Risks of Filtering Requirements for Online Expression and Privacy

By: Corynne McSherry and Christoph Schmon, Electronic Frontier Foundation

Case:	2021Heonma290 Constitutional complaint regarding Article 22-5(2) of the Telecommunications Business Act, etc.
Claimants:	Mr. Kim and 3 others
Interested party:	Korea Internet Corporations Association

The Electronic Frontier Foundation ("EFF") is a member-supported, non-profit civil liberties organization that works to protect digital rights. For over 30 years, EFF has represented the public interest in ensuring that law and technology support human rights. In the US and abroad, we have worked to ensure that internet policy, legislation, and technological measures appropriately balance the rights of all internet users. As a legal services organization, we also counsel individuals and companies whose legitimate activities may be undermined by filtering proposals and requirements.

It is our understanding that the challenged provisions of the Telecommunications Business Act and the Enforcement Decree requires business to take technical measures to prevent circulation of "illegally filmed contents". In our experience, as a practical matter service providers will attempt to meet their legal obligations by reviewing all user content, often called "general monitoring", and using technical measures such as content filters. Below, we outline several principal concerns regarding the potential impact of such requirements on human rights, particularly online expression and privacy, as well as competition. To be clear, we are not experts in South Korean law nor the legislation at issue. We submit this declaration in the hopes that our experience with comparable regulations may be useful to the Court's determination in this case.

A. <u>General Monitoring, Filtering and Online Expression</u>

1. Risk of censorship due to expansive content moderation obligations

The increasingly powerful role of service providers in modern society has prompted a host of policy concerns. One key policy challenge is defining online intermediaries' legal liability for harms caused by content generated or shared by—or activities carried out by—their users or other third parties.

Unfortunately, laws that impose such liability inevitably result in the censorship of lawful and valuable expression. Stringent liability laws for online intermediaries encourage service providers to affirmatively monitor how users behave; filter and check users' content; and remove or locally filter anything that is controversial, objectionable, or potentially illegal to avoid legal responsibility. The effects are especially present where service providers obligations are unclear or broadly defined.

Faced with expansive and vague moderation obligations and major legal consequences if they guess wrong, companies inevitably overcensor. Stricter regulation of and moderation by platforms also results in self-censorship, as users try to avoid negative repercussions for their artistic and political expression. Numerous studies document that when people believe their communication is being monitored, they self-censor both their expression and the content they seek out and read.[1]

2. International experiences – the example of European internet legislation

EU governments confronted this problem in the negotiations over the EU's new internet bill— the Digital Services Act or DSA. The DSA seeks to articulate clear responsibilities for online platforms, with strong enforcement mechanisms, while also protecting users' fundamental rights. The DSA encourages "good Samaritan" content moderation and sets out type- and size-based due diligence obligations, which include obligations relating to transparency in content moderation practices, algorithmic curation, and notice and action procedures.

For instance, the DSA's transparency requirements mandate that users be informed about a platform's content moderation practices. Terms of service of online platforms must contain details about the utilization of automated decision-making processes and the extent of human oversight. In order to address illegal content online, all providers of hosting services, regardless of their size, must put in place notice and action mechanisms that facilitate the notification of potentially illegal content, after which platform providers can decide whether to take action with regard to the notified content. In all cases where content is removed, whether or not the removal decision was based on a notification or on platform's own investigations, users are entitled to know the rationale behind the decision and must be given options to appeal the decision. The DSA sets out several safeguards to ensure that content removal decisions are targeted and consider users' rights to freedom of expression and of information as well as their right to privacy and non-discrimination. Users also enjoy a right to reinstatement if platforms wrongly remove their content.

Several national bills in Europe were presented ahead of the DSA negotiations such as the controversial "Avia Bill" in France. The new law required social media intermediaries to remove obviously illegal content within short time frames and was met with criticism from experts and civil society. An intervention before the French Supreme Court, co-organized by the Electronic Frontier Foundation, proved ultimately successful, as the Court struck down the law's requirements to remove infringing content within 24 hours[2], recognizing that platforms would be encouraged to remove perfectly legal speech. Other national bills, such as the Austrian hate speech law also prompted question about 'overblocking'[3] of legitimate expression and non-compliance with overriding EU principles.

The draft versions of the Digital Services Act showed sympathy for a variety of legal solutions to address problems of online safety and the sharing of illegal content. The initial proposal

---

[1] *See generally* N. Richards, *Why Privacy Matters*, Oxford Press (2022)
[2] https://www.eff.org/press/releases/victory-french-high-court-rules-most-hate-speech-bill-would-undermine-free-expression.
[3] https://www.euractiv.com/section/data-protection/news/austrias-law-against-online-hate-speech-question-marks-in-the-home-stretch/.

suggested to let substantiated user notices suffice to trigger removal obligations for online platforms thus resulting in potential secondary liability for user content. Other versions suggested short deadlines for the removal of content. Ultimately, considering the negative experience with national initiatives and striving to achieve a fair balance between the various interests at stake, including freedom of expression rights by users, the final version abstained from any filter mandates and short inflexible deadlines for the removal of content.

The final legislative deal is widely considered being in line with the principles that underpin the e-Commerce Directive, the EU's previous backbone legislation for internet regulation. The DSA upheld the important principle that States should never mandate platforms to monitor user communication as this would inevitably lead to over-removal of content, undermine free speech rights and ignore users' right to privacy. Addressing the spread of illegal content online and the societal risks posed by the dissemination of disinformation or other harmful content, EU lawmakers opted for a harmonized notice and action system. Online platform providers must establish mechanisms that allow for the submission of detailed and substantiated notices. Upon receiving a notice from users or entities alerting them to the presence of illegal content, they are obligated to act. Platform providers must make a timely decision regarding the restriction or removal of such content and inform users if removal decisions are made. If notices contain detailed and precise information for diligent hosting service providers to identify the content as illegal without conducting a thorough legal examination, these providers will lose the DSA's liability exemption if they fail to act.

From this follows that under EU online platform rules, liability for speech continues to rest with the speaker and not with platforms that host what users post or share online. However, online platform providers are required to put in place processes that help tackle the dissemination of illegal content online.

More recent national bills that don't live up to this standard, such as the draft UK online safety bill, are widely attacked[4] as violating the fundamental rights of user.

3. Human Rights Principles: the Need for Check and Balances

Freedom of expression and online privacy are fundamental rights under the EU Fundamental Rights Charter, the European Human Rights Charter, and they are also protected under other instruments of international human rights law. Any State measure that aims to interfere with these rights for the sake of protecting another public value, such as the avoidance of harmful content online, must seek to strive a fair balance between these objectives.[5] As a general principle, States measures should have a clear legal basis and be necessary in a democratic society and proportionate, meaning that they cannot go beyond of what is necessary to achieve the objectives. They should also be sufficiently safeguarded.

As far as the core of the relevant contested provisions under the Korean Telecommunication Business Act is concerned, it is our understanding that there is a legitimate concern that not enough attention has been paid to fundamental rights aspects, including the right to freedom of

[4] https://www.eff.org/deeplinks/2022/08/uks-online-safety-bill-attacks-free-speech-and-encryption.
[5] ECHR, Perinçek v. Switzerland [GC], § 274; Case C-275/06 Promusicae [2008] ECR I-271.

expression. The bill sets out a positive obligation to monitor and restrict content if they match content reviewed and decided to be illegal by a specific body.

Without sufficient check and balances, such an obligation will likely lead to the indiscriminate general monitoring of all user content and the deployment of error-prone matching technology, such as automated filter systems that automatically prevent relevant content from being uploaded. Under human rights doctrines, any legal framework that provides for or demands blocking measures should ensure that measures strictly target the illegal content and has no arbitrary or excessive effects.[6]

4. Likelihood of Error

Technical measures to screen content are also prone to error, because no algorithm can replace or do the kind of contextual and legal analysis required to distinguish lawful from unlawful uses. This is one of the lessons of the EU Copyright Directive. EU lawmakers tried to mandate use of filters that can block infringement while also permitting lawful use. Everyone who testified in the implementation process, including filter vendors, said they could not do this. This means platforms will have to choose between over- or under-blocking.

Filters can also make purely technical errors, falsely identifying material as a duplicate of a protected work. Even the most expensively developed filters, like YouTube's ContentID, have problems like classical music recordings being falsely matched. Technical analysis of other filters has also found matching problems.

Even when filters perform as intended, they very frequently remove legal content because of human error, usually in the form of rightsholders claiming the wrong content. (Like claiming an entire nightly news broadcast or the film clip used in a movie review.) Many errors with YouTube's ContentID fall in this category.

B. General Monitoring, Filtering, and Privacy

General monitoring also undermines users privacy rights by requiring companies to collect abundant data about users, often without users' knowledge. International experiences show that any bill that requires online platforms to systematically filter user content create serious privacy and security risks. For example, in the realm of copyright enforcement mandated technical measures have shown to be inadequate to deal with risks of data breaches and in the worst case even compel platform operators to break encryption in order to scan the content of messages.[7]

These actions violate human rights standards. The UN High Commissioner for Human Rights has emphasized that any interference should be carried out only when authorized by an independent judicial body, on a case-by-case basis.[8]

---

[6] Cf. ECtHR, Application no. 10795/14 – *Kharitonov v Russia* (website blocking), at 45.
[7] https://www.eff.org/deeplinks/2022/11/filter-mandate-bill-privacy-and-security-mess.
[8] https://www.ohchr.org/en/press-releases/2022/09/spyware-and-surveillance-threats-privacy-and-human-rights-growing-un-report.

Finally, the EU's General Data Protection Regulation (GDPR) recognizes that filter system can have adverse effects on users, whose content is automatically removed or whose data are collected for the purpose of profiling.[9] Data protection principles thus require state measures to ensure safeguards for users' privacy, freedom of speech and other fundamental rights before any uploads are judged, blocked or removed.,

C. General Monitoring, Filtering and Competition

Monitoring requires intensive resource investments, which in turn discourages new companies from entering the field. Aside from the human resources require, technical measures such as upload filters are expensive to build and expensive to license. YouTube's ContentID had already cost that company over $100 million dollars as of five years ago. Audible Magic, which was promoted to EU lawmakers as an affordable solution, is widely reported to cost more in practice than that company represented.

Filters also reduce competition by creating technical lock-in. As Engstrom and Feamster report, a major consequence of adopting particular filtering technology is that companies design their other systems around that technology. Design choices, investments, or new lines of business can be precluded or rendered prohibitively costly as a result. It is one thing for companies to voluntarily put themselves in this position -- but it is quite another for it to be created by government mandate.

D. A Framework for Transparency and Redress

Adverse effects on human rights may be alleviated by a voluntary human rights framework for content moderation. Any decision about which content should or should not be shared online has serious human rights implications and online platforms can benefit from instructions to help operators to act more responsibly. EFF has long worked to provide guidance: In 2015, EFF, as part of an international coalition, helped launch the "Manila Principles on Internet Liability"[10], a framework of baseline safeguards and best practices based on international human rights instruments and other international legal frameworks. In 2018, EFF and partners then launched the "Santa Clara principles[11] on Transparency and Accountability in Content Moderation", which call on intermediaries to voluntarily adopt better practices. In 2021, a new version of the principles was developed, with a focus on adequately addressing fundamental inequities in platforms' due process and transparency practices for different communities and in different markets.

However, EFF recognizes that there's a need to strike a balance between addressing the very real issue of platforms hosting and amplifying illegal content while simultaneously providing enough protection to those platforms so that they are not incentivized to remove protected user speech, thus promoting freedom of expression. Our recommendations on how to best achieve this are guided by the rationale that it is in the best interest of all parties to focus on the regulation of processes on platforms rather than on speech and to make sure that mandatory content

[9] Article 22 GDPR.
[10] https://manilaprinciples.org/index.html.
[11] https://santaclaraprinciples.org/.

restrictions are always ordered by a judicial authority and are applied without resorting to intrusive filter systems.[12]

E. Conclusions

Intermediaries are vital pillars of internet architecture, and fundamental drivers of free speech, as they enable people to share content with audiences at an unprecedented scale. International experiences with internet legislation show the challenges for regulators and legislators to choose the right toolbox when addressing illegal content online. The EFF believes that the adoption of moderation frameworks that are consistent with human rights can best help to meet that challenge.

[12] https://www.eff.org/deeplinks/2022/05/platform-liability-trends-around-globe-conclusions-and-recommendations-moving.