How to fix the over-removal of legitimate content?

Submission to the House of Lords

Martin Husovec (LSE Law)

Context

Scholars have long argued that even the baseline system of intermediary liability — notice and takedown regime — is prone to over-removal of legitimate speech. Faced with potential liability, providers have a rational bias towards over-removal; they err on the side of caution; plus, they face no punishments for removing legitimate content. Affected users do not complain as they see little benefit of doing so. These phenomena have been proven true by rigorous empirical and experimental studies.[1] However, scholars have not been able to measure the aggregate magnitude of the problem that likely varies in different areas of content, such as defamation, hate speech and terrorist content.

When fishing for bad speech, regulators often incentivize providers to use the most inclusive nets, but when good speech gets stuck in the same nets, they provide the speakers only with a chance "to talk" to providers, thus giving them a small prospect of due process. With policies trying to impose more liability or accountability on intermediaries, the risk of over-removal of legitimate content is likely to only grow. When increasing the pressure on the detection and removal side (by prescribing automation, filters and other preventive tools that ought to be scalable), the policies cannot pretend to solve errors by entirely ex-post individual complaint mechanisms that can be overruled by platforms. In lockstep with new intermediary liability rules, there is a need for new speech safeguards.

Incentive approach

The failure is easy to name. The liability system fails to create equally strong incentives for providers to avoid over-removal at scale. Without symmetry in incentives, delegated enforcement by providers is no equal game; and without equality of weapons, there is no due process online. When the speakers (individuals or businesses) have to invest to counter false allegations, they bear all the cost, although they cannot scale up or speed up their defences. Without strong counter incentives, the cost of mistakes will be always borne by the speakers because any correction takes place only ex-post after a lengthy process. Even if somehow legitimate speakers prevail, the system, by definition, defies the legal maxim that justice delayed is justice denied. To correct this imbalance, the platforms need to be equally exposed to incentives to improve the quality of their review ex-ante.

Fiala & Husovec (2018) paper[2] shows one of the ways how to do it. It tests a simple design that follows such an incentive approach. It shows how to reduce the bias against over-removal and restore the symmetry of incentives. The research replicates the over-removal bias under the notice and takedown system in the laboratory experiment and then intervenes by exposing platforms to counter incentives in form of penalties for each wrongful removal. The logic is simple: if platforms bear the costs of their mistakes because over-removal suddenly also has a price tag, they have more incentive to improve by investing resources into the resolution of those errors (false positives). The system works as follows:

---

[1] See for the overview http://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws and for a discussion Fiala and Husovec (2018) p. 5 ff.

[2] Lenka Fiala and Martin Husovec, Using Experimental Evidence to Design Optimal Notice and Takedown Process (July 23, 2018). TILEC Discussion Paper No. 2018-028, Available at SSRN: https://ssrn.com/abstract=3218286

Each user of a platform can complain to an external dispute resolution body (ADR) if the platform fails to resolve his/her complaint. Upon paying a fee, the ADR decides the dispute. If it sides with the platform, the user loses his/her deposit (the deposit serves to filter out the cases). If it sides with the user, the fee is reimbursed, and a multiple of the fee is paid to the ADR body by the platform as a penalty (e.g., the user receives her £50 and the platform pays the ADR body double of the fee, £100). Since the decision of the ADR is binding on the platform, it must reinstate the content. The platform enjoys immunity for implementing it. The fees make ADR self-sustainable as a business. The ADR thus has a dual purpose. For the user, it introduces an effective due process tool as the decisions are binding on the platform. For the ecosystem, the fines in aggregate incentivize higher-quality reviews. Since platforms can learn at scale, each individual mistake is an opportunity for the benefit of *everyone* else, thereby improving the technology, internal review and associated governance processes.

The paper's results show that the introduction of the ADR under this configuration increases the complaints by users from 12% to 77%. The mechanism thus helps to solve the user's apathy created by a previous lack of credible complaint procedure. More importantly, the overall accuracy of decisions increases from 58% to 75%. This happens due to two reasons: (a) the platforms react to the threat of a potential ADR complaint more often by self-correction (14% to 57%) and because (b) the platforms must implement ADR decisions. With the compensated ADR system in place, over-compliance becomes as likely as under-compliance. The symmetry of incentives is therefore restored.

Policy options

There are at least two ways how to implement this ADR system: (a) through encouragement or (b) by a mandate. In the first case, the legislation only creates a post-ADR immunity *in case* a platform decides to designate an independent ADR that is subject to the outlined fee structure. In such a case, the platform voluntarily accepts the arrangement with an external provider of ADR (including potential penalties to be paid for lost cases) because it receives immunity for implemented ADR decisions. Such immunity might be attractive for the platform's users who gain a guarantee of real due process. The platform gains legal certainty. Those who objected to the content can still go to court if they are not convinced by the expert ADR decision. The risk is that platforms might designate no ADR body or one with very expensive fees. Platforms decide the forum.

In the second case, the legislation mandates ADR for certain platforms (e.g., based on size or type of service). In this case, the ADR is not a choice but a must. The fees can be set by the legislature or left to market within some limits. The latter is the approach taken by the European Commission that adopted the paper's approach in its Article 18 of the Digital Services Act proposal (platforms reimburse "any fees and other reasonable expenses"). The ADR bodies are subject to state certification. The benefit of such a system is that it allows ADR bodies to compete on price and quality (users decide the forum and the state controls the quality). The downside is the dispute about the reasonableness might arise and make any reimbursement less straightforward. Plus, without the multiplication of the fee, the platform might end up paying too little as a long term disincentive.

It might be advisable to possibly mix the two approaches. Use the mandated version of the compensated ADR for bigger platforms and the optional version for smaller platforms. To facilitate the complaints also in cases of low-economic value (but potentially high social value) disputes, it is also advisable to allow users to easily refer their cases (e.g., through a simple link) to charities or consumer bodies that could act on their behalf in the ADR process. The internal appeal within the platforms should also be regulated to better inform the affected users of variables, such as typical error rates of the platform and grounds for removal. The more useful information, the better.