Recommended changes to the text of the Draft Amendment to the Information Technology (Intermediary Guidelines) Rules 2018

Submission to MeitY

IT for Change

December 2019

Dear Sirs,

Please find enclosed below, a set of comments (including suggestions for text) on the Draft Amendment to the Information Technology (Intermediary Guidelines) Rules (hereafter, Draft Guidelines). We are grateful for the opportunity to make these inputs.

We note that the Draft Guidelines are responsive to the need to re-think the extent of intermediary liability for third-party content, taking into account the increasing platformisation of online communication in the past two decades. Internet intermediaries are increasingly engaging in algorithmic content curation that enables them to "select the receiver of the transmission" (through personalized targeting) and "modify the information contained in the transmission" (by introducing flags/labels that can influence the users' reaction). Intermediaries therefore need to be held by a higher standard of accountability.

Our primary point of concern here is to address the normalisation of gender-based cyber violence.

The hon'ble Supreme Court's judgment in *Shreya Singhal v. Union of India* (2015) impacted two provisions of the Information Technology Act (2000) – striking down of Section 66A on 'offensive speech' and reading down of Section 79 on 'intermediary liability'. This has created two significant lacunae:
(a) The striking down of Section 66A has made it inordinately difficult to bring to book perpetrators of gendertrolling and online sexist hate speech.
(b) The reading down of "actual knowledge" in Section 79 to mean "receipt of court order" has resulted in delays in the removal of the impugned content in cases of gender-based cyberviolence. What this implies is that where intimate images of a woman are being circulated non-consensually, the offending image/video cannot be expeditiously taken down by notifying the intermediary. The de facto consequence of this is a gross violation of the woman's privacy and dignity.

The process of amending the Information Technology (Intermediary Guidelines) Rules offers an opportunity to remedy this state of affairs.

Our recommendations highlight the need for legal measures in the following directions:
a) content governance standards grounded in privacy, dignity and the prevention of harm
b) clear guidance on intermediary accountability for third party content
c) avenues for speedy redress for victims of sexist hate speech, gendertrolling and gender-based cyberviolence

Overall Comments on Rule 3(2)(b)

In 3(2)(b), we need the Draft Guidelines to be cognizant of:

- content that constitutes hate speech (defining hateful expression appropriately)
- content that undermines democracy and civil discourse
- content that violates privacy – such as in the non-consensual circulation of intimate images and acts of doxxing
- disparaging content that is directed at 'vulnerable groups'.

Content that already has a legal basis, which is currently under 3(2)(b) should be moved from (b) to (e).

It is recommended that terms that reopen Section 66A be avoided; "grossly harmful" for example, is similar to "grossly offensive", a category that was held by the Supreme Court to be "arbitrary" in the *Shreya Singhal* case. The term "harassing" cannot stand on its own. We recommend that these categories be dropped. "Blasphemous" may be covered under "hateful expression".

Proposed Text in Draft Guidelines	**Recommended Text**
3. Due diligence to be observed by intermediary *— The intermediary shall observe following due diligence while discharging his duties, namely: —*	**3. Due diligence to be observed by intermediary** — The intermediary shall observe the following due diligence while discharging his duties, namely: —
(2) Such rules and regulations, privacy policy or user agreement shall inform the users of computer resource not to host, display, upload, modify, publish, transmit, update or share any information that:	(2) Such rules and regulations, privacy policy or user agreement shall inform the users of the computer resource not to host, display, upload, modify, publish, transmit, update or share any information that:
(b) is grossly harmful, harassing, blasphemous, defamatory, obscene, pornographic, paedophilic, libellous, invasive of another's privacy, hateful, or racially, ethnically objectionable, disparaging, relating or encouraging money laundering or gambling, or otherwise unlawful in any manner whatever;"	(b)(i) spreads, incites, promotes or justifies forms of hateful expression based on religion, race, caste, place of birth, sex, sexual orientation, gender or disability. For the purposes of this provision, "hateful expression" comprises violent, dehumanising or denigrating communication. (b)(ii) promotes, endorses and incites gratuitous violence. (b)(iii) violates the privacy of individuals. (b)(iv) erodes the dignity of vulnerable groups.

Rationale and Explanations for the Recommended Text

For 3(2)(b)(i), we have recommended the inclusion of a definition of "hateful expression". The suggested text is in keeping with constitutional categories, also drawing inspiration from:
1. ***Council of Europe's definition of hate speech*** **-** "forms of expression which spread, incite, promote or justify racial hatred, xenophobia"
2. ***Facebook Community Standards*** **-** "We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define "attack" as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation."

For 3(2)(b)(iv), the understanding of "vulnerable groups" is derived from Part III of the Constitution and, inter alia, includes women, scheduled castes, schedules tribes, religious and gender minorities.

Overall Comments on Rule 3(2)(e) We recommend the use of 3(2)(e) for acts listed in 3(2)(b) that clearly reference existing provisions in law or court orders, not relating to hateful expression. This will satisfy the rule of *ejusdem generis*.	
Proposed Text in Draft Guidelines *(e) violates any law for the time being in force;*	**Recommended Text** (e) is pornographic, paedophilic, defamatory, libellous, voyeuristic or featuring rape, gang rape, child pornography or relating to or encouraging sexual harassment, stalking, doxxing, pre-natal sex determination, money laundering, or violates any law for the time being in force;
Rationale and Explanations for the Recommended Text The terms “pornographic, paedophilic, defamatory, libellous” have been placed here from 3(2)(b) of the Draft Guidelines. The term “voyeuristic” is added in light of 354C of the Indian Penal Code. The term “stalking” is added in light of 354D of the Indian Penal Code. The phrase “featuring rape, gang rape, child pornography” references the Supreme Court’s order dated 6th December 2018, in *In Re: Prajwala.* The term “harassing” is replaced by “relating to or encouraging sexual harassment”. The phrase “pre-natal sex determination” is added in light of the Supreme Court’s 2017 ruling in *Sabu Mathew v. Union of India.*	

Overall Comments on Rule 3(5)

Rule 3(5) in the existing draft requires intermediaries to enable "tracing out of originator of information", which assumes that it is possible "to trace the origin of the message without breaking encryption (of the content)".[1] However the following criticisms have emerged:

- The technical feasibility of facilitating such decryption involving originator information alone has not been conclusively proven.[2]
- Tracing origins of messages is likely to lead investigators merely to troll armies/chat bots, rather than the actual originators/conspirators who make decisions for online disinformation campaigns on other forums.[3]
- There is a likelihood that enforcing traceability will be counteracted by the proliferation of commercial services that allow for masking origin, especially when deployed by actors breaking the law.[4]

Matters like traceability that affect encryption-decryption should be covered through appropriate rules under Section 69 of the Information Technology Act rather than the Draft Guidelines, and should be governed in tandem with a strong data protection and privacy law. The phrase - "The intermediary shall enable tracing out of such originator of information on its platform as may be required by government agencies who are legally authorised" - should be redacted.

Proposed Text in Draft Guidelines	**Recommended Text**
"(5) When required by lawful order, the intermediary shall, within 72 hours of communication, provide such information or assistance as asked for by any government agency or assistance concerning security of the State or cyber security; or investigation or detection or prosecution or prevention of offence(s); protective or cyber security and matters connected with or incidental thereto. Any such request can be made in writing or through electronic means stating clearly the purpose of seeking such information or any such assistance. The intermediary shall enable tracing out of such originator of information on its platform as may be required by government agencies who are legally authorised."	(5) When required by lawful order, the intermediary shall, within 72 hours of communication, provide such information or assistance as asked for by any government agency or assistance concerning security of the State or cyber security; or investigation or detection or prosecution or prevention of offence(s); protective or cyber security and matters connected with or incidental thereto. Any such request can be made in writing or through electronic means stating clearly the purpose of seeking such information or any such assistance. ~~The intermediary shall enable tracing out of such originator of information on its platform as may be required by government agencies who are legally authorised."~~

1 https://www.republicworld.com/technology-news/apps/whatsapp-tracability-violation-fundamental-right-privacy.html
2 https://www.medianama.com/2019/08/223-kamakoti-solution-for-traceability-whatsapp-encryption-madras-anand-venkatanarayanan/
3 https://www.medianama.com/2019/08/223-iff-response-kamakoti-submission-traceability-2/
4 https://www.medianama.com/2019/08/223-iff-response-kamakoti-submission-traceability-2/

Overall Comments on Rule 3(8)

a. Amending Section 79

Section 79 of the IT Act is designed to implement a notice-and-takedown regime based on actual knowledge of the intermediary, which is now limited by *Shreya Singhal* to mean that intermediaries will lose safe harbour only on ignoring a notice in the form of a court order. However, in copyright violations, the Delhi High Court (*Myspace v Super Cassettes*) has interpreted the requirement of awareness to mean 'actual knowledge' of infringing works, not necessarily by way of a court order.

It is recommended that Section 79 be amended to encompass a notice-notice-takedown regime (modelled after the system adopted in New Zealand[5]) that is based on the following due process:

- On receipt of a complaint from any user that the Internet intermediary is hosting content in contravention of Rule 3(2) of the Draft Guidelines, the intermediary must, within 48 hours, send a copy of the complaint to the author of the content (while protecting the identity of the complainant, if requested to do so). If the intermediary is unable to contact the author after taking reasonable steps to do so (for instance because author identity is unknown), the intermediary should takedown the content within 48 hours after receiving notice of complaint.
- If the intermediary is able to contact the author, it has the duty to notify the author that they have a right to respond with a counter-notice within 48 hours of receiving the notice. The author has the option of submitting a counter-notice either consenting to remove the content or refusing to do so.
- If the author refuses to takedown the content, the intermediary must forward the counter-notice to the complainant. The complainant could then approach the appropriate court.
- If the author of the content fails to provide a counter-notice within 48 hours of receiving notice of the complaint, the intermediary must takedown or disable access to the content within 48 hours, after notifying the author.

Rationale and Explanations for the Suggested Amendment

The 'notice-notice-takedown' regime effectively balances the right to free speech with the right to freedom from gender-based cyberviolence, for the reasons described below:

1. Anonymous trolls will, in most cases, not respond with a counter-notice, thereby facilitating speedy takedown of the impugned content.
2. Trolling by bots can be speedily addressed.
3. The intermediary is not expected to rely wholly upon its own discretion for determining whether access to a piece of content deserves to be disabled. Instead, there are clear guidelines provided for the nature of such proscribed content (through Rule 3(2)), and the process to be followed in cases where takedown is contested/refused by the author of the content. When the illegality of a piece of content is disputed, it is up to complainants to take the dispute to court for adjudication.
4. Where the complainant takes an anonymous abuser to court, the intermediary can be asked by the concerned court to present evidence of meta data about similar notices served to the said abuser. Where multiple complainants have been harmed by the abuser, such a trail of evidence will ensure justice for multiple victim-survivors.

5 *See* in New Zealand Harmful Digital Communications Act, 2015.

Overall Comments on Rule 3(8)

b. The test of legality

The Draft Guidelines conflates the principles laid out in Article 19(2) of the Constitution against which the limits to free speech can be tested with the limits themselves. Any elucidation of instances where the intermediary is liable for content moderation must therefore be specifically spelt out in Rule 3(2).

Proposed Text in Draft Guidelines	**Recommended Text**
(8) The intermediary upon receiving actual knowledge in the form of a court order, or on being notified by the appropriate Government or its agency under section 79(3) (b) of Act shall remove or disable access to that unlawful acts relatable to Article 19(2) of the Constitution of India such as in the interests of the sovereignty and integrity of India, the security of the State, friendly relations with foreign States, public order, decency or morality, or in relation to contempt of court, defamation or incitement to an offence, on its computer resource without vitiating the evidence in any manner, as far as possible immediately, but in no case later than twenty-four hours in accordance with sub-rule (6) of Rule 3.	(8) The intermediary upon receiving actual knowledge in the form of a court order, or on being notified by the appropriate Government or its agency under section 79(3) (b) of the Act shall remove or disable access to ~~that unlawful acts relatable to Article 19(2) of the Constitution of India such as in the interests of the sovereignty and integrity of India, the security of the State, friendly relations with foreign States, public order, decency or morality, or in relation to contempt of court, defamation or incitement to an offence,~~ content that is in contravention to sub-rule (2) of Rule 3, on its computer resource without vitiating the evidence in any manner, as far as possible immediately, but in no case later than twenty-four hours in accordance with sub-rule (6) of Rule 3.

Rationale and Explanations for the Recommended Text

The phrase, “unlawful acts relatable to Article 19(2) of the Constitution of India such as in the interests of the sovereignty and integrity of India, the security of the State, friendly relations with foreign States, public order, decency or morality, or in relation to contempt of court, defamation or incitement to an offence” is the constitutionally envisioned standard against which any limitations to freedom of speech and expression should be tested. The limitations themselves need to be elucidated specifically. Using constitutional provisions in the place of specific restrictions on free speech will result in the delegation of law making to intermediaries, policemen and judges (as highlighted in Kartar Singh v. State of Punjab, 1994).

The Draft Guidelines therefore need to spell out the instantiations that vitiate this cornerstone. We suggest the phrase is replaced by “content that is in contravention to sub-rule (2) of Rule 3”, since Rule 3(2) is where pertinent requirements on content moderation have been placed on intermediaries for due diligence under the Draft Guidelines (failing which they may lose immunity provided under Section 79.)

Overall Comments on Rule 3(9)

The consensus of the progressive technological community is that algorithms are currently not well-equipped to judge the appropriateness of content. Although algorithms should be deployed for identifying potentially violating content, algorithms should not take the decision of actioning content. Any action taken upon unlawful content should be based on human discretion.

Algorithmic audit standards will need to be operationalised and notified by the Ministry of Electronics and Information Technology. These audit standards must be followed by intermediaries who design algorithms for content moderation, taking into account gender-based experiences of violations and abuses online. Once algorithms are deployed, there must also be a dedicated institutional mechanism for public scrutiny and process audit for appropriate application of content related laws.

Proposed Text in Draft Guidelines	**Recommended Text**
(9) The Intermediary shall deploy technology based automated tools or appropriate mechanisms, with appropriate controls, for proactively identifying and removing or disabling public access to unlawful information or content.	(9) The Intermediary shall deploy technology based automated tools or appropriate mechanisms, with appropriate controls, for proactively identifying and removing or disabling public access to content in contravention to 3(2) in accordance with 3(8) and involving human intervention. *Explanation*: "appropriate controls" for the purposes of this provision are as described below: (i) Intermediaries shall publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of the content guidelines. (ii) Intermediaries shall provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension. (iii) Intermediaries shall provide a meaningful opportunity for timely appeal of any content removal or account suspension. (iv) Intermediaries shall ensure that their algorithms comply with standards and processes of public scrutiny for appropriate application of content related laws.

Rationale and Explanations for the Recommended Text

The phrase "unlawful information or content" is replaced with "content in contravention to 3(2) in accordance with 3(8)", as 3(2) is where content proscribed under these guidelines is specified, and 3(8) is the mechanism for takedown of content.

The phrase "involving human intervention" is added.

The phrase "appropriate controls" is described in accordance with the three Santa Clara Principles on content moderation (https://www.santaclaraprinciples.org/), plus a requirement on intermediaries to present their algorithms for public scrutiny. An example of such a mechanism is New York city's algorithmic task force[6] which has been set up to examine automation systems derived from machine learning, data processing or AI techniques used to "make or assist in making decisions concerning rules, policies or actions implemented that impact the public."

6 https://www.theverge.com/2019/4/15/18309437/new-york-city-accountability-task-force-law-algorithm-transparency-automation