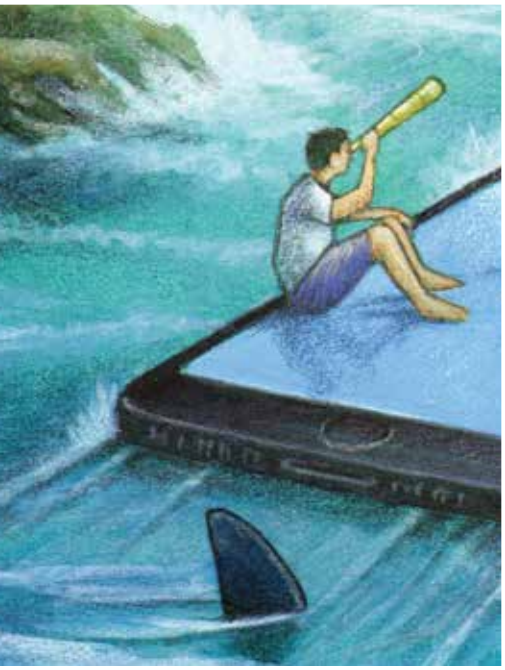
One in a Series of Working Papers from the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression



Design Principles for Intermediary Liability Laws

Joris van Hoboken
Vrije Universiteit Brussels and University of Amsterdam

Daphne Keller
Stanford Center for Internet and Society

October 8, 2019

TRANSATLANTIC WORKING GROUP

The Transatlantic Working Group Papers Series

Co-Chairs Reports

Co-Chairs Reports from TWG's Three Sessions: Ditchley Park, Santa Monica, and Bellagio.

Freedom of Expression and Intermediary Liability

Freedom of Expression: A Comparative Summary of United States and European Law
B. Heller & J. van Hoboken, May 3, 2019.

Design Principles for Intermediary Liability Laws
J. van Hoboken & D. Keller, October 8, 2019.

Existing Legislative Initiatives

An Analysis of Germany's NetzDG Law
H. Tworek & P. Leerssen, April 15, 2019.

The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications
J. van Hoboken, May 3, 2019.

Combating Terrorist-Related Content Through AI and Information Sharing
B. Heller, April 26, 2019.

The European Commission's Code of Conduct for Countering Illegal Hate Speech Online: An Analysis of Freedom of Expression Implications
B. Bukovská, May 7, 2019.

The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem
P.H. Chase, August 29, 2019.

A Cycle of Censorship: The UK White Paper on Online Harms and the Dangers of Regulating Disinformation
P. Pomerantsev, October 1, 2019.

U.S. Initiatives to Counter Harmful Speech and Disinformation on Social Media
A. Shahbaz, June 11, 2019.

ABC Framework to Address Disinformation

Actors, Behaviors, Content: A Disinformation ABC: Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses
C. François, September 20, 2019.

Transparency and Accountability Solutions

Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry
M. MacCarthy, February 12, 2020.

Dispute Resolution and Content Moderation: Fair, Accountable, Independent, Transparent, and Effective
H. Tworek, R. Ó Fathaigh, L. Bruggeman & C. Tenove, January 14, 2020.

Algorithms and Artificial Intelligence

An Examination of the Algorithmic Accountability Act of 2019
M. MacCarthy, October 24, 2019.

Artificial Intelligence, Content Moderation, and Freedom of Expression
E. Llansó, J. van Hoboken, P. Leerssen & J. Harambam, February 26, 2020.

www.annenbergpublicpolicycenter.org/twg

A project of the Annenberg Public Policy Center of the University of Pennsylvania in partnership with
The Annenberg Foundation Trust at Sunnylands and the Institute for Information Law of the University of Amsterdam,
with support from the Embassy of the Kingdom of the Netherlands in Washington, D.C.

Design Principles for Intermediary Liability Laws[†]

Joris van Hoboken, Vrije Universiteit Brussels and University of Amsterdam[1]
Daphne Keller, Stanford Center for Internet and Society[2]

October 8, 2019

Contents

I. Introduction

The goal of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (TWG) is "to identify and encourage adoption of scalable solutions to reduce hate speech, violent extremism and viral deception online, while protecting freedom of expression and a vibrant, global internet." This goal raises two central questions:

- what are optimal *policies* with respect to hate speech, violent extremism and viral deception in the private sector?
- what is the optimal *legal framework* to promote such policies in the private sector, while protecting freedom of expression online?

As explored in this discussion paper,[3] intermediary liability (IL) frameworks provide answers that are at the intersection of these two questions. They define platforms' legal responsibilities in moderating and managing content posted by internet users. Specific intermediary liability laws, such as the U.S. Communications Decency Act of 1996, Section 230 (CDA 230), and Articles 12-15 of the EU's e-

[†] One in a series: A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

Commerce Directive (ECD), were put in place in response to the rise of the internet in the 1990s. They provide internet services acting as intermediaries with so-called safe harbors from liability for the activities of third parties using their services. Such laws may, for example, define or restrict the liability that can be imposed on a social media service for defamatory comments of users or on broadband providers for giving access to a website that facilitates illegal file-sharing.

There were a number of reasons why intermediary liability laws were adopted at the time. During the 1990s, intermediary service providers became the targets of litigation for the behavior of users. The open nature of these services, whose providers typically would not exercise prior control over the contents of information and communication of users, raised complex questions about the allocation of legal responsibility for harmful and/or illegal behavior.[4] Intermediaries became litigation targets and such litigation about the precise responsibilities of different intermediaries under existing laws led to legal uncertainty. These disputes raised business risks for the nascent internet service industry, caused legal fragmentation among different countries or regions, and raised concerns about the proper balance between effective remedies for harm and the protection of freedom of expression.

In response, statutory intermediary liability laws were adopted that sought to balance three goals: preventing harm; protecting free expression and information access; and encouraging technical innovation and economic growth more generally.[5]

CDA 230 is by far the strongest safe harbor provision internationally because it immunizes online intermediaries unconditionally for the speech of others (outside of the area of intellectual property, sex trafficking, and federal criminal offenses). The strength of these protections against both liability and the considerable cost to platforms of even successful litigation is hard to overstate.[6] CDA 230 also immunizes online intermediaries for decisions to remove content, including lawful speech, from their services. CDA 230 thereby serves two parallel goals. First, to support the development of the internet ecosystem and freedom of expression by limiting risk and liability for relevant services acting as intermediaries. Second, to provide maximum space for such services to apply voluntary mechanisms to address potentially illegal and harmful content. Another important U.S. law, the Digital Millennium Copyright Act (DMCA), creates a detailed "notice-and-takedown" system for content alleged to infringe copyright.

In the EU, the ECD provides similar safe harbors for the activities of specific intermediaries, across a wider set of legal issues. The European safe harbors effectively create a notice-and-takedown system (or "notice and action," since an intermediary may respond in other ways besides taking content down) for content ranging from copyright infringement to hate speech. Rules vary from country to country, and are generally not spelled out in detailed statutes. Once an online intermediary obtains knowledge or awareness about illegal content, it loses immunity under the ECD and risks becoming liable. European courts have denied statutory immunities to intermediaries that were too "active" in engaging with user content – contrasting with CDA 230's encouragement of moderation and editorial control. Notably, the ECD leaves room for injunctions and duties of care at the national level with respect to unlawful content (including removal). Such injunctions are limited by Article 15, which prohibits national lawmakers from imposing general monitoring obligations on intermediaries for illegal content.

II. The Same Principles in Changed Circumstances

Clearly, the core principles underlying these frameworks, including the tackling of harm, protection of freedom of expression, and support for innovation, remain valid today. Still, the environment has

changed significantly since the adoption of these laws. First, intermediary service providers such as Google Search, YouTube, and Facebook may be considered dominant players and have been found to be unprepared to tackle emergent phenomena. This, and the broader political "techlash" we are witnessing today, undercuts one of the political rationales underlying the safe harbors: minimizing business risks to emerging companies and technologies. The safe harbors are now often portrayed as sweetheart deals for already dominant companies. The legal certainty they provide for (potential) new entrants and smaller service providers, however, remains important as a source of support for innovation and competition. Overall, it's easy to critique existing laws, but replacing them with something better will require considerable attention to the real mechanics of intermediary liability law, and to the doctrinal choices discussed in this essay.

There is a clear need to guard against oversimplifications in the discussion. For instance, simply scrapping the current protection from the law will not provide a clean slate for the determination of intermediary liability. It will leave internet users and online expression subject to a complicated, fragmented, and uncertain mix of traditional legal doctrines. While years of litigation might ultimately lead to reasonable and workable rules, the harm to both innovation and internet users' rights in the interim would be significant. Any regime that imposes liability on speech intermediaries should comply with constitutional and human rights safeguards. Intermediary liability laws' restrictions on core democratic freedoms such as freedom of communication, speech, and association, as well as the right to privacy, must be necessary, proportionate, and provided for by law.

The second important shift in the environment for intermediary liability laws involves platforms' role in society. Concerns about the impact of out-of-control online speech dynamics and challenges posed to our democracies abound. Generally, online platforms and associated technologies and practices have become catalysts in much wider economic, cultural and social change. But online platforms are also essential entities in the online ecosystem for freedom of expression, transforming the setting for the regulation of speech and harm into a triangle of platforms, users, and regulators. Considering these circumstances, a balanced answer to the questions about their proper roles and responsibilities with respect to societal impacts, including harmful ones, remains an essential legal and regulatory challenge.

In addressing this challenge, a warning is due with respect to a singular focus on the roles and responsibilities of online platforms and infrastructural services. Although online platforms are attractive targets for regulation, due both to their ability to exercise control as well as to market concentration, such regulation can have a number of clear downsides, such as privatized censorship. Intermediary liability frameworks should be assessed in light of their impact on both platforms and end users, and weighed against the option of laws targeting the primary speakers that are ultimately responsible for the publication of illegal and/or harmful content or activity.

As a result of the changes discussed above, thought leaders and policy makers on both sides of the Atlantic have started to question whether service providers, in particular large ones, should be expected and required to do more, and prevented from "hiding behind" first-generation internet regulations. In addition, political pressure is mounting on platforms to be more restrictive toward speech that is not necessarily illegal but is considered harmful, such as viral deception and certain forms of hate speech that are protected under the First Amendment to the U.S. Constitution, Article 10 of the European Convention on Human Rights, and Article 11 of the EU Charter of Fundamental Rights.

Thus, intermediary liability laws in the U.S. and Europe are a key issue for the TWG to consider and examine, as the debates about the possible revision of relevant statutory regimes are in need of robust guidance and well-informed recommendations. To provide input to the debates in Europe and the

United States and provide the basis for higher-level recommendations, the subsequent sections lay out the key components of intermediary liability laws, seen from a transatlantic perspective.

Specifically, we review the "dials and knobs" available to lawmakers seeking to update intermediary liability laws to account for present-day concerns. We break down key doctrines or provisions from existing law or current policy discussions into modular elements, focusing on their ramifications for free expression in particular. Of course, in real-world legislation, these elements rarely occur in isolation – and combining them can produce new effects. For example, a law holding platforms liable for deceptive speech they "know" about may mean something different depending on whether the law lets users explain and defend their posts. For the purposes of our discussion, however, isolating them can help in identifying options for well-designed intermediary liability laws.

III. Central Considerations

Platform liability laws affect internet users' free expression and access to information in two big ways. First, users' rights suffer if platforms are incentivized to "over-remove," taking down lawful speech in order to avoid liability or reduce costs. Over-removal in notice-and-takedown systems is well-documented.[7] Second, liability risks can deter innovators from building – or investors from funding – open speech platforms in the first place. As a result, strict liability standards for users' expression on platforms have generally been considered incompatible with freedom of expression. Legal and human rights literature identifies the following as particularly critical tools to mitigate threats to free expression:

No monitoring: Not requiring platforms to proactively filter or police user expression

Human rights literature includes strong warnings against making platforms proactively monitor, police, or filter their users' expression. Many intermediary liability laws expressly bar such requirements, though they have gained traction in recent European legislation such as the EU Copyright Directive for the Digital Single Market and some drafts of the Terrorist Content Regulation. One concern is that technical filters, which may range from simple hash-based systems for recognizing duplicate files to more sophisticated AI-based processes, are likely to over-remove because they are bad at recognizing context – like news reporting or parody – or accommodating changing legal interpretations. (Filtering, though, is relatively accepted for child sexual abuse images, which are unlawful in every context.) Another is that when platforms have to review and face over-removal incentives for every word users post, the number of unnecessary takedowns can be expected to rise. Under a law that requires monitoring, legal exposure and enforcement costs may also give platforms reason to allow only approved, pre-screened speakers, or to use Terms of Service (TOS) to prohibit controversial or legal gray-area speech. The liability risk and enforcement costs may also deter new market entrants from challenging incumbents.

Public due process: Using courts or other public authorities, not platforms, to decide what expression is illegal in most cases

As a protection for internet user expression rights, some countries reserve the responsibility for assessing certain claims against online content to courts or other government authorities. Platforms are immune until informed of the authority's legal determination that specific content is illegal. The government may also be subject to transparency obligations when it requires or suggests removal of content. This speech-protective standard typically has exceptions, requiring platforms to act of their own volition against highly recognizable and dangerous content such as child sexual abuse images. Lawmakers who want to move the dial toward harm prevention without having platforms adjudicate

questions of speech law can also create accelerated administrative or court processes or give platforms other responsibilities, such as educating users, developing streamlined tools, or providing information to authorities. Judicial review is particularly important and valuable for borderline cases involving disputed facts or nuanced legal doctrine.

Private due process: Requiring procedural protections for speakers when platforms take action against content

Building procedural protections into platforms' internal notice-and-takedown systems and terms of service enforcement can protect against over-removal. A widely supported civil society document, the Manila Principles, provides a menu of procedural protections with respect to notice and action. For example, a platform can be required or incentivized to notify the affected speaker, provide sufficient reasoning for its actions affecting her speech, and let her defend herself. The existence of such procedures may deter bad-faith notices in the first place. Claimants or accusers can also be required to include adequate information in notices and face penalties for bad-faith allegations. And platforms can be required to disclose raw or aggregate data about actions against content to facilitate public review and correction. Procedural protections for users affected by TOS enforcement – i.e., "private due process" – may also be required as a matter of consumer contract law. Self-regulation initiatives, which may have the partial aim to encourage reliance on TOS and prevent actual regulation from being passed, should come with robust private due process safeguards.

Public rule-setting: Regulating platforms' use of private Terms of Service enforcement

Platforms often take down disfavored but legal speech based on their TOS or Community Guidelines. To protect users' free expression rights and prevent undue bias in content moderation practices, a law might try to impose limitations on TOS enforcement against protected expression (possibly in combination with requiring the private due process discussed in the previous section). To ensure that governments do not fail in their own human rights obligations, they might also be prevented from relying on companies' TOS instead of publicly enacted law to regulate speech. However, limiting TOS enforcement would have some clear downsides from the perspective of tackling illegal and harmful activity. Rules designed to protect users' rights to expression, information, and due process may need to be balanced with Good Samaritan defenses (discussed below), which are designed to encourage platforms to moderate lawful but harmful content. TOS enforcement may also effectively serve to protect the free expression rights of vulnerable users. For example, reducing legal but abusive comments on platforms like Twitter can, as a practical matter, enable attacked or marginalized users to speak more freely. It can also make the platform attractive to other users, preserving its value as a forum for civil discourse. These arguments are particularly salient in countries like the U.S., where the law permits speech that violates widely held social norms or moral beliefs. Finally, in the U.S., the law also likely protects TOS enforcement as an exercise of platforms' own editorial rights. In Europe, too, freedom of expression and media pluralism warrant care in imposing community standards on platforms instead of allowing services to choose their standards freely within the boundaries of the law.

Remedies for speakers: Equal access to courts for speakers and victims of harmful content

Platform over-removal incentives come in part from asymmetry between the legal rights of accusers and those of speakers. Under most intermediary liability systems, including Europe's, victims of speech-based harms can sue platforms to get content taken down. Speakers, by contrast, can very rarely sue to get content reinstated or be compensated. (To our knowledge, such claims have succeeded only in Germany, Poland, and Brazil, and only very recently.[8]) This means that outside of

strict immunity regimes like the U.S.'s CDA 230, liability concerns consistently push toward removal. This asymmetry also distorts courts' opportunities to clarify the law, particularly when platforms enforce novel legal standards by allowing courts to review the claims of people seeking more content removal, but not the claims of people defending their expression rights. A few untested new laws in Europe, including the Audio Visual Media Services Directive and Copyright Directive, try to remedy this.[9] It is unclear how these new mechanisms will work in practice or how speakers' claims will intersect with platforms' power to take down speech using their TOS.

Consistent speech laws: Protecting the same expression online and offline

Content that might do only modest harm offline – like political disinformation spread by word-of-mouth to a few people – may do greater harm online, where it persists over time and can spread virally. Some lawmakers have responded to this concern by pressuring platforms to prohibit harmful-but-legal content voluntarily under their terms of service.[10] This approach reduces public rule making and due process for sensitive free expression issues. Others have proposed or enacted laws restricting online dissemination of speech that is otherwise legal.[11] This approach – which has long been strongly disfavored in human rights literature – resembles regulation of older media like broadcast or cable, which, for instance, have rules to protect minors. Applying such rules to internet platforms would put a greater burden on free expression rights, though, because it would affect everyday speech by internet users who rely on platforms to communicate.

Legal predictability: Bright-line rules versus fuzzy standards

Intermediary liability rules can hold platforms to flexible standards like "reasonableness" in responding to potentially unlawful user content, or prescribe specific steps. Both platforms and free expression advocates typically favor the latter because it increases predictability and reduces the role of platform judgment. Poorly calibrated process rules may encourage over-removal – if, for example, platforms automatically honor all takedown demands -- but this can be somewhat offset with private due process requirements, like counter-notice or transparency.

Private vs. public speech: Respecting communications privacy and targeting public (illegal and harmful) speech

Appropriate IL rules may be different for fully public communications (like a blog post or tweet) as compared with private communications (like email or a post to a small closed Facebook group). Existing legal frameworks for communications services include protections for communications privacy, with some of these protections also applying to new services (beyond traditional telecommunications). Internet services have also increasingly integrated technical protections, including end-to-end encryption in services such as WhatsApp, Signal, and Telegram. As internet users migrate toward these private platforms (potentially as a result of content moderation practices), there is increasing pressure on service providers to police private communications or even build encryption backdoors. IL laws targeting private communications services should only do so while respecting communications privacy and security. The distinction between private and public speech is not a sharp one, and appropriate rules may vary depending on the type of illegal or harmful content. (In some cases, disseminating content may be legal in a private communication but not in a public one.)

Cross-border dimension: Respecting the global nature of internet speech and platforms

Across jurisdictions, IL laws or underlying law on issues such as hate speech or disinformation may set different standards that can create cross-border conflicts since global platforms are subject to legal pressure from around the world. For instance, one country may set a Good Samaritan defense

permitting platforms to remove lawful but harmful speech, while another country may limit a company's freedom to do so. Or a court in one jurisdiction may order the global removal of speech that is legal elsewhere. Intermediary liability laws should be respectful of the global nature of internet speech platforms and minimize cross-border conflicts limiting freedom of expression. Jurisdictions that want to increase enforcement of their local laws by tightening IL standards (like NetzDG) generally should not require the entire worldwide platform to operate according to their standards.

IV. Standards for Platforms' General Conduct

Some recently proposed standards focus on platforms' responsibility in their overall operations, rather than on case-by-case liability for individual items of unlawful content.

"Due diligence" or "duty of care" standards

At their core, IL laws are concerned with the question of what liability exists for *specific instances* of illegal content or activity. IL laws have established that such liability should be limited in nature (e.g., only after the service has actual knowledge) and not strict (liability imposed regardless of knowledge, establishing de facto proactive duties to monitor to prevent liability). Considering the current direction of discussions about IL, the question is whether and what policy options exist between those conditions. Some recent European laws and proposals move away from penalizing platforms for individual incorrect decisions about specific user expression, and instead seek to regulate platforms' overall content management operations and create new forms of administrative oversight with respect to these frameworks.[12] This approach might crudely be analogized to food safety standards that accept a certain number of insect or other contaminant parts-per-million, on the basis that requiring a smaller margin of error would impose disproportionate costs on both the regulated entity and society. For instance, under this model, a platform that meets a "duty of care" or "due diligence" standard in its overall content moderation system would not be punished for one-off mistakes. One can also imagine more specific targets for content moderation practices. In countries willing to accept significant regulatory review and standard setting for platforms, these approaches may represent an important new way forward. They could build on existing approaches with respect to risk management and fundamental rights impact assessment, requiring platforms to consider risks to freedom of expression, due process, non-discrimination and minority participation in public discourse.

Transparency requirements

Transparency reporting has emerged as a practice to create more accountability for removal of content by platforms. Industry transparency reporting practices have developed over the last decade, with reports providing insights into the number of requests to take down allegedly unlawful content in different categories. Transparency reporting is required in some new intermediary liability laws, such as Germany's NetzDG and the EU's proposal for a Terrorist Content Regulation, and co-regulatory frameworks for hate speech and disinformation. The reports have become important sources of evidence, but it remains difficult to compare different platforms' reports because of differences in their reporting procedures and standards.

V. Liability Based on Knowledge or Control

Traditional tort doctrines typically held publishers or distributors liable based on their editorial control or knowledge about unlawful content. Similar standards appear in many IL laws, although platforms

often differ from pre-internet publishers or distributors in the volume of third-party expression they handle and in their relatively weak incentives to defend it.

Knowledge and other "mental state" standards

Many legal systems hold platforms liable for continuing to host or transmit illegal content once they "know" or "should know" about it. This is for instance the case under the European intermediary liability framework, Article 14 of the e-Commerce Directive specifically. Similar standards exist in U.S. criminal law and copyright law.[13] Others reject this standard, considering it too likely to incentivize over-removal. Laws that use knowledge standards can reduce this problem somewhat by defining "knowledge" narrowly or adding elements like private due process.

Controlling or "active" platform standards

Most IL laws strip immunity from platforms that are too actively involved in user content, e.g., because they help create or solicit particular material, or optimize, select and/or promote it in a commercial context or for profit-making purposes.[14] Some version of this rule is necessary to distinguish platforms from content creators. But laws that reward passivity also generate legal uncertainty – and may deter innovation or lead to over-removal – as platforms consider new features that go beyond bare-bones hosting or transmission. They also risk deterring platforms from moderating content at all, for fear of losing immunities.

"Good Samaritan" rules to encourage moderation

Platforms that want (for economic or other reasons) to weed out illegal or offensive content may be deterred by both "knowledge" and "control" liability standards. Plaintiffs can use platforms' moderation efforts as evidence of editorial control, or argue that the platform knew about content that a moderator saw but did not take down. This concern underlies the broad immunities the U.S. established in CDA 230. The current European framework lacks a Good Samaritan defense and platforms also risk losing their safe harbor protection if they more proactively address illegal and harmful content. This has complicated the development of self- and co-regulation to tackle illegal and harmful content online.

VI. Using Different Rules for Different Problems

Real-world laws typically combine elements listed here, which allows lawmakers to more carefully calibrate trade-offs affecting free expression. The potential downside is that complex laws can increase operational costs for platforms, potentially leading them to simplify by being too restrictive.

Variations based on legal claim

IL laws often require special or more urgent treatment for particularly harmful or highly recognizable content, such as child sexual abuse material. By contrast, they may provide stronger free expression protections for claims that platforms cannot reasonably assess because of nuanced legal doctrines or disputed facts, such as in the case of defamation.

Variations based on a platform's technical function and relation to user expression

Many IL laws put the risk of liability on the entities most capable of carrying out targeted removals – like taking down a single comment instead of a whole page or website. This is also consistent with the internet's "end to end" technical design principles. Thus, infrastructure providers like ISPs or domain registries generally have stronger legal immunities than consumer-facing platforms like YouTube.

Many recently proposed IL laws, like the 2018 amendments to CDA 230 in the U.S., have not reflected this principle.[15]

Variations based on a platform's size

Recently, experts have raised the possibility of special obligations for mega-platforms like Google or Facebook. Drafting such provisions without distorting market incentives, driving bad actors to less strictly policed, smaller platforms or punishing unusual platforms like Wikipedia would be challenging. In principle, though, it might improve protections on the most popular forums for online expression without imposing such onerous requirements that smaller market entrants couldn't compete.

VII. Judicial Actions Against Platforms

Platforms' actions against user-generated content can be shaped both by *direct* legal mandates (like injunctions) or the *indirect* influence of potential future claims. In deciding whether to remove legal gray-area content like crude parodies, for example, platforms may act based on their expectations or fears of what a court *might* do if the content stays up and a plaintiff or prosecutor brings a legal claim. Thoughtfully tailoring the availability of financial damages or injunctive relief in the platform context can help protect lawful expression.

Cost to platforms

Over-removal incentives (or incentives to stop offering services completely) are likely to be greatest when platforms fear high damages, regulatory attention that can lead to other costs or business impact, or business-altering injunctions (like having to turn off popular features).

Scope of injunctions

Because countries vary in their laws regarding expression, a platform takedown order issued in one country can affect speech and information that is legal in another. Geographically limited orders can mitigate this problem, but mean that harms may be addressed less effectively. Courts can also issue time-limited orders, allowing content to become accessible again after a certain period.

Options other than taking down or reinstating content

Increasingly, content-related issues in online platforms are dealt with through measures short of simply removing the material. For example, controversial but lawful content may be demoted in rankings, demonetized, or delisted in search results. IL laws can replace binary take-down/leave-up outcomes by stipulating more tailored remedies. For example, platforms can show users a warning before viewing certain content, cut it off from ad revenue, or show it in response to some search queries but not others. In principle, IL law could also regulate the algorithms that platforms use to rank, recommend, or otherwise amplify or suppress user content. Such a law, however, would be complex to define, enforce, and administer.

VIII. Conclusions

There are several structural reasons for revisiting intermediary liability laws that were adopted in the 1990s. When doing so, lawmakers should continue to be informed by the principles underlying these laws, including freedom of expression, as well as more than two decades of experience in providing for intermediary liability provisions and associated policies. Simply scrapping existing safe harbor provisions currently in place would in no way resolve many of the issues outlined here, and would

inevitably cause significant legal uncertainty and harm to competition and fundamental rights of internet users. In view of this, this discussion paper offers a systematic overview of key elements to consider in potential revisions and design of new IL laws and ways in which these elements can be approached in a balanced manner.

Lawmakers considering potential revisions to IL laws should craft any amendments carefully to avoid incentivizing platforms to act against the rights and interests of their users. The concerns and doctrinal tools listed under Central Considerations in section III are particularly key for this purpose, and should serve as guiding parameters. For example, under laws requiring platforms to remove unlawful content, "private due process" protections such as notice and appeals for the affected users can serve to protect expression and information rights. Lawmakers can further refine IL laws using transparency requirements, legal obligations tailored to intermediaries' technical functions, and other doctrinal tools discussed in sections IV-VII. By adjusting the "knobs and dials" set forth in this discussion paper, lawmakers can strike an appropriate and proportionate balance between reducing online harms, protecting fundamental rights, and promoting innovation and competition.

Notes

[1] Professor of Law, appointed to the Chair "Fundamental Rights and the Digital Transformation," established at the Interdisciplinary Research Group on Law Science Technology & Society (LSTS), Vrije Universiteit Brussels (VUB), with the support of Microsoft; Senior Researcher, Institute for Information Law (IViR), Faculty of Law, University of Amsterdam.

[2] Director of Intermediary Liability, Stanford Center for Internet and Society (CIS); former Associate General Counsel, Google. Stanford CIS funding information is available at http://cyberlaw.stanford.edu/about-us.

[3] An earlier version of this TWG discussion paper was prepared for the Santa Monica meeting of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression in May 2019. Sections III-VII are adapted and expanded from Daphne Keller, *Build Your Own Intermediary Liability Law: A Kit for Policy Wonks of All Ages* (June 2019), https://balkin.blogspot.com/2019/06/build-your-own-intermediary-liability.html?m=1. We would like to thank Sherwin Siy and Jan Gerlach of the Wikimedia Foundation and João Pedro Quintais of the University of Amsterdam for their input in drafting the initial document. Results of the discussions at the Santa Monica meeting have been incorporated into the text.

[4] The distinction between *illegal* content or activity and *harmful* content or activity is important. Harmful content and behavior includes speech and activity that the law permits but that still cause harm.

[5] Intermediary liability laws can also raise questions regarding other fundamental rights (including due process, communications freedom and confidentiality, data privacy, freedom of association and non-discrimination) and policy goals (e.g., competition, trade policy).

[6] Liability in the narrow sense can be about civil liability for damages caused by or criminal liability for third-party content and activity. In addition, there is an important question about the possibility to impose injunctions, for instance in a possible situation of negligence.

[7] Listing of empirical studies documenting platform removal of lawful speech under notice-and-takedown systems (last updated 2018): http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws.

[8] There may be many more cases in which an initial threat of a lawsuit by a speaker affected by a removal gets resolved by the platform reinstating the content.

[9] The AVMSD does so by giving users recourse to administrative review when platforms remove allegedly unlawful content.

[10] In Europe, for instance, lawmakers are discussing referral mechanisms in combination with calls on service providers to ban terrorism content through their terms of service. For a discussion, see Van Hoboken, "The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications," TWG discussion paper, Institute for Information Law, Amsterdam, 3 May 2019. Available at https://www.ivir.nl/publicaties/download/TERREG_FoE-ANALYSIS.pdf.

[11] See for instance the UK's Online Harms White Paper. Available at https://www.gov.uk/government/consultations/online-harms-white-paper.

[12] See in particular the NetzDG law in Germany. For a discussion of NetzDG, see Tworek and Leerssen, "An Analysis of Germany's NetzDG Law," TWG discussion paper, 15 April 2019. Available at https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf.

[13] 18 U.S.C. 2252, 2258A, 2258B (knowledge-based liability and obligations for intermediaries regarding child sexual abuse material); 17 U.S.C. 512 (intermediaries lose DMCA immunity based on actual or "red flag" knowledge).

[14] This is the case for the e-Commerce Directive in the European Union. U.S. intermediaries lose CDA 230 immunity if they materially contribute to the unlawfulness of content, *Fair Housing Council of San Fernando Valley v. Roommates.com, LLC,* 521 F. 3d 1157, 1168 (9th Cir. 2008), and lose DMCA immunity based on right and ability to control infringing content from which they directly benefit. 17 U.S.C. 512.

[15] The Allow States and Victims to Fight Online Sex Trafficking Act, commonly known as FOSTA, which was enacted in 2018, draws no distinctions between obligations of infrastructure providers and those of edge-of-network, user-facing platforms. H.R. 1865, 115th Cong. (2018).