REGULATING ONLINE CONTENT – the way forward

WHAT IS CONTENT MODERATION?

When someone posts, or searches for, content on online platforms, a company manages that process and decides how information is stored and made available, and in which order it appears. As more and more people use the internet to access information, to express their views, protest and mobilise others, how private platforms and States restrict or incentivize pieces of online content has immense consequences for public debate and public participation.

LEGITIMATE CONCERNS, PROBLEMATIC SOLUTIONS

As digital technologies have revolutionized how we receive and share information, technology companies and state regulators have faced complex challenges. While initial debates focused on copyright protection and online child sexual abuse, the discussion has now increasingly shifted to how to prevent the spreading of extremist content, hate speech and disinformation.

States' regulatory efforts typically have encountered challenges:

Poor definitions of what constitutes unlawful or harmful content

Often based on loosely demarcated risks of "harm" and without providing evidence of the effectiveness and adequacy of the restrictions, laws in a variety of states have sought to outlaw poorly defined categories of "harmful" content and speech ("extremist", "hate speech", "fake news"). The resulting lack of clarity has made it difficult for companies to devise clear rules and likely breaches the requirement that offences must be clearly defined in law (the principle of legality).

Outsourcing regulatory functions to companies

Imposing liabilities on companies often means transferring the responsibility to adjudicate and implement restrictions on expression to private entities and bypassing justice systems. Platforms are ill-equipped to make some of these decisions and business incentives may drive filtering and censoring, rather than taking the risks of fines and more regulation.

Over-emphasis on take downs and the imposition of unrealistic timelines

Legislative efforts generally over-emphasize take-downs of content as the central option to deal with undesired content on platforms. This over-focus is unhelpful and potentially counter-productive, as supressing information tends to amplify the impact of certain messages. Depending on the context, other measures, such as giving prominence to reliable information side by side or flagging posts, can be more effective. Additionally, the imposition of short time frames for content detection and removal drives companies to privilege automated content moderation (and thus automated censorship insensitive to context) over human assessment of posted content.

Over-reliance on Artificial intelligence (AI) mechanisms

Platforms heavily rely on AI for managing large volumes of content, which is problematic when dealing with complex forms of speech such as "hate speech", which are extremely context-dependent concepts. In addition, the complex algorithms used in AI may reproduce and amplify bias of those involved in their development (e.g. researchers have noted a racial bias in some AI hate speech detectors).

> *"Smart regulation, not heavy-handed viewpoint-based regulation, should be the norm, focused on ensuring company transparency and remediation to enable the public to make choices about how and whether to engage in online forums." "Companies must embark on radically different approaches to transparency at all stages of their operations, from rule-making to implementation and development of "case law" framing the interpretation of private rules."*
>
> UN **Special Rapporteur on Freedom of expression** report to the HRC on online content regulation, 2018

REGULATING ONLINE CONTENT – the way forward

WHEN STATES REGULATE ONLINE EXPRESSION…– 7 KEY ASKS

1) ANCHOR LAWS IN HUMAN RIGHTS

International human rights law is the only recognised transnational set of rules on freedom of expression, and should underlie regulation in an area where cross-border coherence is key.

2) FOCUS ON PROCESS RATHER THAN CONTENT

Instead of trying to combat different types of speech that could be harmful but are hard to define, regulations should focus on platforms' processes that determine whether and how content is amplified or restricted, ensuring human review for complex decisions. This includes the approach to content moderation, making the process transparent, and creating effective channels for users to challenge restrictions.

3) IF REGULATING CONTENT, AVOID AMBIGUITY

While process regulation is preferable to content-based approaches, any attempts to regulate content on platforms must be based on laws and clear and unambiguous definitions, and any restrictions to expression must be necessary and proportional.

4) REQUIRE TRANSPARENCY

Full transparency on how companies and States intervene regarding online content is key to enable effective responses. While some companies publish transparency reports, these often offer selective and aggregated data. States must require full transparency from companies in relation to the way they curate and moderate content and share information with others and be transparent about their own requests to platforms (e.g. for eliminating content and for user data).

5) CONSULT STAKEHOLDERS

Unless regulatory processes benefit from meaningful participation of civil society in the design and evaluation of regulatory instruments, they can result in flawed outcomes. Too often, regulatory efforts rely on opaque or speedy processes where only companies and State authorities are engaged.

6) ENSURE ATTENTION TO CONTEXT

Given that expression is always deeply contextual and decision-making on speech necessitates a solid understanding of social, cultural, and linguistic nuances, regulation should require adequately resourced human review of content moderation processes to complement automated review and accessible channels for exchange with communities.

7) ENSURE REMEDIES

Measures to block or limit online content should be subjected to scrutiny and users should have effective opportunities to appeal against content related decisions they consider to be unfair. Implementation of systems by which users are expressly notified about content interventions and which offer the possibility for reversal of decisions are key to ensuring procedural fairness. Companies' procedures can assist to solve immediate concerns, but independent courts should have the final say over the lawfulness of content.