegal, BUT the application of specific domestic laws on the local territory is considered acceptable by other countries, in particular for historic reasons (example: criminalization of Holocaust denial);

d. The content/behavior IS NOT universally considered as objectionable AND some countries even consider that it should not be allowed to make it illegal (example: laws discriminating against or criminalizing certain sexual orientation).

The frontiers between these different categories are however not rigid. Debates exist regarding where some topics fall.

3. Types of regulation

Some providers aim for Community Guidelines as uniform as possible for all their users. This produces a de facto global harmonization of applicable rules on their respective spaces. Others however rely on community-driven moderation, for instance organized by topic (e.g. Reddit) or by language (e.g. Wikipedia).

Private providers have sometimes initiated rules on new issues, e.g. on non-consensual posting of adult content (aka "revenge porn"), with some legislative initiatives emerging as a result.

A conceptual framework could be envisaged to distinguish more clearly between:

a. Regulation/moderation ON a platform: under the responsibility of the administrators of sub-forums and groups, such group rules can be more restrictive than the rules of the overall platform.

b. Regulation BY a platform: Community Guidelines and decision-making rules establish the general framework for all content on the corresponding space, including potentially the latitude given to group administrators.

c. Regulation OF platforms: domestic laws, regional legislation or international agreements defining the general responsibilities of providers in terms of content moderation, including how to reconcile their business capacity to determine their ToS and the duties that could result from an extensive market position.

[2] *This may include relevant laws related to media as applicable.*

PART II - DETECTION

CRITERIA C - THIRD-PARTY NOTICES

1. **Public authorities issue:**
   a. Formal orders on the basis of domestic laws. However, due to constraints of volume and timely action, this is sometimes done without validation by a local court that could clearly establish the illegality of the content with all appropriate procedural guarantees. This puts on private entities the responsibility to make this determination, with incentives to over-restrict content in situations of uncertainty. Additional procedures (with appropriate protections) for expedited domestic evaluation could be developed.
   b. More informal requests on the basis of ToS or Community Guidelines, for instance through so-called internet referral units. Clearer procedures are necessary to ensure transparency and accountability regarding the use of such a channel by public authorities.
2. **Private notices can come from:**
   a. Specialized notifiers, for instance for copyright or child abuse material. More clarity is however needed regarding, inter alia, their procedures, decision-making criteria, due diligence requirements and avenues for recourse if their notices are to be taken prima facie.
   b. Media, inter alia for fact-checking (e.g. during electoral periods), or reporting cyber harassment against journalists targeted because of their professional activities.
   c. Individual flaggers (including "trusted flaggers") via platform tools. The role of user reports should be seen alongside the increasing role of automated detection means for means for identifying and removing violating content. Easy-to-use flagging tools remain nonetheless necessary.

CRITERIA D - PROVIDER DETECTION

Responding to pressure, major providers increasingly implement proactive detection and this can only be done through intensive use of algorithmic tools, including Artificial Intelligence. The use of hash databases to prevent re-upload of content previously detected as justifying restriction already detected illegal content is also spreading.

The performance of such tools however strongly varies according to the different types of problematic content: performing well for images with easily recognizable elements, it remains much less accurate for anything requiring a strong evaluation of context.

In spite of significant progress, major challenges must be addressed, as automated tools:

1. Still lack the required accuracy to detect all infringing content or correctly identify objectionable content, risking under- or over-restriction.
2. Largely ignore contextual considerations, including external context, culture, and intention.
3. Risk making decisions without a proper balance between competing interests, by ignoring legal rules, legal interpretations, nuances in platform ToS, etc.
4. Raise serious transparency issues: automated removal or restriction may provide insufficient information about its rationale, making it difficult to either understand the restriction decision or challenge it.
5. May still be circumvented through technical means, for example, by changing metadata or

encrypting content, thus allowing certain harmful content to stay online or be viralized further through encrypted channels.

6. Can exhibit undetected biases due to the datasets they were trained on.

Human review remains a necessity in the decision-making process for individual restrictions and a significant collaborative effort is needed more generally to allow proper evaluation and oversight of algorithmic tools.

The following Typology of Detection Modalities is a list of recognized actions available to service providers and network / hosting intermediaries to identify allegedly harmful or illegal content. This list was developed to illustrate and map a spectrum of possible responses and is not meant to endorse any specific actions.

ACTIONS	DESCRIPTION/TECHNICAL TOOLS
CONTENT PROVIDER/PLATFORM AND SEARCH ENGINES	
Account Authentication/Verification	Authentication is to ensure that the person is who s/he claims to be and to verify the identity data. It may include confirming email address, date account was established, whether the profile is complete, etc. Authentication relates more to an internal process where the verification is about external data.
Content Monitoring	Content monitoring involves the process of implementing procedures and filters to identify content or online behaviors which may be violative of ToS, Community Guidelines or local laws. Some monitoring is conducted by humans but much of it is done automatically through algorithms or AI. This may include evaluating behavior of users (e.g. who they follow or what they share), how other accounts interact with them (e.g. who mutes, follows, shares, or blocks actor), or if there are coordinated actions taken by groups or across platforms with the intent to harm. Once potentially harmful content/behavior is identified, it may be flagged for review.
Hashing (and hash databases)	This technology creates a unique digital signature (or "hash") of an image or video, which can then be compared against hashes of other photos or videos. This can help detect and remove or prevent the upload of a new image or video if its hash matches the hash stored in a database of items previously identified as justifying restrictions. Hash databases have been used for instance regarding child sexual abuse material, violent extremism, unauthorized dissemination of intimate images ("revenge porn") or copyright.
Notice and Take Down: Temporary or Permanent Removal	Mechanism where an individual can issue a legal request to a content host that requires the host to take down, delete or restrict access to allegedly harmful content. Examples include "Right to be Forgotten" policies in the EU and the copyright-oriented notice and takedown regime of the United States Digital Millennium Copyright Act.
NETWORK/ HOSTING INTERMEDIARIES	
Notice and Take Down: Temporary or Permanent Removal	Hosting intermediaries may take content offline, based on company policies and procedures, court orders, or state regulations.

PART III - PROPORTIONATE ACTION

CRITERIA E - TIMELINESS

Recent legislative efforts have put an increasing emphasis on short response times[1] for removing specific types of content, in particular regarding terrorism and violent extremism.

1. **Tensions**
   a. Response time can be measured by reference to different operational events, including: time of upload, notification to the platform or notification to the user. Clarity regarding this factor is necessary in any rule establishing compulsory response time.
   b. The tension between short response times, accuracy of the measure and the need to ensure the protection of rights can be summarized in the following statements:
      i. Some decisions on content restrictions have to be made quickly to prevent harm;
      ii. The faster the decision, the greater the risk of errors or inaccuracies in restriction decisions, or its impact on users' rights;
      iii. Ensuring accuracy of the decision and the full respect of users' rights and interests requires careful evaluation and thus time.
2. **Rationale supporting quick decisions**
   There are a number of reasons incentivizing quick decisions:
   a. Normative incentives, including the respect of national laws, the limitation of service provider liability, and the prospect of fines.
   b. Economic incentives, including the volume of content and requests to be handled (as a matter of resources), the satisfaction of users and other interested parties (such as advertisers).
   c. Operational considerations, including the nature and volume of content and restriction requests, the distinction between clear cut and harder cases, and the consideration of harm from restriction versus harm from keeping the content accessible.
   d. Interest considerations, including for whom the rationale for restrictions are more pressing: respect of national laws, or national security, may be more pressing as an interest for governments than for users.
3. **Potential risks of quick decisions**
   a. Incorrect decision-making:
      i. False positives, leading to over-restriction. These include wrongly identified infringing content based on competing values (e.g. nudity as art v. norms against nudity), contextual analysis (i.e. external situations where there is no offense from certain content, e.g. breastfeeding), analytical errors (i.e., content was misidentified).
      ii. False negatives, leading to content unduly remaining accessible, thus materializing harm, as the counterpart to false positives.
   b. Procedural fairness:
      i. Limited capacity of users to challenge a decision before the restriction takes place. Ex ante capacity to contest is time intensive. These risks differ with regard to the type of content and related harms.
      ii. Limited transparency to challenge restriction decisions before or after they take

place, when automated detection systems are used, and there is a lack of information on the process for detection or the cause of restriction.

c. Substantive harm: risks according to the harm that too rapid decisions could produce, for example:
   i. On fundamental rights, such as freedom of expression or privacy.
   ii. On users' interests in the correct functioning of the platform, such as collaboration and discussion.
   iii. On larger public interests, such as democracy and public debate.

d. Different types of content (e.g. child abuse material vs. political content) have different risks of harm from false positives or false negatives in restriction decisions.

e. Downstream consequences of wrong decisions must also be taken into account as possible harms. A wrong decision of restriction can impact the user that generated the content but can also impact the audiences, the information environments and political decisions. The pressure to act quickly can disproportionately impact particular user groups based on language or type of content.

f. The need for quick decisions, even when justified, may have other negative effects with regard to people reviewing content, their preparation for decisions on restrictions, and their protection from harm produced by exposure to harmful content.

4. **Criteria affecting how quickly a decision can be made** include:
   a. Whether it is a clear-cut case or a hard/disputed issue.
   b. Whether restriction is pursuant to international human rights law, domestic local law, or ToS and Community Guidelines, and whether there are some potential conflicts between these norms.
   c. What the type of content is, both in terms of format (text, picture, video) and subject matter.
   d. What the type and amount of harm would be in case of restriction or not, related to the impact the delay may have.
   e. What the different effects may be in relation to users at large and those allegedly affected by certain content.
   f. What action will be implemented, among the diversity of measures that could be used to restrict access to material. For instance, content removal restricts access to all users, while placing content behind a login system restricts it from users who have not registered for a particular website. These measures and others have varying impacts on accuracy and effects on user rights and defining the least restrictive one in difficult cases takes additional time.

CRITERIA F - EVALUATION

The information available in notices needs to be sufficient for decision-makers to understand inter alia what prohibition is being referred to, what specific content is allegedly violating it and whether the content does violate the prohibition. When assessment is made that the content violates the prohibition, the action implemented needs to respect the standard of proportionality. Ensuring proportionate action on individual items of content requires evaluation of a diversity of factors and a broader appreciation of the potential impact of the measure.

1. **Multi-factor evaluation**

   a. **What is the context of the content at issue?**

   Content posted online is, by default, globally available. Nevertheless, the user making the content available and those accessing it perceive it within specific contexts (history, references, orientation, linguistic community, etc.).

   In order to take this fundamental tension into account, decision-makers can identify where and from whom the specific piece of content originates. A larger discussion on the methods for and difficulties in identifying origination would be useful to fully understand the challenges behind the identification of context. This issue is compounded when taking into account situations where the content itself is "mirrored" across multiple different websites/platforms.

   In addition, to fully understand context, decision-makers can first try to determine where the content is hosted and displayed. In other words, it is crucial to unpack the potential differences between where the website's domain is registered, the website/platform owner's country of incorporation, where the content is hosted/hashed, and where it is available.

   Finally, evaluation of the context needs to be conducted by people with capacity to understand the language and corresponding cultural environment.

   b. **What are the motives of those who have posted/re-posted this content?**

   The motive of the users who have posted or re-posted the content is important to consider. Decision-makers should keep in mind that there can be more than one motive, including the following: economic, political, humor, satire, social commentary. Those motives need to be evaluated within their linguistic and cultural environment. Where motive can be (or has been) ascertained, this may help decision-makers as they think through the various options available for restriction.

   c. **What motives might other actors have in "receiving"/having access to this content?**

   Decision-makers can consider what motives other actors can have in "receiving" or having access to this content. These can align, or be independent from, the intent of the user(s) posting it. It is also important to address the risk associated with the content, including the imminence of danger associated with it.

   d. **Are there particular jurisdictions/actors that may have an interest in and/or be impacted by this decision, and if so what do their laws/rules say about this kind of content?**

   Other jurisdictions / actors may have an interest in the decision, or be impacted by it. If the decision-makers identify that this is a possibility, it is important to consider what these interests may be, and in particular which of their specific laws or rules could apply. The above points pertaining to context and motives of users posting and receiving content may inform the identification of other relevant jurisdictions whose interests can be considered in a potential comity or conflict-of-law analysis.

   e. **Are prohibitions of this kind of content universal/widely shared/inconsistent across jurisdictions?**

   The decision-maker can determine the level of international normative consistency, understood as a basic assessment of the degree of global consensus on the unacceptability / illegality of such content. As a general matter, it may be useful to consider whether the content fits into one of the four categories described above in Normative Basis.

f. What is the format of the content at issue? How does the format of the content at issue impact its potential virality?

The format of the content at hand (e.g. text, image, video, hyperlink, etc.) can often determine the virality of the content and is an important criterion to analyze with respect to the ability of the content to be shared across pages, platforms and devices. In addition, the format of the content is an important (but not unique) characteristic in determining the file size and its implications in of accessibility and storage.

2. Impact analysis

Evaluation includes the consideration of a range of potential impact, including (but not limited to):

a. Impacts on freedom of expression

Laws that restrict freedom of expression must meet the legality, legitimacy, and necessity tests drawn from Articles 19 of the Universal Declaration of Human Rights (UDHR) and the International Covenant on Civil and Political Rights (ICCPR). Laws on expression nevertheless vary across jurisdictions, and as a result decisions about how to restrict digital content must pay particular attention to potential conflicts of laws.

In particular, it is important for decision-makers to discern, where possible, what other jurisdictions may have connections to the content at issue. This can include for instance the identification of:

i. where the user(s) responsible for the content is situated;
ii. where the platform hosting the content is headquartered; and/or
iii. where significant audiences for the content are located.

The stronger the identifiable connections to countries whose laws could be read to protect the content at issue, the more cautious decision-makers should be before ordering restrictions that could have impacts in those jurisdictions.

b. Impacts on privacy

Enforcement of laws restricting expression online is imperfect and never complete. The extent to which authorities seek to limit this imperfection tends to correspond to the degree to which it frustrates an authority's legitimate interests and/or results in harm to other individuals in its jurisdiction.

Allowing content that has been determined to violate one country's laws to remain available elsewhere online opens up the possibility that individuals in the censoring country may continue to access it by circumventing technical restrictions. It is however important to recognize that most Internet users do not use circumvention tools, and those that do tend to use them episodically. As a result, the "harm" that may flow from the possibility of circumvention should be scrutinized carefully and on a case-by-case basis.

Efforts to eliminate digital content, including efforts to prevent "re-posting", tend to have broad extraterritorial impacts that extend beyond freedom of expression. In particular, efforts to proactively identify possibly infringing content often creates conditions that can lead to privacy infringements. Decision makers ordering content restrictions should be aware that their orders could impact the privacy and data protection rights of individuals both inside and outside their jurisdiction, and they should take steps to ensure such orders avoid or minimize infringing these rights.

c. **Economic impacts**

The hosting, display, and transmission of digital content can implicate a wide range of private enterprises, including web hosts, Internet registries, Internet service providers, mobile network operators, content-delivery networks, social media platforms, and financial intermediaries. Orders to restrict content, depending on their formulation, often impact multiple such entities directly or indirectly.

Decision makers ordering content restrictions should consider the extent to which private actors on the receiving end of such orders have the technological and economic means to implement them. Given the importance of fostering competition and innovation in the ICT sector, particular attention should be paid to the impacts such restrictions may have on smaller or start-up actors.

d. **Setting precedent**

Decisions to restrict a particular piece of content are rarely made in isolation. Such decisions tend to build on previous actions taken with respect to similar content, and also to impact future decisions. Decision makers should be cognizant of the possibility that any given restriction could be cited as precedent for future decisions by other decision makers in other contexts.

To the extent restriction decisions cumulatively reveal patterns, they can also impact the decisions of individuals whether or not to post content. While this can constitute effective "deterrence" against future violations, where restriction patterns are vague and protections for expression are unclear it can also lead to the "chilling" of legitimate expression.

The extent to which decisions can be contextualized and narrowly applied will help mitigate against misreading or misapplication by individuals or other decision makers.

CRITERIA G - GEOGRAPHICALLY PROPORTIONATE ACTION

1. **Different normative basis**

Content restriction decisions are increasingly based upon two distinct sets of rules: (1) national laws and (2) providers' ToS and Community Guidelines.

a. National laws

National laws pertaining to illegal content vary widely on a country-by-country basis. As a consequence, content that is legal in a country may be illegal in another. In addition, some countries may have national laws barring expression that is protected under international human rights standards. In recognition of the uniquely un-territorial nature of digital content, and out of respect for the laws of other sovereign nations (comity), as a general rule, where a determination is made that digital content violates the laws of a particular jurisdiction, any related decision or action to restrict that content should be limited – to the extent possible – to that jurisdiction. An assessment of the level of international normative consistency (see Operational Criteria B - Normative Basis) can help ensure that restrictions on the basis of national laws are geographically proportionate.

[1] *It was highlighted that the core and foundational purpose of companies' ToS and Community Guidelines is to maintain the nature and benefits of the service for the user community.*

b. Providers' ToS / Community Guidelines

By nature, providers' ToS / Community Guidelines[1] generally apply to the entirety of their services, irrespective of users' nationality or location. Content restriction decisions on the basis of these norms are therefore generally global by default. Ensuring the respect of proportionality standards in those instances remains nonetheless crucial, to preserve the broadest availability of legitimate content. A particular act or form of expression may violate relevant rules at a given time in a specific place. Yet, it may not be prohibited in another place or time, or even in the same place and time under different circumstances. Community Guidelines should allow for such a nuanced analysis.

2. **Default approach and exceptions**

The normative basis invoked for a content restriction has a direct relation with its geographic extent, as illustrated by the table below, which can help identify the default action associated with each case:

	Geographically limited restriction	**Global restriction**
Illegal according to local laws	Unless the rationale for the request is clearly contrary to international human rights standards (no restriction by provider), by default, the content item is restricted locally by the provider (for instance through geo-IP filtering).	A global restriction can exceptionally be implemented by the provider in response to a request / order if a proper justification is provided (e.g. high international normative consistency).
Content contrary to ToS / Community Guidelines	According to the diversity of local circumstances, the content is restricted in the most geographically proportionate manner.	The content is generally globally restricted when clearly in violation of the ToS / Community Guidelines, except if a court issued a local stay-up order.

CRITERIA H - CHOICE OF ACTION

The actions that can be implemented to deal with content that is illegal, harmful or contrary to ToS/Community Guidelines are increasingly diversified. The choice of the appropriate measure in each case is an important component to achieve the least restrictive effect.

The following Typology of Actions is a list of recognized actions available to platforms, intermediaries or states to block allegedly harmful or illegal content. This list was developed to illustrate and map a spectrum of possible responses and should not be understood as a normative index of actions that should be considered as equally valid.

[1] *Among primary sources used to compile this list were:*

Internet Society - Perspectives on Internet Content Blocking: An Overview

Daphne Keller - A Glossary of Internet Content Blocking Tools

Internet Society - Summary of Content Blocking Techniques

IETF - A Survey of Worldwide Censorship Techniques

ACTIONS	DESCRIPTION
CONTENT PROVIDER/PLATFORM AND SEARCH ENGINES	
Additional context	Additional context may be required for posting certain types of content, and may include explanatory information or URLS to additional sources of information and alternative perspectives. For instance, graphic images of historical, artistic or scientific significance, might requires context for the users to understand or appreciate the image. It may also be used in situations where content is deemed to be extremist or a form of disinformation.
Labelling	Label content with a warning for a specific type of content (i.e. violent content).
Age Verification / Age-gating	Age verification is undertaken by platforms to ensure that content is accessed only by users of the appropriate age. Age-gating prevents access to content and services by underage users according to national, regional or international laws.
Right of reply	Response for alleged defamatory content where the publisher/poster has opportunity to post a reply, counter-speech, or disclaimer.
Account suspension	Accounts may be de-activated or suspended for a temporary period of time due to policy violations or invalid traffic. During this time, the account may not be accessible (i.e. error message will show), or visible to the public, or key functionality may be de-activated (ability to post, comment, read data). Alerts may be sent to give account holder time to address the issue. If the issues are not resolved, the account may be disabled.
Account disabled	Users or content providers who are not in compliance with the relevant governing policy, may have their account permanently disabled so that it is no longer visible or active. Some platforms may not allow the user to create a new account on the same platform.
Anonymizing source documents	In cases involving alleged defamatory information (i.e. “Right to be Forgotten”), names may be removed from source documents such as newspaper articles or public documents, and replaced with initials or a random letter (i.e. X or Y).
Block search indexing	In cases involving allegedly defamatory information, content from individual pages can be de-indexed/de-referenced so it cannot be found through internal (i.e. news archive) or external search engines. This is done either by including a noindex meta tag in the page's HTML code, or by returning a 'noindex' header in the HTTP request.
Block keywords	Content providers and search engines can block specific keyword search terms to prevent associated content from being found via search results. For example, on Tumblr, searches for keywords associated with adult content will come back with no results, even if there are matches.
Take down: temporary or permanent removal	Mechanism where an individual can issue a legal request to a content host that requires the host to take down, delete or restrict access to allegedly harmful content. Examples include "Right to be Forgotten" policies in the EU and the copyright-oriented notice and takedown regime of the United States Digital Millennium Copyright Act.

Down-ranking /voting (modifying the visibility of content)	Down ranking is used to demote content visibility (as Google web search has done on DMCA grounds) for content posted by confirmed bad-faith actors who intend to manipulate or divide the conversation.
Quarantining	Potentially harmful content may be quarantined to prevent it from being viewed by users. Quarantined content usually will display a warning for users who may not wish to view it, or require users' opt-in to view it.
Geo-blocking/Geo-IP-filtering/Withholding Content	Platforms can "withhold content" or block target content, or users and content at once. This can be done by blocking all users from a geographic region, from specific IP addresses, or other applications. An example of geographic blocking is "country withheld content" (CWC) which could happen, for instance, if a tweet violates local laws or if it is blocked due to a court order.
Shadow banning	Shadow banning restricts the visibility and reach of a user's content without their knowledge. This discreet ban allows the user to perform all the normal activities on a site but may prevent his/her profile or posted content from being visible to others or restrict the reach of the content by preventing it from appearing in feeds or showing in search results. This might be done to allow problematic or possibly harmful content to remain up while preventing those not seeking it from finding it. It may prevent bad actors from simply starting a new account if they knew of the ban, or alternately, it may encourage bad actors to leave a platform due to lack of engagement. It is also a common technique for combating bots and trolls. Other terms for this include, stealth banning, ghost banning or comment ghosting.
Platform-based blocking	In cooperation with platform, content or specified search results are blocked from coming back from the search engine. This is often initiated by national authorities to block "illegal" content within a geographic region and thereby avoid blocking an entire platform. In some cases, it may be done by platforms to block content that violates its ToS or points to malware.
NETWORK / HOSTING INTERMEDIARIES	
Deep packet inspection-based blocking	A device is inserted in the network that blocks based on keywords and/or other content (e.g. file name). This technique is often used for data protection, anti-spam and anti-malware (anti-virus), and traffic prioritization.
Keyword block lists	*A keyword block list is a tool used by hosting intermediaries to filter keywords, and other forms of ID for video or audio. The filtering can be automated or done in combination with human monitoring. States also employ keyword blocking to censor content.*
URL or HTTP Header Based Blocking	*A device is inserted in the network that intercepts web requests and looks up URLs against a block list.*
IP and protocol-based blocking	A device is inserted in the network that blocks traffic based on IP address and/or application (e.g. VPN) between the end user and the content.

Internet Service Providers (ISPs) [point of control]	*ISPs are very effective points of control as they are easily identifiable and can readily identify the regional and international traffic of all users. Filtration mechanisms can be placed on an ISP via governmental mandates, ownership, or voluntary/coercive influence. ISPs can stop all its users from going to a website or using an app. Blocking can be done based on a URL, IP address (all content associated with IP or partial); technical specifications (such as blocking a port to prevent use of VOIP).*
Geographic IP-filtering	A website can partially or fully block users with IP addresses from a certain country or based on GPS, Wi-Fi network identification, or other technical information.
Performance degradation	Performance degradation involves the intentional decrease in connectivity and response speed throughout a given network. "Bandwidth throttling," for instance, may be done to manage network congestion or to partially block a percentage of traffic from specified IP addresses or other applications.
Packet dropping	Packet dropping interrupts traffic flow by not properly forwarding packets associated with the harmful content. This technique is most effective when the packet contains transparent identifiers linked to the specified content, such as the destination IP. It often results in over blocking.
DNS-based blocking / Geographic TLD blocking	At the network or ISP level, Domain Name System (DNS) traffic is funneled to a modified DNS server that can block lookups of certain domain names.
DNS Interference	*DNS interference results in an incorrect IP address being returned in response to a DNS query to a censored destination. Users may receive an error message.*
Domain name reallocation / seizure	Domain names may be reallocated or seized legally (i.e. criminal copyright violations) or extrajudicially when a top-level domain (TLD) deregisters a domain name to prevent DNS servers from forwarding and caching the site.
Network disconnection or adversarial route announcement	This is a form of technical interference where a whole network can be cut off in a specified region when a censoring body withdraws all of the Border Gateway Protocol (BGP) prefixes routing through the censor's country. This is an extreme and extensive form of blocking usually only undertaken for short periods under dire circumstances.
Server Takedown	If undesirable content is hosted in the censoring country the servers can be physically seized or the hosting provider can be required to prevent access.
EXTRALEGAL BLOCKING	
Blocking/ Interference/ RST Packet Injection	A specific type of packet injection attack that is used to interrupt an established stream by sending RST packets to both sides of a TCP connection; as each receiver thinks the other has dropped the connection, the session is terminated. This is also known as a "man in the middle" attack.

PART IV - NOTIFICATION / RECOURSE

CRITERIA I - USER NOTIFICATION

1. **Timing of the notification**
   a. **User notification before action**
   When a decision to restrict content has been taken on the basis of a national law or of a provider's ToS or Community Guidelines, user notification should occur prior to the decision being implemented. This can allow the user to decide either to modify the content so that it does not infringe on the relevant normative basis and/or contest the decision.

   b. **User notification simultaneous to or after action**
   In some situations, it may not be practical, advisable or permissible that notification occurs ahead of the content restriction decision's implementation. This can include inter alia the providers' best evaluation of how to minimize potential harm.

   c. **Exceptions to user notification**
   Exceptionally, a situation may justify an exception to the general principle of user notification. This can include inter alia cases when the user cannot be identified, local legal requirements for confidentiality and the need to avoid thwarting ongoing investigations.

2. **Content of the notification**
   The notification should contain information pertaining to the normative basis and rationale for restriction along with the specific/respective channels, information and applicable timelines for recourse. For content restricted on the basis of the providers' ToS/Community Guidelines, notification also contains information pertaining to the specific clause/guideline that was violated.

CRITERIA J - RECOURSE

Recourse and appeal have emerged as important issues regarding online content restrictions. Two of the approaches currently considered are addressed here: company-established independent review bodies, and country-based self-regulation councils. Discussions on these approaches can be organized around the structuring questions in the two concept notes below. These concept notes do not however address or prejudge the level of support for either of these approaches.

I. Company-Established Review Bodies

The following note explores elements related to the potential creation by companies of mechanisms to provide an independent appeal of their content restriction decisions made on the basis of their Community Guidelines. It is understood as a company-specific instrument[1] with binding authority at the third level of a decision-making escalation path following initial first instance decisions and reconsideration[2].

[1] *A specific body as opposed to something that would be established at a national level (such as Social Media Councils) or for a diverse group of companies.*

[2] *See Annex 2, infographic regarding the three stages.*

As detailed below, analogies with national Supreme Courts (or equivalent) have some validity but can only be pushed so far, given very significant differences in situations. Other inspirations might be also relevant and inventiveness is required in this radically new transnational environment.

The following preliminary questions could help structure "what if" discussions on the creation of such a mechanism:

COMPETENCE	DUE PROCESS	BODY	OTHER
• Topics covered • Normative reference • Initial source (AI/notices) • Cases filtering ("cert") • Mandate focus/limitation • Applicants • Remedies	• Limited steps/duration • Written/oral procedure • Adversarial process • Role of third parties • Decision-making • Production of rationale • Dissenting opinions • Expedited mechanisms • Transparency • Suspensive procedure	• Size • Composition • Members profiles • Designation • Mandate duration • Meeting frequency • Independence • Secretariat support • Funding	• Charter • Name • Advisory role(s) • Geographic scope • Thematic chambers • Mutualization • Electronic tools • Liability protection • Jurisprudence coherence

The elements in the table above are briefly detailed below:

1. **Competence**

   a. **Topics covered**: Community Guidelines cover diverse topics with different volumes of content restrictions and levels of automatic detection[3]. To keep the volume of expected appeal requests manageable, *should such a mechanism initially be open only for certain topics?* Would an option in that regard be to focus on issues (e.g. hate speech and bullying) where the impact on freedom of expression and the need for nuance are maximum, while the number of initial actions is relatively smaller (see Annex 1)?

   b. **Normative reference**: Community Guidelines would clearly constitute the primary normative source. *Should however the Charter of such a body also reference other sources*, such as international law (e.g. human rights principles and specific conventions) or even national laws (particularly if the body's remit also covers at some point requests from public authorities on the basis of national law)?

   c. **Initial detection**: Content restriction decisions are taken on the basis of 1) Artificial Intelligence detection, or 2) flagging by users or notices by public authorities. Appeals in the first case only involve the posting user and the company, while the second case creates a tripartite interaction, with potential impact on the procedure. *Should the envisaged mechanism only concern the first case, for the sake of simplicity, or cover both situations? Could this be established in phases?*

   d. **Cases filtering ("cert")**: Preventing the docket overload plaguing many high courts around the world is a fundamental success factor. The US Supreme Court only "grants cert" to 1,4 % of the annual submissions it receives (100 out of 7.000) and other jurisdictions have a similar practice. Case filtering is probably needed here but there is uncertainty as to the actual number of cases to be handled. *Beyond weeding out clearly frivolous cases, does this require a rapid early selection mechanism? What latitude should the body have in selecting the cases*

[3] See for instance the relevant data in Facebook's transparency report, in Annex 1.

it handles? Should there be specific expedited provisions to prioritize issues with a timeliness factor, such as urgency or tense local situations?

e. **Mandate focus/limitation**: Most Supreme Courts adopt or are subject to a restrictive approach in their case selection, in order for instance to focus on: conflicts between lower jurisdictions; clear challenges of interpretation of the law or a Constitution; or procedural dimensions of a lower judgment. *Should a similar approach be applied here and documented in the body's Charter and online submission forms?* Yet, it is important to note that in the examples above, the ultimate review mechanism sits on top of two levels of lower independent courts and a large body of public jurisprudence while the exercise here starts de novo, and is expected to intervene directly as a follow up to a simple internal process of reconsideration.

f. **Applicants**: An ultimate independent review is envisaged only against restriction decisions made by the existing appeal/reconsideration stage and not against an initial decision (rule of exhaustion of previous avenues for recourse). It is clearly intended to be open for a user whose content has been restricted. *Should this appeal also be open - and if yes, when - to notifiers whose requests for removal have been denied? In that case, should distinctions be made between public authorities and individual flaggers? And among the latter, between people directly targeted by the post and more general flaggers?* Opening recourse to notifiers adds procedural complexities.

g. **Remedies**: *What range of remedies can be ordered: mere reversal of the platform decision or also more granular and nuanced alternatives* (e.g. technical, geographic scope, warnings...)*? Can the body order the company to post a public correction?*

2. **Due process**

To fully respect human rights, an independent review body must take inspiration from elaborate due process requirements developed in various nations for courts dealing with freedom of expression. However, the expected large volume of cases and the need to keep the process manageable call for some adaptation. This means, inter alia, making choices on the following elements:

a. **Limited steps/duration**: Rather than several iterative phases, *should the procedure have a limited number of steps and/or duration? Can dedicated online formats for appeal help in that regard?*

b. **Written/oral procedure**: *Would the procedure be based exclusively on written briefs or on oral arguments as well? Would this vary in any way depending on cases?*

c. **Adversarial process**: This procedure can be seen in two different ways. Either as arbitrating a dispute between the company and the user, or as reviewing a lower level decision. *Would company representatives be a party in the procedure or only provide reasoning for the initial decision? Should individual notifiers directly affected by the posting under evaluation (if applicable) be part of the process?*

d. **Third parties**: *What is the possibility of their intervention in the procedure (e.g. legal representation, amicus provided by a supporting NGO or other parties) and conditions thereof?*

e. **Decision-making**: *What would be the majority rules for the body and any subsets of it?*

f. **Production of rationale**: Producing a rationale for every decision is potentially burdensome but an important contribution in setting up a coherent jurisprudence, as it establishes precedent. *Should this be implemented and if so, for all or only certain decisions (for instance in larger formations)?*

g. **Dissenting opinions**: *Can they be envisaged, and if so, under which conditions?*

h. **Expedited mechanisms**: Irrespective of the overall body size, *can most decisions be made by a limited number of members*, keeping larger formations for more delicate cases? Likewise, *can procedural guarantees vary* according to the perceived importance or complexity of the case, with for instance a mere one-step written procedure for the simpler ones?

i. **Transparency**: *What would be the level of publicity of deliberations and decisions? Should a repository of such decisions be put in place, and if so, by whom? Should some precautions be taken to preserve the rights of users, such as anonymization of decisions?*

3. **Body**

a. **Size**: If a single review body is formed per company, determination of its appropriate size can be informed by an analysis of practices in Supreme Courts or equivalent not only in the US, but also in Europe (e.g. France, the UK, Germany, ...), India, Brazil and other countries, as well as some regional ones (e.g. ECHR). These bodies significantly vary in size: 9 for the US Supreme Court, 30 in India, 47 for the ECHR, with varying formations sizes. In light of the expected large number of cases, the size and diversity of user communities, and the absence of a vast network of lower courts, *can a small number of members be viable?* On the basis of practices in arbitration and ADR, *should the establishment of a large roster of available arbiters or mediators to compose ad hoc panels be explored?*

b. **Balanced composition**: Community Guidelines cover several topics, requiring diversified expertise. Also, a company's global geographic range means a diversity of scripts[4], languages, cultural and political local contexts, that call for linguistic and local knowledge. *How to ensure the geographic, stakeholder, gender, age, cultural and competence balances* that will be key to allow nuanced decisions and establish legitimacy of such a body? *Should the composition be organized around specific constituencies? How can the interests of the user community be represented?*

c. **Members profiles**: A spontaneous impulse could be to mainly call on the expertise of former judges or lawyers. Yet, most are proficient in one particular body of national law and might keep a particular bias in that regard while the main body of norms to be enforced here is likely to be the Community Guidelines. *How to ensure a broad diversity of professional profiles and experiences, privileging people with exposure to a variety of environments?*

d. **Designation**: This may be one of the most delicate issues. Modalities of designation of superior court judges in countries are difficult to transpose here. A designation by company management alone - whatever the level - would probably not be perceived as fully legitimate. Yet, a full selection by the community itself raises many conceptual and operational challenges. *What innovative mechanisms can be designed* to enable the selection of people of high integrity, competence and dedication that will be seen by the community and the general public as forming a legitimate and trusted college? *Should diverse sources of designation be combined? Should specific constituencies play a role and if yes, how to form them if it is at a global level?*

e. **Mandate duration**: The lifelong mandate of US Supreme Court Justices is an outlier case and a limited mandate seems more appropriate. *What would be its appropriate duration (2, 4, more years)? Should there be a limited number of renewals (1, 2, more)?* Would partial rotation help avoid brutal changes in the composition of the group? Could a large group be established progressively (e.g. with 1/3 of the members every 2 years in the first 6 years, if such duration is retained)?

f. **Meeting frequency**: How frequently should the Body be in session? This should take into account the expected amount of cases, as a result of responses to the above questions of scope (in part 1) and procedure (in part 2).

g. **Independence**: This is a critical factor, whatever role the company might play in the designation of members. Given the expected volume of activity, *should members be expected to be fully dedicated to this mission for the duration of their mandate or not?* In any case, what should be the conflict of interest policies limiting their past or current activities, including their potential relations with the company?

h. **Secretariat support**: Such a body will need secretariat support to manage the process and conduct research. *Could automation reduce the overall burden in comparison with existing judicial processes?*

i. **Funding**: Will the financing of such a body be under the exclusive responsibility of the company? Or should there be other sources?

4. **Other**

a. **Charter**: A dedicated Charter for this independent review body will be necessary, detailing inter alia its mandate, normative reference basis, procedures, composition and mode of designation. *How should it be developed and what role can the corresponding user community play in that regard?*

b. **Name**: This note uses the expression "Independent Review Body" by default. Alternative names can be envisaged, such as Panel, Council, Committee, or equivalent. *What could be a proper name*, given that the term "Independent Review" has a strong benefit in terms of clarity.

c. **Advisory role(s)**: In addition to the appellate role on individual decisions envisaged above, should the following additional advisory roles could also be envisaged for such a body:
   - On a case by case basis early on, upon spontaneous request by the company in difficult or sensitive situations, even before a decision is made or the user is notified,
   - In a more general way, to provide guidance on best practices and refining of Community Guidelines, on the basis the cases it handles or some that would be shared by the company.

   In the first case and maybe also in the second one, would the company have the option to either follow the advice/recommendation (without any further justification), or not (in which case it may have to provide an explanation to the body, to help refine its jurisprudence)?

d. **Thematic chambers**: Many jurisdictions around the world have specialized chambers for different topics. *Should in due time different sub-groupings of the body be formed*, according to members' competences or personal preferences, to specialize in particular types of cases?

e. **Geographic scope**: Likewise, in order to ensure maximum understanding of local context, cultures and legal frameworks, *should sub-groupings of members* - in due time - *compose*

specialized chambers, for instance on a script or linguistic basis (rather than a purely geographic or national one)? How to maintain cultural diversity in such groupings to preserve some jurisprudential coherence? Can such independent review bodies alternatively be created at national level?

f. **Mutualization**: Implementation of an Independent Review Body by every single company might be difficult, especially for small ones. *Could some groupings be envisaged among several companies?*

g. **Electronic tools**: Like in the general moderation, significant automation of this independent review process can be achieved to manage the workflow of a large number of cases. *Can innovative mechanisms be explored* to enable collegial decision-making among people likely to be distributed in various locations around the world?

h. **Liability protection**: *Would the establishment of such an Independent Review body impact the liability regime of the corresponding company?*

i. **Jurisprudence coherence**: *How to ensure compatibility between the decisions of a diversity of Independent Review Bodies from different companies?*

NOTE: This note purposefully focuses mainly on decisions taken on the basis of AI detection. Independent review of decisions taken on the basis of notification raise additional procedural challenges and should also be envisaged in particular in two cases: flagging by individuals directly targeted or impacted by specific posts and notices by public authorities. This would require additional procedural steps and more analysis.

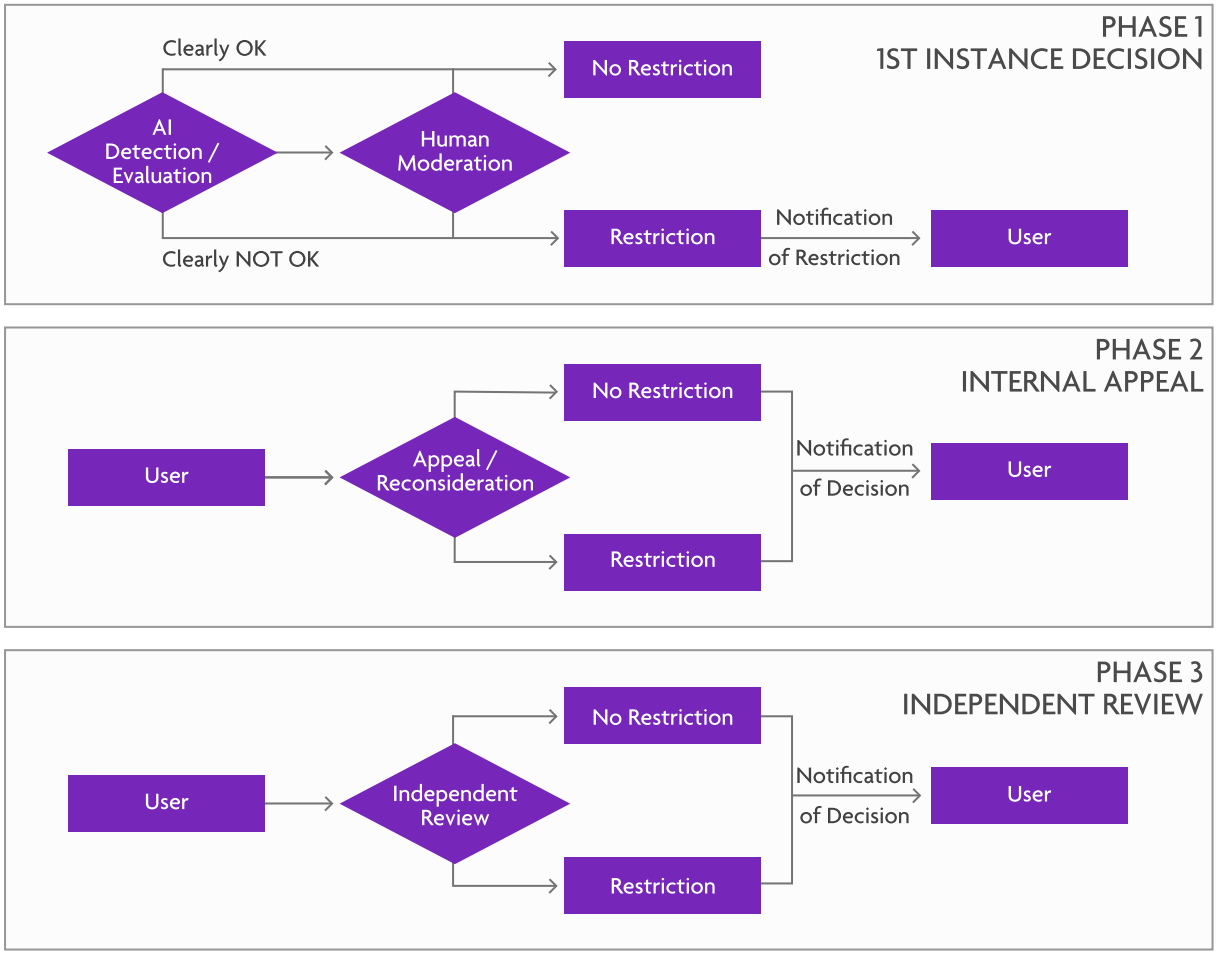
ANNEX 1: Typology of actions by topic

A recent transparency report by Facebook shows that actions taken can be grouped in three different clusters:

Very high annual volumes, with a strong security component and almost 100% detection by Facebook itself	Significant annual volumes, with a strong visual component and almost 100% detection by Facebook itself	More limited annual volumes, with a strong free speech component and lower detection levels by Facebook itself
• Spam (circa 4 billion actions/year) • Fake accounts (circa 2,4 billion/year)	• Nudity (100 M actions/year) • Child exploitation (36 M actions/year) • Graphic violence (25 M actions/year) • Terrorism (15 M actions/year)	• Hate speech (9,5 M actions/year, with 50 % automated detection) • Bullying (8 M actions/year, with 20 % automated detection)

It could make sense to focus the scope of this independent review system, at least initially, on the third categories (hate speech and bullying) where the impact on freedom of expression and the need for nuance are maximum, while the number of initial actions is - relatively - smaller.

ANNEX 2: Independent review as third stage in the escalation path



The process workflow corresponding to situations where decisions were taken upon notification (by users or public authorities) is not described here and involves additional interactions.

II. Country-Based Self-Regulation Councils

The present note identifies elements to help discuss the concept of Social Media Councils (SMCs). Creation of SMCs is proposed as independent self-regulatory bodies at the national level providing inter alia appeal mechanisms against content moderation decisions by Social Media Platforms (SMPs).

This approach emerged by analogy with existing self-regulatory practices in print media, in particular Press Councils, but significant differences must be taken into account. Other references might also be relevant and inventiveness is required in this radically new transnational environment.

The following questions could help structure "what if" discussions on the creation of such SMCs. They adapt and expand those introduced in this document to discuss another potential type of appeal: company-instituted Independent Review Bodies. Questions can be grouped in the following clusters:

COMPETENCE	DUE PROCESS	BODY	OTHER
• Companies covered • Topics covered (scope) • Normative reference • Initial source (AI/notices) • Cases filtering ("cert") • Mandate focus/limitation • Applicants • Authority • Remedies • Penalties	• Limited steps/duration • Written/oral procedure • Adversarial process • Role of third parties • Decision-making • Production of rationale • Dissenting opinions • Expedited mechanisms • Transparency • Suspensive procedure	• Size • Composition • Members profiles • Designation • Mandate duration • Meeting frequency • Independence • Secretariat support • Funding	• Creation • Charter • Name • Advisory role(s) • Geographic scope • Thematic chambers • Mutualization • Electronic tools • Liability protection • Jurisprudence coherence

The elements in the table above are briefly detailed below:

1. **Competence**
   a. **Companies covered**: *Would a SMC only accept recourse against decisions by companies having voluntarily accepted its authority or could it also be tasked to oversee any company "providing services in the country"?*
   b. **Topics covered**: Content moderation by SMPs covers a broad range of topics with different volumes of restrictions and levels of automatic detection. *Should such a mechanism initially be open only for certain topics*, to keep the volume of expected appeal requests manageable? Would an option be to focus on issues (e.g. hate speech and bullying) where the impact on freedom of expression and the need for nuance are maximum, while the number of initial actions is relatively smaller?
   c. **Normative reference**: *On what documents would an SMC decisions be based?* Existing public authorities' sources, such as national law(s), specific conventions, international human rights principles? Should the relevant companies' Community Guidelines be taken into account? Should new document(s) be developed ad hoc (e.g. a "Code of Ethics" or equivalent)? If so, should they be specific to each SMC or aspire to be broader (regional, even global)? And how should they be developed? Note: the expression "Code of Ethics" may be a wrong label if this deals with substantive norms (harmonization) and not the internal procedures of the Council.
   d. **Initial detection**: Content restriction decisions are taken by SMPs on the basis of 1) Artificial Intelligence, 2) user flagging or 3) notices by national public authorities, including court decisions. Appeals in the first case only involve the posting user and the company, while the two other ones create a tripartite interaction with consequences on the appeal procedure. Tensions might arise with public authorities in the third case. *Should the envisaged mechanism only concern the first case, for the sake of simplicity, or cover one or both of the other situations? Could this be established in phases?*
   e. **Cases filtering ("cert")**: Given the vast volumes of content posted and the numerous decisions made by SMPs, keeping the number of appeal cases manageable for an SMC can be a major challenge. *What filtering criteria and mechanisms* can help weeding out clearly frivolous cases, rapidly address repeat situations where precedent exists and focus on the difficult and potentially precedent-setting issues?
   f. **Mandate focus/limitation**: S*hould the area of competence of the body be explicitly framed in a limitative manner*, for instance in its Charter? This could cover for instance, beyond the topical scope covered above, requirements for: compulsory prior reconsideration within the company (e.g. by an ombudsperson), or a substantial connection

of the case to the country (with or without enumerative criteria). Should decisions only address procedural aspects of the company decision or the substantive part of the case?

g. **Applicants**: *Is the nationality or residence of the applicant to be taken into account, in particular to avoid forum shopping?* Furthermore, such an appeal is clearly envisaged as redress avenue for a user whose content has been restricted. Opening recourse to notifiers adds some procedural complexities. Reviewing decisions made on the basis of public authorities' request/order can raise questions of hierarchy of norms. *Should the procedure also be open - and if yes, under which conditions - to notifiers whose requests for removal have been denied? In that case, should distinctions be made between public authorities and individual flaggers, and among the latter, between people directly targeted by a post and more general flaggers?*

h. **Authority**: *How binding would the decisions of such a Council be for the participating platforms? In that regard, can there be different models in different countries and/or for different providers in the same country? Can there be two situations: binding authority for formal Charter adherents and non-binding for the others, even if this creates disincentives?*

i. **Remedies**: *What range of remedies can be ordered*: mere reversal of the platform decision or also more granular and nuanced alternatives (e.g. technical, geographic scope, warnings...)?

j. **Penalties**: Could a SMC, under certain circumstances, be able to impose penalties (monetary or not, for instance public excuses) on platforms as part of their decision?

2. **Due process**

To fully respect human rights, an appeal mechanisms such an SMC can take inspiration from practices in Press Councils but also from the elaborate due process requirements developed in various nations for courts dealing with freedom of expression. However, the expected large volume of cases and the need to keep the process manageable call for some adaptation.

This means, inter alia, making choices on the following elements:

a. **Limited steps/duration**: Rather than several iterative phases, *should the procedure have a limited number of steps and/or duration? Can dedicated online formats for submission help in that regard?*

b. **Written/oral procedure**: *Would the procedure be based exclusively on written briefs or on oral arguments as well? Would this vary in any way depending on cases?*

c. **Adversarial process**: *Would company representatives be involved in the procedure and if yes, how: as full party or only to explain the rationale of the first decision and the previous steps taken? Would public authorities or individual notifiers directly related to the posting under evaluation (if applicable) be part of the process? If yes, how?*

d. **Third parties**: *What is the possibility of their intervention in the procedure (e.g. legal representation, amicus provided by a supporting NGO or other parties) and conditions thereof?*

e. **Decision-making**: *What would be the majority rules for SMC decisions and any subsets of it?*

f. **Production of rationale**: Producing a rationale for every decision is potentially burdensome but an important contribution in setting up a coherent jurisprudence, as it establishes precedent. *Should this be implemented and if so, for all or only certain decisions (for instance in larger formations)?*

g. **Dissenting opinions**: *Can they be envisaged, and if so, under which conditions?*

h. **Expedited mechanisms**: Irrespective of the overall body size, *can most decisions be made by a limited number of members, keeping larger formations for more delicate cases? Likewise, can procedural guarantees vary according to the perceived importance or complexity of the case, with for instance a mere one-step written procedure for the "simpler" ones?*

i. **Transparency**: *What would be the level of publicity of decisions? And deliberations?*

j. **Suspensive procedure**: If the company action remains in effect during appeal, *should a specific procedure allow content to be reinstated pending the decision? If yes, under which circumstances?*

3. **Structure**

a. **Size**: Smaller bodies are more manageable but constraints of balanced composition (see below) go in the opposite direction. *Would the size of an SMC vary in relation to the size of the country and the number of cases it is likely to address*, in light of the answers on scope and mandate questions above?

b. **Composition**: The concept of Social Media Councils is based on the representation of different categories of actors, in particular companies and different types of civil society organizations. *How to identify the relevant constituencies and define the balance between the different groups? Should local authorities have a representation, and if yes, of which nature: full decision-making role or not? How does the composition vary in relation to local circumstances and would some common minimal guidelines exist?*

c. **Members profiles**: Content moderation covers several topics, requiring diversified expertise. Also, the national approach calls for linguistic capacity and knowledge of the local context and legal system. *How to ensure the stakeholder, gender, age and competence balances that will be key to allow nuanced decisions and establish legitimacy of such a body? How can the interests of the user community be represented?*

d. **Designation**: Press Councils usually rely for their formation on pre-existing professional associations (e.g. media, journalists, etc.) that can designate by elections the occupiers of the respective seats. The new field of Social Media may however not be as structured as traditional media. *How should constituencies be determined? How diversified can the modes of designation be? Can innovative mechanisms be designed to enable the selection of people of high integrity, competence and dedication? Should diverse modes of designation be combined?*

e. **Mandate duration**: *What is the appropriate mandate duration for Council members? Should there be limits on renewals? Should renewal be on a rotating basis to ensure continuity?*

f. **Meeting frequency**: *How frequently should the Council be in session?* This should take into account the expected amount of cases, as a result of responses to the above questions of scope (in part 1) and procedure (in part 2). *Would too infrequent meetings place a large and potentially disproportionate responsibility on a Secretariat?*

g. **Independence**: *What degree of independence should be established, 1) for the institution as a whole, in particular vis-à-vis the national government, and 2) for each member of the Council? In the second case, should members be expected to be fully dedicated to this mission for the duration of their mandate or not? In any case, what should be the conflict of interest policies limiting their past or current activities? Should members be remunerated?*

h. **Secretariat support**: Such a body will need secretariat support to manage the process and conduct research. *Could automation reduce the overall burden in comparison with existing judicial processes?*

i. **Funding**: *Would the financing of such a body be only based on contributions by participating SMPs? According to what criteria (various size metrics, related activity in the country, ...)? Would criteria and levels be set by the Council itself? Should there be other sources of financing, including from the corresponding government?*

4. **Other**

   a. **Creation**: *Who sets up such an SMC: a spontaneous coalition of companies and civil society actors? The same, but incentivized by the local government? The local government though a formal legislation still guaranteeing independence (or not)? Can this vary from country to country (see analogy with the different regimes of ccTLDs)?*

   b. **Charter**: A dedicated Charter for this independent review body will be necessary, detailing inter alia its mandate, normative reference basis, procedures, composition and mode of designation. *How should it be developed? What form would the commitment of participating SMPs to implement decisions take?*

   c. **Name**: This note uses the expression "Social Media Council" by default. *Can alternative names be envisaged, corresponding to potential different approaches among countries?*

   d. **Advisory role(s)**: In addition to the appellate role on individual decisions envisaged above, should the following additional advisory roles also be envisaged for such a body:
      - On a case by case basis early on, upon spontaneous request by a company in difficult or sensitive situations, even before a decision is made or the user is notified,
      - In a more general way, to provide guidance on best practices and refining of Community Guidelines, on the basis of the cases it handles or some that would be shared by companies.

      In the first case and maybe also in the second one, *would the company have the option to either follow the advice/recommendation (without any further justification), or not (in which case it may have to provide an explanation to the body, to help refine its jurisprudence)?*

   e. **Thematic chambers**: Many jurisdictions around the world have specialized chambers for different topics. *Should in due time different sub-groupings of a Council be formed, according to members' competences or personal preferences, to specialize in particular types of cases?*

   f. **Mutualization**: Implementation of Social Media Councils in every country might be difficult, especially for small ones. *Could some groupings be organized, on a geographic (e.g. regionally), linguistic or cultural basis? Would this require the existence of some degree of harmonized reference framework (e.g. the European Charter of Human Rights)?*

   g. **Electronic tools**: Like in the general moderation, significant automation of this independent review process can be achieved to manage the workflow of a large number of cases. *Can innovative mechanisms be explored to enable collegial decision-making?*

   h. **Liability protection**: *Would committing to a voluntary self-regulation regime such as a SMC impact the liability regime of the participating companies?*

   i. **Jurisprudence coherence**: *How to ensure compatibility between the decisions of a diversity of Social Media Councils in different countries, especially when they concern transnational cases?*

Note: This note primarily addresses appeals against content restrictions. Mechanisms for appeals against account suspensions are also needed and may raise additional questions.

PART V - SCALABILITY

CRITERIA K - CAPACITY OF SMALL PROVIDERS/COUNTRIES

1. **Small or atypical providers** are confronted with specific challenges in a rapidly evolving landscape regarding content moderation and restriction, including:
   a. Legal rules often established by reference to large well-known global actors, without sufficient criteria or thresholds for differentiated treatment of smaller actors or different types of services,
   b. Limited availability of human and financial resources to handle increased responsibilities regarding content evaluation,
   c. Difficulties to develop on their own or even access performing algorithmic and AI tools to comply with increasing detection responsibilities and short response times,
   d. Dependency on external hash database to prevent re-upload,
   e. Limited capacity to put in place recourse mechanisms of their own.

   Differentiated treatment and mutualization efforts must be envisaged as part of the general ecosystem of content moderation and restriction.
2. **Small countries**, likewise, are confronted with, inter alia:
   a. Globally uniform Community Guidelines developed by reference to major markets with limited consideration to domestic laws or sensitivities,
   b. Insufficient knowledge of local context or language by moderation teams, which makes accurate and proportionate action more difficult,
   c. Less established channels of communication with the major providers than larger countries, which can be harmful in situations of emergency or imminent threat to public order,
   d. Difficulties to establish local recourse mechanisms.

OPERATIONAL MECHANISM

NEW APPROACHES PERTAINING TO RECOURSE AFTER CONTENT RESTRICTION

CONTEXT

Every day, hundreds of millions of posts and hundreds of thousands of hours of videos are uploaded on the major internet platforms and made globally accessible, greatly facilitating freedom of expression. At the same time, legitimate concerns are raised regarding increasing harmful behaviors. Addressing abuses while protecting human rights has become a central issue of the global digital society.

Service providers have an important role to play in the identification and moderation of content that is illegal or not in compliance with their Terms of Service (ToS) and Community Guidelines. This has been translated into various normative frameworks, including self-regulation, codes of conduct or hard regulation. In addition, the numerous decisions on content restriction taken by providers are expected to be made in short timeframes to limit potential harm.

The use of automated tools increasingly allows detection at scale of potentially infringing content, but entails risks of bias and false positives of negatives. The increased reliance on ToS / Community Guidelines as the basis for content restriction decisions has in parallel magnified the norm-setting and decision-making roles of providers.

In order to ensure that content moderation and restrictions are proportionate and conducted responsibly, renewed attention is being paid to recourse mechanisms allowing users to contest a decision to restrict their content. New approaches at various degrees of development have emerged in recent years, including:

- **Company-established independent review**
  [As detailed in Operational Criteria J - Recourse] - Some companies explore mechanisms to provide an independent appeal of their content restriction decisions made on the basis of their Community Guidelines. It is understood as a company-specific instrument with binding authority at the third level of a decision-making escalation path following initial first instance decisions and reconsideration.
- **Country-based self-regulation councils**
  [As detailed in Operational Criteria J - Recourse] The establishment of independent self-regulatory bodies (Social Media Councils) at the national level is proposed to provide inter alia review mechanisms against content moderation decisions by providers.
- **Review by national authorities**
  Some actors have proposed that specific public authorities at the national level may have a formal role in reviewing content restriction decisions made by providers. The opinion of such bodies would be binding on the company, and geographically limited to the country.
- **Global advisory council**
  Finally, proposals for a global council with advisory power on companies' Terms of Service (ToS) and Community Guidelines have emerged, to increase transparency and accountability regarding this important normative basis.

NOTE: The list above is non-exhaustive, and does not address or prejudge the degree of support for any of those proposals.

RECOURSE MECHANISMS INTEROPERABILITY

The recent multiplication of initiatives and approaches to recourse mechanisms illustrates that actors have identified this issue as important and express the desire to address it. On the other hand, this proliferation raises major questions of interoperability, including:

1. **Jurisprudence coherence**: How can instances in which a decision made within one recourse mechanism contradicts the conclusions of another one be addressed? Should cases decided in such a manner have an impact on providers' ToS / Community Guidelines?
2. **Overlap**: How can duplication of efforts be avoided? In particular, if various separate mechanisms consider the same content restriction decision, how can coordination be best fostered?
3. **Liability**: How would potentially competing or complementing decision by various recourse mechanisms impact providers' liability? What consequences for providers' liability do conflicting decisions infer?
4. **Relation with national courts**: How can multiple recourse mechanisms interact with national courts? In particular, could decisions by independent review mechanisms be appealed before national courts?
5. **Respective responsibilities of actors**: What roles can each type of actor play in the various recourse mechanisms, to ensure that users' rights are respected, that processes remain efficient, and that excessive burdens are not created?

Every recourse mechanism that is implemented will partly address these issues. Yet, unless frameworks for coordination and cooperation between actors are established, there are significant risks that uncoordinated actions lead to unintended consequences, including a lesser protection of users' rights, duplication of efforts and high costs. Jointly developed norms and criteria can help structure the interactions between various mechanisms and ensure that interoperability is included by default in the implemented approaches.

EXPECTED BENEFITS

The creation of a dedicated group within the Internet & Jurisdiction Policy Network bringing together the diverse stakeholders addressing the issue of recourse mechanisms could provide the following benefits:

- Allowing actors developing proposals for recourse mechanisms to refine their project to ensure maximal utility. Relevant stakeholders can give and gather feedback in a neutral safe space, to ensure that standards of accountability and transparency are respected.
- Developing norms and criteria on cross-cutting subjects that need to be addressed collectively. These can be introduced in individual initiatives to foster interoperability.

NEXT STEPS

The 3rd Global Conference of the Internet & Jurisdiction Policy Network in Berlin can discuss the validity of this proposal, the potential mandate and timeline of such a group, as well as ways to ensure involvement of the most relevant stakeholders.