Online harms: Content moderation and models of regulation

Dialogue 3 in the Series “Frontiers of Digital Economy”

Report on discussions of the Expert Round Table on “Content moderation and models of regulation”

27th October 2022, 0830 – 1030 UTC via Zoom

LIRNEasia Solutions is the affiliate of LIRNEasia, a pro-poor, pro-market think tank whose mission is catalyzing policy change through research to improve the lives of people in the Asia and Pacific using knowledge, information, and technology.
Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160. info@lirneasia.net
www.lirneasia.net
The “Frontiers of Digital Economy” series is supported and sponsored by Meta.

Executive summary

Social media, or platforms that make possible the publishing and sharing of user-generated content are a ubiquitous element today. These new media present a set of novel challenges because of their velocity and articulation of information dissemination. As platforms grapple with the challenge of mitigating harms, content moderation has become increasingly important, and has attracted controversy. As platforms regulate their respective spaces, the consequences of such content moderation decisions have raised concerns. In some instances, states have attempted to step in and co-regulate the content moderation of platforms.

Soft co-regulation is a recent development that presents an alternative. The New Zealand Code of Practice for Online Safety and Harms is an example of this approach. Its systems-based and whole-of-society approach seeks to anticipate and mitigate harms. Such a code is not a "one size fits all" solution. However, the flexibility of such codes allows for adaptation in diverse contexts and as technologies and speech practices change. Industry codes may be seen as "regulatory sandboxes" that allow much needed learning as other solutions are considered. Codes have the potential to enable nuanced dialogue among citizens, policymakers, and platforms as they collectively grapple to understand the constantly evolving and complex factors that contribute to online harms. They can provide valuable learnings and thereby assist in developing balanced and permanent solutions.

This report captures the gist of a virtual dialogue among actors from government, civil society, private sector, and academia in countries that share some commonalties on existing content moderation frameworks, tools, and regulatory options. The recommendations are to:

- Consider broadening opportunities for input by stakeholders not only when industry codes are formulated, but also when they are periodically revised in response to changed technology and other conditions.

- Consider soft regulatory industry codes that can be utilized as "regulatory sandboxes." Such an approach will allow the state and stakeholders to learn about the effects of regulation on the dissemination of harmful content and on freedom of expression. This will allow for informed legislation, if considered necessary. For example, such a code could consider including provisions for remedies for content generators/disseminators dissatisfied by content moderation practices of platform companies in industry codes.

- In parallel, develop capacity among those are expected to take action against illegal content in the respective countries to distinguish between illegal content and content that is undesirable, but not illegal.

- Maintain engagement in the form of dialogue between governments and platforms. Open feedback and a collaborative approach between parties when approaching content moderation in specific jurisdictions will help establish trust between the two parties. This could be especially useful when local needs, customs and traditions need to be taken into consideration.

The problem

Many forms of online harm are perceived, and remedies are sought from platform providers and from the state. The harms range from online bullying and intellectual property violations to incitement to, or facilitation of, violence.

Because platforms enable extremely rapid and articulated dissemination of user-generated content (UGC), remedies obtained through court orders or administrative actions fail to be fully responsive to the aggrieved parties and state authorities. Therefore, the responsibility for remedial action to remove or reduce the reach of UGC perceived as causing harm tends to fall on the platform providers.

Most platform providers have therefore put in place various modalities of content moderation, ranging from algorithmic takedowns and de-prioritization through moderation by human agents to bans and suspensions of those deemed to be repeat offenders. This may be in the form of "private regulation" (hereafter described as self-regulation) and soft or hard co-regulation whereby the state requires the platform providers to act in specified ways.

The state also engages in efforts to directly control UGC, usually in the form of ex-post prosecution of content originators and disseminators deemed to have committed an offense set out in a law. While the Virtual Dialogue sought to focus attention on content moderation and the regulation thereof, it was found in the course of the Dialogue that stakeholder positions on the regulation of content moderation practices were influenced by the actions or plans of state authorities with regard to direct control of UGC, therefore it will be mentioned as relevant in the report.

The Virtual Dialogue

The Virtual Dialogue was designed to focus on content moderation by platform providers and the regulation thereof by the state (co-regulation). Content moderation by platform providers is a "wicked problem,"[1] where different parties cannot even agree on the nature of the problem; none of the solutions can make everyone happy. Inaction will draw blame; action will be challenged as being biased in one direction or another. The objective was to assist the participants in making decisions on the solutions most appropriate to their circumstances, based on the experiences of other countries and relevant studies.

The composition of the participants of the roundtable by country from among the BBNMAPS (Bangladesh, Bhutan, Nepal, Maldives, Afghanistan, Pakistan, Sri Lanka) countries and by affiliation are given in Annex 1. The list of participants in given in Annex 2. The roundtable was conducted under Chatham House Rules.

The panelists for the Virtual Dialogue, moderated by Rohan Samarajiva of LIRNEasia, were:

- Brent Carey, CEO of Netsafe, New Zealand, spoke on how the Aotearoa New Zealand Code of Practice for Online Safety and Harms had been adopted, the platforms that had agreed to join it, and how it had worked in the past few months.

[1] Head, B.W. (2022). Wicked Problems in Public Policy. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-94580-0_2

- Meg Chang, Content Regulation Policy Lead, Head of APAC at Meta explained how Meta approached content regulation and what were considered content moderation best practices in a space where both technology and speech practices are dynamic.
- Professor Ershadul Karim of University of Malaya, Malaysia (and Bangladesh) spoke of the regulatory options for user-generated content on social media platforms.
- Sofyan Sultan of Soch Fact Check, Pakistan presented the fact checker perspective on regulation of content on social media platforms.
- Dr Gehan Gunatilleke, founding partner at LexAG, Sri Lanka & Research Fellow at University of Oxford situated content regulation in the context of an overreaching state.

The menu of solutions

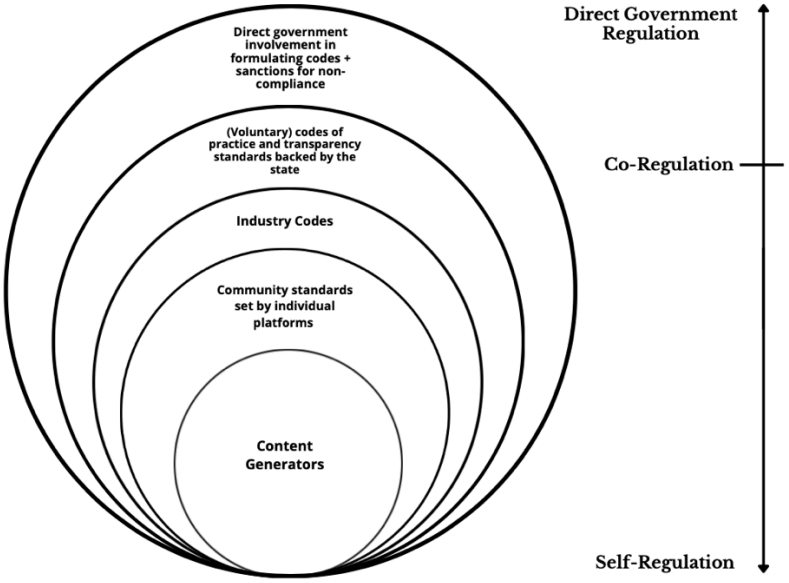
It was agreed that there is no one-size-fits-all solution. It was acknowledged that although genuine differences do exist, it is possible to use the lessons learned from other regions in the world, which can be critically evaluated in relation to specific circumstances in individual countries. There is no “one-size-fits-all" regulatory solution, especially because technology and speech practices are continually evolving. However, lessons and failings of different systems can and should be observed. Commonalities may be found in various aspects even under dissimilar conditions. There is no point in reinventing the wheel. This is indeed the underlying justification for this series of Virtual Dialogues involving key decision makers from BBNMAPS and those with relevant experience to share.

The principal solutions are given below. It is not necessary to pick one and exclude all others. For example, a “self-regulatory” solution can be implemented in a regulatory sandbox (see explanation in Box 1 below) that will permit learning about what works and what does not, while working toward a different solution.

- Soft “self-regulation.” In the face of demand from users for remedies against harms caused by UGC, platforms developed their own methods of moderating content. The criteria were based on community standards (mostly applicable to all countries) and national laws (specific to countries). The details of the criteria of what was unacceptable were not widely known. The publication of these criteria and the extent of consultation used to develop them varied from platform to platform. The actions taken when the standards are violated include algorithmic deprioritization and downranking of pages/individuals.
- Hard “self-regulation” is based on standards developed and enforced as above, but the sanctions are more severe: warnings, content takedowns, and prohibitions against posting for specified periods.
- Co-regulation involves parties other than the entity doing the moderation. In the soft form, co-regulation would require platform companies to adhere to codes of practice that are developed and enforced by non-governmental or industry-specific bodies. The sanctions for non-adherence are specified in the codes that the participating platform companies sign on to.
- In the hard form, the government authorities are involved in approving the codes, and in ensuring that the codes are followed. Sanctions are set out in law. Though conceptually distinct, the outcomes of hard co-regulation of UGC are indistinguishable from those of direct state regulation. The difference is that the platform provider acts as a proxy of the state, but without the usual safeguards of due process or natural justice associated with administrative actions by

the state. In fact, this form of regulation may be more severe, in that the platform company may engage in overbroad regulation because of uncertainty about views of the state authorities.[2]

Figure 1: Menu of content moderation solutions



Box 1: Regulatory Sandboxes – via *'Artificial Intelligence and Regulatory Sandboxes' - briefing by the European Parliamentary Research Service (ERPS)*[3]

A regulatory sandbox is an experimental space that allows entities to innovate and test out new products or services under a regulator's supervision within a given time period.

Regulatory sandboxes permit business learning, the development and testing of innovations in a real-world environment; and regulatory learning, the formulation of experimental legal regimes to guide and support businesses in their innovation activities under the supervision of a regulatory authority.

Regulatory sandboxes allow for "customization of regulation" while innovators navigate complex regulatory landscapes.[4] At the same time, it gives regulators time to deepen their understanding of new technologies before they attempt to regulate the same. Innovation takes place in a framework of controlled risk and supervision, if designed effectively. The use of regulatory sandboxes is common in the financial sector, increasingly so in the process of testing new developments in fintech. Countries including, but not limited to, the UK, Norway and Japan have opted for this approach, by utilizing "fintech sandboxes." This can similarly be extended to other areas, providing an agile environment for innovation and the regulation of the same in highly dynamic and technology intensive sectors.

[2] Canaan, I. (2022). NetzDG and the German Precedent for Authoritarian Creep and Authoritarian Learning. *Columbia Journal of European Law*, p. 101 on.
[3] https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf
[4] Ranchordas, Sofia, (2021) Experimental lawmaking in the EU: Regulatory Sandboxes. EU Law Live [Weekend Edition, 22 October 2021], University of Groningen Faculty of Law Research Paper No. 12/2021

Criteria for moderation

Each of the solutions outlined above requires those engaged in content moderation to have some criteria that will guide their activities. In the case of industry codes of content, the criteria will be spelled out and, ideally, be public. This could be the case with company specific self-regulation too. For effective co-regulation, agreement on the part of the platforms doing the moderation and the state authorities engaged in regulating that activity would be useful. Agreement cannot exist without explicitly stated criteria. Indeed, one of the main shortcomings of extant hard co-regulatory regimes is the fact that those engaged in content moderation under time pressure must keep guessing what criteria will be applied ex post by the legislatively empowered authorities.

The criteria used in content moderation, tend to be described as community standards,[5] a term drawn from US court rulings that specified that content that could be prohibited as "obscene" (illegal), or to which access could be limited because they were "indecent" (not illegal), or simply as policies. That practice was anchored on communities defined by physical proximity, which tended to overlap with municipal, state, or national jurisdictions. There are obvious difficulties in translating this approach to online spaces where communities tend to be defined by criteria other than physical proximity or presence within a legally constituted governance area.

The policies of major platform companies contain broad definitions and an indication that uploaded content should fall within local laws and norms. However, it is not that simple. For example, when governments believe content on Facebook or Instagram goes against local law, they may request the content to be restricted or there may be court orders to restrict content. Allegations that content is unlawful may be made by non-government entities and members of the public. Facebook and Instagram will respond to these requests in line with their commitments as members of the Global Network Initiative and the Corporate Human Rights Policy.[6]

Platforms such Facebook, Twitter, TikTok, and Reddit state that content may be restricted when a request that meets the above criteria is made by the authorities in a particular country. This is relevant especially when content, or the functioning of a group is legal in one country, but not in another. Instances where country specific laws were applied by platforms to moderate content include:

●Germany: Twitter blocked access to a Neo-Nazi group banned in the country, acting on their "country-withheld content" policy for the first time.

●India. Facebook restricted access to 337 items of user generated content, specifically for users in the country in response to directions from the Ministry of Electronics and Information Technology for violating Section 69A of the Information Technology Act between July to December in 2021.

●Pakistan. NSFW (Not Safe for Work) subreddits are banned by default.

The formulation of community standards and the industry codes require extensive participation of, and dialogue with, a range of stakeholders. In the case of the New Zealand Code, the "whole of society collaboration and cooperation" principle brings together a holistic, multi stakeholder approach which is more balanced than the alternative, which is a single entity acting on its own. While this includes a

[5] https://transparency.fb.com/en-gb/policies/community-standards/
[6] https://transparency.fb.com/data/content-restrictions/content-violating-local-law

variety of groups, the Dialogue pointed to two groups that may be unrepresented: the stateless such as refugees, and citizens belonging to a particular state who are untrusting of their authorities.

To establish frameworks that articulate clear, predictable, and balanced policies, it was deemed essential to consult as wide a group as possible – creating an environment where all these stakeholders share the responsibility of keeping users safe online by means of effective collaboration. Participants of the dialogue suggested that marginalized parties who are affected disproportionately by platform moderation decisions should be involved in teams that set user policies.

Additionally, a few participants of the dialogue expressed concern over algorithms used to make content moderation decisions. They were skeptical of the ability of algorithms to appropriately detect problematic content in some languages or recognize local nuances, and of the ability of moderators to identify or correct such content. However, it was stated that platforms are investing more in the development of processes and technology in hitherto neglected languages.

Self-regulation

"Self-regulation" may occur in two forms. In the first, the platform provider sets its own community standards and enforces them. It has greater flexibility in terms of modifying the standards and transparency of enforcement actions. In the second form, multiple platform providers agree to adhere to a consultatively developed industry code. Here, there is less flexibility regarding modifications and transparency.

Self-regulation by individual platforms provides greater flexibility to those engaged in content moderation, compared to an industry-wide self-regulatory scheme. From the perspective of the user, self-regulation based on an industry code provides greater certainty and is more conducive to generating trust in the process, especially when an independent entity is responsible for ensuring that the companies that have signed on to the code keep to their commitments.

It is not that flexibility is eliminated by industry codes. It was pointed out that rapidly evolving market and technological conditions made it necessary to keep improving the codes. If this was done using a holistic, multi-stakeholder approach, and the changes were well communicated to those engaged in content moderation and those affected by their actions, certainty and trust will not be negatively affected.

Co-regulation

Any content moderation policy is subject to challenge from multiple parties. Those who are seeking remedies for perceived harms will claim the moderation policies are too lax. Those sanctioned because of the policies will be unhappy about the standards and about their implementation being too strict. They will complain about the lack of due process and appeals. Government authorities may question the content moderation policies and their implementation for their own reasons or on behalf of either or both the above two parties.

In the mildest form of oversight of online content moderation, the state may mandate that the criteria and procedures used in content moderation be made public. Additional transparency requirements may include the publication of aggregate data on actions taken based on the standards. These actions need not be mandated by the state but may be voluntarily undertaken by the platform company.

Provisions for obtaining public input on the formulation and revision of community standards may be made by platform companies on their own, or as mandated by the relevant state authorities/statutes. Another layer would be appeal mechanisms being made available, within the platform company itself, or outside.

Once the government gets involved in hearing appeals from decisions emanating from the content-moderation process, it is likely that the government may also require some form of authority over the standards used for content moderation. In the strongest form, legislation specifying response time and penalties for original decisions that are second-guessed by government authorities may be put in place.[7]

The danger is that these actions will shade into government authorities making decisions on what content is allowed on the basis of opaque and possibly political criteria. Those who do not wish to see this happen within and outside government may wish to devise good safeguards, based on experience with different forms of self-regulation.

In light of this, it was posited that soft co-regulation can act as a placeholder until more comprehensive solutions are devised, if necessary. This gives an opportunity for states to improve the capacity of investigators, prosecutors, and other affiliated parties, while building citizen trust. Depoliticized and independent functions for managing online harms might be key in establishing trust and faith of citizens in the system. This could be an alternative to enforcing arbitrary and misinformed laws lacking adequate capacity to enforce them effectively, which could make matters worse.

Role of stakeholders in formulation of content moderation related policies

The formulation of community standards and the industry codes require extensive participation of, and dialogue with, a range of stakeholders. In the case of the New Zealand code, the "whole of society collaboration and cooperation" principle of the code brings together a holistic, multi stakeholder approach which is more balanced than the alternative – which is, a single entity acting as the arbiter of truth. While this includes a variety of groups, the dialogue raised attention to two groups that were mostly absent in the discussion: the first group being the stateless such as refugees, and the second being citizens belonging to a particular state who are less than comfortable with being represented by their respective state.

In order to establish frameworks that articulate clear, predictable and balanced policies, it was deemed essential to consult as wide a group as possible – creating an environment where all these stakeholders share the responsibility of keeping users safe online by means of effective collaboration.

[7] https://www.loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/#:~:text=Background%20on%20the%20Network%20Enforcement%20Act&text=The%20Network%20Enforcement%20Act%20is,after%20receiving%20a%20user%20complaint.

Recommendations

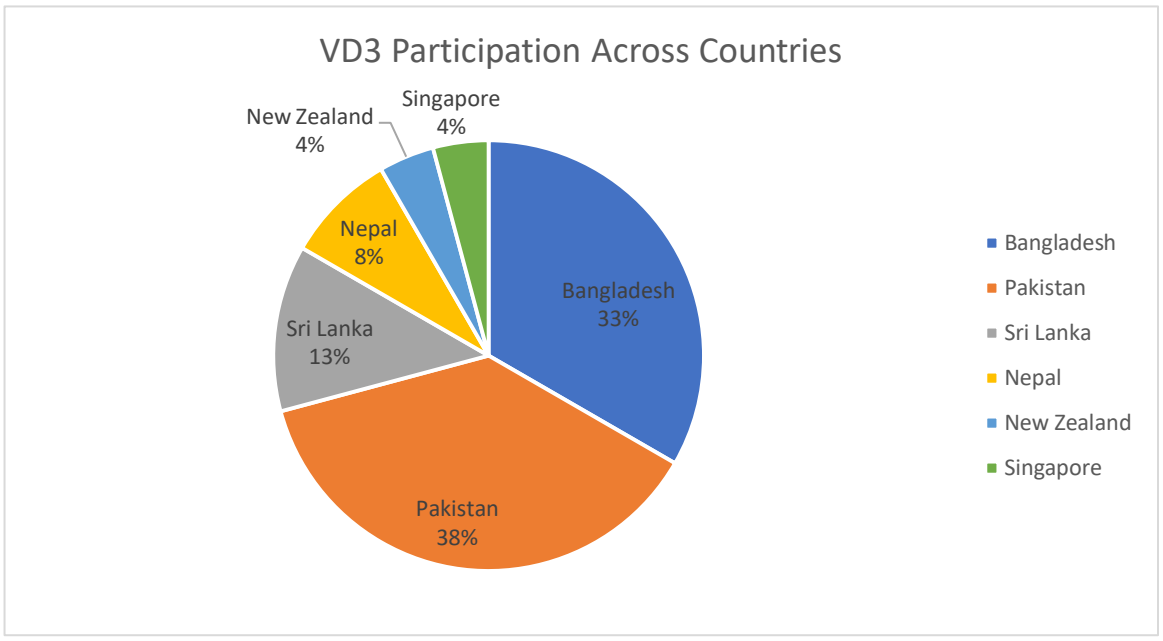
- Consider broadening opportunities for input by stakeholders not only when industry codes are formulated, but also when they are periodically revised in response to changed technology and other conditions.

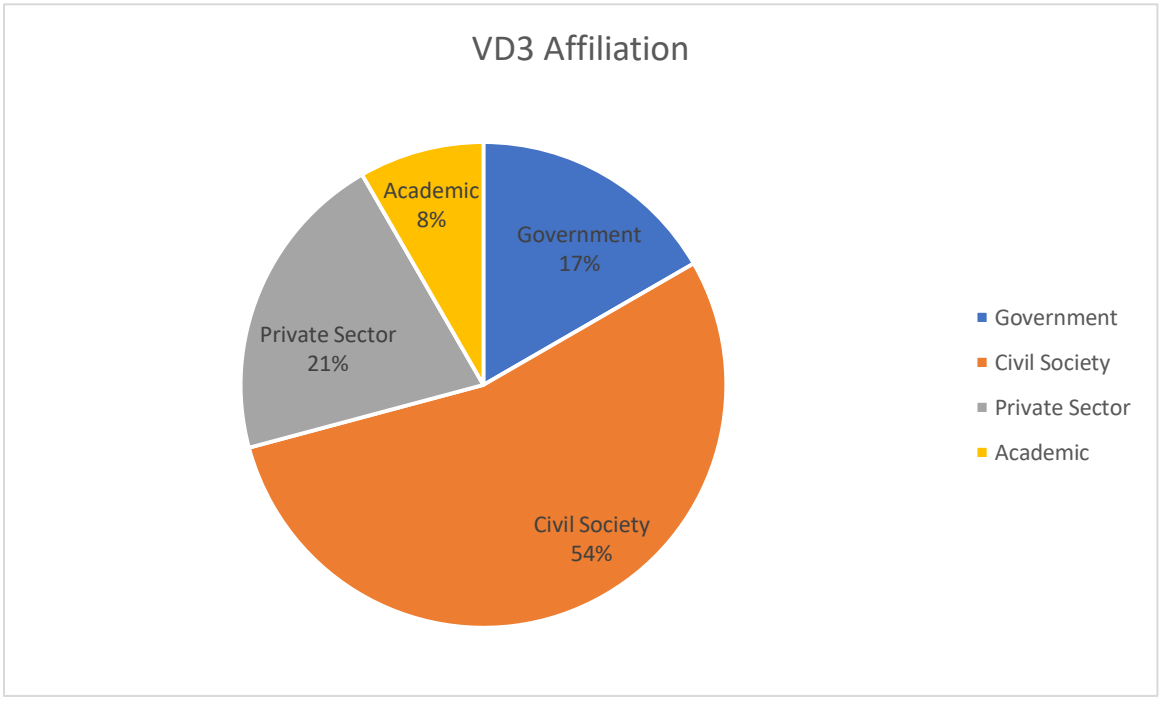
- Consider industry self-regulatory codes that can be utilized as “regulatory sandboxes.” Such an approach will allow the state and stakeholders to learn about the effects of content moderation on the dissemination of harmful content and on the rights of citizens freedom of expression. This will allow for informed legislation, if considered necessary. For example, such a code could consider including provisions for remedies for content generators/disseminators dissatisfied by content moderation practices of platform companies in industry codes.

- In parallel, develop capacity among those are expected to take action against illegal content in the respective countries to distinguish between illegal content and content that is undesirable, but not illegal.

- Maintain engagement in the form of dialogue between governments and platforms. Open feedback and a collaborative approach between parties when approaching content moderation in specific jurisdictions will help establish trust between the two parties. This could be especially useful when local needs, customs and traditions need to be taken into consideration.

Annex 1: VD3 Participation Analysis





Annex 2: List of Participants

Panelists		
Brent Carey	Netsafe	New Zealand
Meg Chang	Meta	Singapore
Gehan Gunatilleke	LexAG	Sri Lanka
Dr. Md Ershadul Karim	University of Malaya	Bangladesh
Sofyan Sultan	Soch Fact Check	Pakistan
Participants		
Sadiul Islam Antor	Bangladesh Legal Aid and Services Trust (BLAST)	Bangladesh
Babu Ram Aryal	Internet Governance Institute	Nepal
Atifa Asghar	Prime Institute	Pakistan
Miraj Ahmed Chowdhury	Digitally Right Limited	Bangladesh
Sheikh Manjur E Alam	Transparency International Bangladesh	Bangladesh
Arosha Fernando	President's Media Unit - Counter Disinformation Unit	Sri Lanka
Yasser Latif Hamdani	Irfan and Irfan Law firm	Pakistan
Faheem Hussain	School for the Future of Innovation in Society, Arizona State University	Bangladesh
Hija Kamran	Association for Progressive Communications (APC)	Pakistan
Usama Khilji	Bolo Bhi	Pakistan
Mr. Khalid Latif	Ministry of Human Rights	Pakistan
Aisha Moriani	Ministry of IT and Telecommunications	Pakistan
Prihesh Ratnayake	Hashtag Generation	Sri Lanka
Detepriya Roy	BRAC	Bangladesh
Muhammad Saad	Prime Institute	Pakistan
Quadaruddin Shishir	AFP Factcheck, Bangladesh	Bangladesh
Prabesh Subedi	Digital Media Foundation	Nepal
Mr. Babur Suhail	Ministry of IT and Telecommunications	Pakistan
Saimum Reza Talukdar	Member, Artificial Intelligence Working Group, HAC	Bangladesh