STRUCTURING QUESTIONS: CONTENT MODERATION IN RELATION TO COVID-19

REF: 20-104 | April 20, 2020

The COVID-19 crisis sheds a new light on the role of internet intermediaries as information service providers and commerce platforms. In this context, the Secretariat of the Internet & Jurisdiction Policy Network has identified 3 key issues that the current crisis has brought to the forefront:

I. **Increased reliance on AI in content moderation**
II. **Scope of Content Moderation**
III. **De facto role of internet platforms as public service information providers**

The Internet & Jurisdiction Policy Network[1] and its three Programs[2] have for more than five years facilitated discussions among public, private and civil society actors on jurisdictional challenges on the internet. This extensive experience has demonstrated the importance of a common frame of reference to properly address an issue, design solutions and evaluate their impact.

In that spirit, this Framing Brief from the I&J Secretariat, building on interactions with Policy Network members, presents a **list of structuring questions** regarding new challenges regarding cross-border content moderation in the context of the COVID-19 crisis. It aims to assist the various stakeholders in analyzing them in this exceptional context and contribute to the design of the most appropriate solutions.

Once this crisis is over, **a thorough evaluation** of the steps taken will need to analyse the consequences and lessons learned. This requires preserving all relevant information, and for internet intermediaries to consider special transparency reports on actions taken. An assessment of the best practices, and benefits of coordination and cooperation will help better prepare for the next crisis. Part of this analysis should include an evaluation of what elements of the procedures and practices introduced in this exceptional period will remain.

Structuring Questions for New Challenges

I. Role of AI in Content Moderation

In the wake of the health crisis major internet platforms announced that due to the inability of their moderators to access the necessary tools from home, they will need to rely more on AI moderation tools. This is a change from the practice where the technology is used as an identification and flagging tool to facilitate human moderators' decisions. It raises the question of the distribution of priorities for the reduced number of moderators to deal with child safety, terrorism, suicide and self-injury, and harmful content related to COVID-19. Reduced human capacity for review also leads to longer response times. Could this risk discouraging individual flaggers?

The Content & Jurisdiction Program Contact Group has previously discussed how increased AI moderation affects each of the four stages for content moderation - identification, evaluation, choice of action and recourse.

Identification

Criteria for AI identification

- **Substantive analysis:** Keywords, image recognition.
- **Virality:** engagement, speed of retransmission.
- **Prevalence** (e.g. numbers of followers, speed of dissemination).

[1] https://internetjurisdiction.net/

[2] https://www.internetjurisdiction.net/work

Algorithms

- **Addressing new topics:** How to program on a new topic? (what are the relevant data sets)
- **Smaller entities:** Availability of tools and datasets to smaller entities?
- **Hash databases:** How are they set up and what is the degree of mutualization between actors?

Evaluation

- **False positives/negatives:** How is the initial level of acceptable false positives/negatives set, taking into account reduced human capacity?
- **AI Training:** How is human feedback provided to train algorithms and revise these parameters?

Choice of action

- **Decision-Making:** What are the options for AI-based decision, only binary up/down or a more refined set of options (reduced virality, demonetisation, …)?

Recourse

- **Recourse/Remediation:** Ways to improve existing mechanisms or build new ones?
- **Accessibility:** Making content non-accessible vs permanent deletion to allow for reconsideration?

II. Scope of content moderation for e-commerce intermediaries

In relation to COVID-19, advertising, online marketplaces and e-commerce infrastructure platforms are facing a proliferation of blatantly fake cures and profiteering. Countries banning sales of medical equipment online. How can these internet intermediaries deal with this challenge? Most of the issues they have to deal with are already covered by their terms of service.

This situation necessitates increased coordination and sharing of information on violators between law enforcement and internet intermediaries, which raises the following questions:

- **Decision making:** Who decides on what products or services (eg. fake cures) are dangerous to public health in the current context?
- **Coordination:** What are the best practices in coordination for the determination on what is illegal or manifestly illegal and how to ensure consistency of content moderation across platforms?
- **Capacity Building:** How to build content moderation competences and institutionalize recourse/remediation mechanisms for actors who traditionally do not engage in such a role? The scale of the crisis also requires a change in the existing processes on how they deal with content, using a mix of manual tools, AI and screening while protecting fundamental human rights.
- **Scalability:** How can practices and solutions be scalable across smaller platforms, given the diversity of intermediaries? How can cooperation and best practices sharing be facilitated between platforms?

III. Public Information Announcements in times of crisis - proactive approach by platforms - vs obligations.

Internet platforms have become important information sources. Some have taken action to promote information from official sources. Unlike traditional media that have a clear role/responsibility/obligation for public service announcements, the role of internet platforms in that regard is undefined. Current unprecedented actions on the part of the platforms gives rise to the following questions:

- **Geographic scope of announcements:** How to determine it, given platforms' global reach?
- **Conflicting information:** How to reconcile conflicting information from different public authorities, between and within countries?
- **Global principles:** Is there value in establishing global principles for public information announcements?
- **Internet access:** How to ensure public information flow in locations with low internet bandwidth or prevalent Zero-Rating regimes?

REF: 20-104