USER NOTIFICATION IN ONLINE CONTENT MODERATION

REF: 20-116 | November 24, 2020

User notification is a critical part of moderation of any type of online content and an essential element of due process. It may allow users to provide additional information or modify their upload or posting before any restrictive measure[1] is decided or implemented. It is in any case a precondition to reconsideration or recourse processes.

Building on the work of the Content & Jurisdiction Contact Group in 2019, the present document intends to reconcile the practical constraints regarding the timing of user notification with strengthening due process.

1. I&J Operational Norm and Criteria

The Operational Approaches[2] developed by the Content & Jurisdiction Contact Group in 2019 identified the importance of early user notification and laid out the following Operational Norm:

"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."

Regarding the content of notification, Criteria I of the Operational Approaches further clarified that:

"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."

However, the implementation of the above Operational Norm requires taking into account important constraints.

2. Implementation constraints

A considerable volume of potentially objectionable[3] posts must be detected, reviewed and decided upon expeditiously to ensure the timely prevention and remediation of online harm.

However, public authorities have a limited human and technical capacity to ensure detection of illegality at scale. Moreover, applying to each case the elaborate court procedures developed for traditional publishing would create a considerable burden on the judicial systems and introduce delays incompatible with the rapid and potential global propagation of illegal content. In this context, national (or regional) regulations increasingly impose upon operators the responsibility to address illegal content, with short action timelines under penalties.

The same volume and time constraints apply to service providers, compounded by the need to also address violations of their own increasingly detailed rules. While artificial intelligence (AI) tools increasingly supplement flagging systems for the detection of content to review, human intervention is necessary for decision-making.

Notifying users before a content restriction is decided and implemented naturally implies delays.

[1] See Criteria H: Choice of Action in Content & Jurisdiction Operational Approaches

[2] https://www.internetjurisdiction.net/uploads/pdfs/Papers/Content-Jurisdiction-Program-Operational-Approaches.pdf

[3] Although contentious posts represent only a very small proportion (less than 1%) of overall activity, hundreds of millions per year (for the largest operators) have to be reviewed for illegality or violation of Terms of Use.

It is thus frequently not practical when time is of the essence to address harm, or the illegality (respectively: violation of companies rules) is sufficiently evident on rapid review. Prior notification may not be permissible by law, and is inadvisable in case of ongoing or imminent real world harm.

Likewise, notification in the course of review is only meaningful if it allows the user to either modify the posting so that it does not infringe any more on the relevant normative basis, or provide relevant context or information useful for the evaluation.

An important distinction must therefore be made between:

- Content that is **manifestly illegal or clearly contrary to the terms of use** of the intermediary, and
- Content that **requires a more extensive evaluation** in light of the context, to establish the right balance between prevention of potential harm and protection of freedom of expression.

Indeed, some recent legislations (e.g. the German NetzDG) do recognize different response times on operators according to how manifest the illegality is.[4] Likewise, internet companies' internal escalation paths allow more time to handle non-manifest violations of their rules. This allows for a time-bound extensive evaluation by the company.

3. Timing of user notification

In light of the above, user notification can be meaningfully implemented by the company at different stages of the evaluation, under the following conditions:

At upload before posting, as a warning prompting[5] users to reconsider potentially harmful comments.	**As early as possible** if a significantly extended evaluation is warranted, in order to allow the user, within a limited time frame, to provide useful context information, to immediately modify the post, or to accept a specific restrictive measure.
On posting, if re-uploading of hashed[6] content is detected.	**Simultaneously with the action,** if the content is manifestly illegal or clearly contrary to the Terms of Use, or if the imminence and extent of the potential harm justifies rapid action.
When the evaluation process is completed, to open avenues for recourse.	

In the second situation, the notification should also indicate if the service provider is applying temporary measures to limit the distribution or virality of the content during the extended evaluation. The company can also separately solicit advice from a competent third body.

As notification is a precondition to reconsideration or recourse processes, in all cases it needs to contain, as indicated in Criteria I above, precise information regarding the available reconsideration and appeal mechanisms, and, in some jurisdictions, the possibility of recourse to a higher authority.

[4] Under Netz DG, companies must take down or block access to manifestly unlawful content within 24 hours of receiving a complaint. Other illegal content must be taken down or blocked within 7 days of receiving a complaint. Alternatively, social networks may refer the content concerned to a "recognized institution of regulated self-governance" (the FSM) with the understanding that they will accept the decision of that institution. The institution must then decide on whether the content is unlawful within 7 days. Social networks can exceed the 7day deadline when determining the illegality of the content depends on "the falsity of a factual allegation" or "other factual circumstances".

[5] See for instance Twitter's "Want to revise this?" and Instagram's "do you really want to post this?"

[6] Hash databases contain references to already evaluated content that has been previously determined as very dangerous or harmful (e.g. CSAM or terrorist content)