Harvard Journal of Law & Technology
Volume 34, Digest Spring 2021

PLATFORM RESPONSIBILITY AND REGULATION IN CANADA: CONSIDERATIONS ON TRANSPARENCY, LEGISLATIVE CLARITY, AND DESIGN

*Sonja Solomun, Maryna Polataiko, Helen A. Hayes**

I. INTRODUCTION

Platforms and online intermediaries play a crucial role in mediating online discourse, which necessarily involves shaping what freedom of expression means on the Internet.[1] In light of the recent policy focus on mis- and dis-information, polarization, and harmful/hateful speech on the Internet—and prompted by calls from

*Sonja Solomun is the Research Director at the Centre for Media, Technology, and Democracy at the Max Bell School of Public Policy, McGill University, and a PhD Candidate at McGill University. Maryna Polataiko, B.C.L./LL.B., is a lawyer and was a Legal Fellow at the Centre for Media, Technology, and Democracy at the Max Bell School of Public Policy, McGill University. Helen A. Hayes is a Policy Fellow at the Centre for Media, Technology, and Democracy at the Max Bell School of Public Policy, McGill University, and a PhD Student at McGill University. Special thanks to Dr. Taylor Owen and James A. Hayes, J.D., for helpful comments and insights.

[1] Lex Gill, *Legal Aspects of Hate Speech in Canada*, CTR. FOR MEDIA, TECH., AND DEMOCRACY (2020), https://www.mediatechdemocracy.com/work/legal-aspects-of-hate-speech-in-canada; *see also* Jack Balkin, *The Future of Free Expression in a Digital Age*, 36 PEPP. L. REV. 428 (2009), https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1222&context=fss_papers; *see also* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2937985.

Canadians for more robust regulation of social media companies[2]—the federal government plans to introduce legislation imposing obligations on Internet platforms to remove unlawful speech. In doing so, policymakers and legal experts have had to interrogate complex policy questions about how best to regulate speech without infringing on guaranteed rights and freedoms and due process. Without weighing in on the merits of take-down legislation, this note outlines two broad considerations that may help protect expressive freedom and due process in the context of online speech regulation: 1) transparency reporting regarding content removals; 2) suggestions on drafting and legislative design, namely a) clearly defined categories of speech and notice procedures; and b) clear notice and counter-notice requirements for complainants and respondents.

II. CANADA'S INCOMING LEGISLATION

Following direction from the Prime Minister of Canada "to take action on combating hate groups and online hate and harassment, ideologically motivated violent extremism and terrorist organizations," the Minister of Canadian Heritage recently announced a plan to table legislation addressing harmful online speech.[3] According to the Heritage Minister, the incoming bill has been a joint initiative with ministers including the Minister of Public Safety and Emergency Preparedness and the Minister of Innovation, Science and Industry.[4] Together, they have assessed online speech regulation from other jurisdictions and have met with representatives therefrom in an effort to develop a uniquely "Canadian approach" to online speech regulation.[5]

The Heritage Minister explained that the new bill will outline a regulatory framework applying to online platforms that

---

[2] Canadian Commission on Democratic Expression, *Harms Reduction: A Six-Step Program to Protect Democratic Expression Online*, PUB. POL'Y F. (2021), https://ppforum.ca/wp-content/uploads/2021/01/CanadianCommissionOnDemocraticExpression-PPF-JAN2021-EN.pdf

[3] Standing Committee on Canadian Heritage, *Meeting No. 12 CHPC*, House of Commons (2021), https://parlvu.parl.gc.ca/Harmony/en/PowerBrowser/PowerBrowserV2/20210129/-1/34603?Language=English&Stream=Video#info_.

[4] *Id.*

[5] *Id.*

includes the establishment of a regulator[6] meant to oversee platforms' management of unlawful online speech.[7] The regulator will have the authority to impose financial penalties on platforms for failure to comply.[8] Under the new legislation, the definition of "speech" is said to include hate speech, child pornography, incitement of violence, incitement of terrorism, and non-consensual sharing of sexual images.[9] The bill will define these categories in detail.[10]

III. CONSIDERATIONS

A. Transparency in Content Moderation

Platform companies and digital intermediaries routinely screen and make decisions about the content being shared on and with their products and services. This practice of "content moderation"—understood as "the detection of, assessment of, and interventions taken on content or behavior deemed unacceptable by platforms or other information intermediaries, including the rules they impose, the human labor and technologies required, and the institutional mechanisms of adjudication, enforcement, and appeal that support it"[11]—has increasingly become a prominent priority for

---

[6] Note that the UK's Online Harms White Paper establishes Ofcom as an "online harms regulator" to produce an annual report detailing online harms and accounts before Parliament. For more information, see *Online Harms White Paper: Full Government Response to the Consultation*, DEP'T FOR DIGI.L, CULTURE, MEDIA & SPORT (2020), https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response#part-3-the-regulator.

[7] Standing Committee on Canadian Heritage, *Meeting No. 12 CHPC*, HOUSE OF COMMONS (2021), https://parlvu.parl.gc.ca/Harmony/en/PowerBrowser/PowerBrowserV2/20210129/-1/34603?Language=English&Stream=Video#info_.

[8] *Id*.

[9] *Id*; *see also* Elizabeth Thompson, *Canada Not Exempt from Social Media Forces that Created U.S. Capitol Riot, Heritage Minister Says*, CBC NEWS (2021), https://www.cbc.ca/news/politics/facebook-twitter-canada-regulation-1.5894301.

[10] Standing Committee on Canadian Heritage, *supra* note 7.

[11] Tarleton Gillespie and Patricia Aufderheide, *Introduction* in *Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates*, 9 INTERNET POL'Y REV. 4 (2020), https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy#footnoteref2_apdze36.

law and policymakers around the world. Although a popular policy focus within the recently coalesced but burgeoning field of platform governance,[12] forms of content moderation have been in existence since the conception of online communication[13] and pre-date digital communication.[14] The advent of commercial, rather than community-based, moderation, however, was spurred by the sheer scale at which online platforms have come to operate.[15] As such, regulatory considerations about content moderation now encompass both moderators—human or AI-powered—that have the ability to take down and demote content or de-platform (suspend or ban) users, and the corporate/commercial decisions that organize how such content moderation is structured.

Given the increasing power that platform companies hold in determining the contours of public discourse and online expression,[16] and the opacity of both the technological[17] and human processes and decisions behind them[18] (including in the ad-hoc way

---

[12] Robert Gorwa, *What is Platform Governance?*, 22 INFO. COMMC'N & SOC'Y 6 (2018), https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1573914?journalCode=rics20; *see also* Taylor Owen, *The Case for Platform Governance*, CTR. FOR INT'L GOVERNANCE INNOVATION, Paper No. 231 (2019), https://www.cigionline.org/publications/case-platform-governance, *see also* Taylor Owen et al., *Models for Platform Governance*, CTR. FOR INT'L GOVERNANCE INNOVATION (2019), https://www.cigionline.org/publications/models-platform-governanc; Taylor Owen, *The Case for Platform Governance*, CTR. FOR MEDIA, TECH., AND DEMOCRACY (2020), https://www.mediatechdemocracy.com/work/case-for-platform-governance.

[13] Sarah T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*, YALE UNIVERSITY PRESS (2019); *see also* Mary L. Gray and Siddarth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*, HOUGHTON MIFFLIN HARCOURT (2019).

[14] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*, YALE UNIVERSITY PRESS (2018).

[15] Cliff Lampe and Paul Resnick, *Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space*, PROC. OF ACM COMPU. HUM. INTERACTION CONF. (2004), http://www.presnick.people.si.umich.edu/papers/chi04/LampeResnick.pdf.

[16] Kate Klonick, *supra* note 1.

[17] Jenne Burell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y (2016), https://journals.sagepub.com/doi/pdf/10.1177/2053951715622512.

[18] Suzor et al., *Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda*, 80 INT'L COMMC'N GAZETTE 4 (2018), https://journals.sagepub.com/doi/abs/10.1177/1748048518757142, *see also* Robyn Caplan, *Content or context moderation? Artisanal, community-*

that they are inconsistently enforced),[19] it is perhaps unsurprising that demands for increased disclosure about content moderation have steadily increased.[20] While companies have begun to provide greater transparency around their content policy in recent efforts to regain public and government trust,[21] a breadth of research and independent audits have revealed that significant problems remain.[22]

i. Transparency Reporting

Transparency reporting has become a key mechanism of various regulatory and accountability frameworks, especially as it provides realistic expectations about platform capabilities on which

reliant, and industrial approaches, Data & Society (2019), https://datasociety.net/library/content-or-context-moderation/, *see also* Elinor Carmi, *The Hidden listeners: Regulating the Line from Telephone Operators to Content Moderator*s, 13 INT'L J. COMMC'N 440 (2019), https://ijoc.org/index.php/ijoc/arti cle/view/8588/0; *see also* Sarah T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*, YALE UNIVERSITY PRESS (2019).

[19] Bridget Barrett et al., *Enforcers of Truth: Social Media Platforms and Misinformation*, UNC CTR. FOR INFO., TECH., AND PUB. LIFE (2020), https://citapdigitalpolitics.com/wp-content/uploads/2020/05/Enforcers-of-Truth-CITAP-Reporot.pdf.

[20] Suzor et al., *What Do We Mean When We Talk About Transparency? Towards Meaningful Transparency in Commercial Content Moderation*, 13 INT'L J. COMMC'N 1532 (2019), https://ijoc.org/index.php/ijoc/article/view/9736.

[21] Robert Gorwa and Timothy Garton Ash, *Democratic Transparency in the Platform Society*, *in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD AND PROSPECTS FOR REFORM, CAMBRIDGE UNIVERSITY PRESS 286 (2020), https://www.cambridge.org/core/services/aop-cambridge-core/content/view/E79E2BBF03C18C3A56A5CC393698F117/9781108835558AR.pdf/Social_Media_and_Democracy.pdf?event-type=FTLA#%FE%FF%00B%00K%00C%00N%00-%00b%00p%00-%001%002.

[22] House of Lords Select Committee on Communications, *The Internet: To Regulate or Not to Regulate?*, Article 19 (2018), https://www.article19.org/wp-content/uploads/2018/06/HOL-inquiry-Internet-regulation-A19-written-evidence-18-May-2018-.pdf; *see also* Tarleton Gillespie et al., *Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates*, 9 INTERNET POL'Y REV. 4 (2020), https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy#footnoteref2_apdze36.

law and policymakers rely.[23] Most calls for robust transparency[24] reflect the Manila Principles on Intermediary Liability[25] and the Santa Clara Principles on Transparency and Accountability in Content Moderation.[26] The latter established a baseline for due process and transparency regarding individual notice, including a call for regularly aggregated information on such notices. Jointly declared by academics and civil society groups, the Santa Clara principles advise inclusion of "the total numbers of posts and accounts flagged or reported and the proportion of content removed or accounts suspended"[27] in order to understand the patterns and impact of content moderation "not only to the individual user, but also to the broader community." [28] Other scholars have noted that this may inform an integral "public right to hear" that may include, under democratic governance, a right to encounter diverse perspectives alongside an individual right to speak.[29]

While the Santa Clara Principles outline provisions for commercial content moderation, which likely require revision to

---

[23] *Democratic Transparency in the Platform Society*, *supra* note 21; *see also* Daphne Keller and Paddy Leerssen, *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation, in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD AND PROSPECTS FOR REFORM, CAMBRIDGE UNIVERSITY PRESS (2020), https://www.cambridge.org/core/services/aop-cambridge-core/content/view/E79E2BBF03C18C3A56A5CC393698F117/9781108835558AR.pdf/Social_Media_and_Democracy.pdf?event-type=FTLA#%FE%FF%00B%00K%00C%00N%00-%00b%00p%00-%001%002.

[24] House of Lords Select Committee on Communications, *supra* note 22; *see also* Jillian C. York and Corynne McSherry, *Automated Moderation Mst be Temporary, Transparent, and Easily Appealable*, ELEC. FRONTIER FOUND. (2020), https://www.eff.org/deeplinks/2020/04/automated-moderation-must-be-temporary-transparent-and-easily-appealable.

[25] *The Manila Principles on Intermediary Liability Background Paper*, ELEC. FRONTIER FOUND. (2015), https://www.manilaprinciples.org.

[26] *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, NEW AMERICA (2018), https://newamericadotorg.s3.amazonaws.com/documents/Santa_Clara_Principles.pdf.

[27] *Id*.

[28] Suzor et al., *What Do We Mean When We Talk About Transparency? Towards Meaningful Transparency in Commercial Content Moderation*, 13 INT'L J. COMMC'N 1526 (2019), https://ijoc.org/index.php/ijoc/article/view/9736.

[29] Mike Ananny, *Networked Press Freedom: Creating Infrastructures for a Public's Right to Hear*, MIT PRESS (2018).

meet the increasingly automated nature of content moderation,[30] the Manila Principles advise that transparency reports comprise "actions taken on government requests, court orders, private complainant requests, and enforcement of content restriction policies."[31] Given Canada's stated aims to oversee company decisions by an external regulator, it follows that the upcoming bill ought to draw from the available preceding principles to protect freedom of expression and information online.

While we recognize the value and limitation of existing voluntary transparency mechanisms, we are here concerned with mandatory transparency requirements by governments,[32] especially in conjunction with take-down legislation. Due to mounting concerns over excessive content removal and censorship of free speech coming from take-down legislation such as Canada's upcoming proposal, mandatory transparency reporting will be even more important for the country to consider.

To date, the German *Network Enforcement Act* (NetzDG) is the only legislation to make public transparency reporting[33] mandatory for major platforms, obliging German firms with more than 2 million users to report "details about operational procedures, content removals across various sections of the German Criminal Code, and the way in which users were notified about content takedowns"[34] The European *Digital Markets Act* (DMA) and *Digital Services Act* (DSA), although both in the draft stage, likewise update regulations for online platforms, including requiring transparency,

---

[30] *See EFF Seeks Public Comment About Expanding and Improving Santa Clara Principles*, ELEC. FRONTIER FOUND. (2020), https://www.eff.org/press/releases/eff-seeks-public-comment-about-expanding-and-improving-santa-clara-principles.

[31] *The Manila Principles on Intermediary*, supra note 25, at 52–6.

[32] Transparency can also be mandated through voluntary membership in informal governance groups or civil society organizations, such as the Global Network Initiative. *See, e.g.*, *Enhancing Diversity and Participation*, GLOB. NETWORK INITIATIVE (2019), https://globalnetworkinitiative.org/wp-content/uploads/2020/07/GNI-Annual-Report-2019.pdf.

[33] Although several governments have mandated public disclosure of political advertising, especially during election intervals, Canada has already joined other countries including France (and debated in the UK and US) in mandating public registries.

[34] Robert Gorwa, *supra* note 21 at 300; *see also* Ben Wagner et al., *Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act*, PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (2020), https://dl.acm.org/doi/abs/10.1145/3351095.3372856.

user safety, and platform accountability.[35] The DSA specifically outlines due diligence and transparency obligations that include conducting "annual risk assessments regarding illegal content, negative effects on fundamental rights, and intentional manipulation of their services."[36] Similarly to NetzDG, both the DMA and DSA account for platform size in their requirements: the DMA only affecting "gatekeepers"—platforms with at least 45 million monthly active users; the DSA applying only to intermediaries used by more than 10% of EU consumers.[37] Other countries, such as the U.S., are proposing mandatory transparency and due process be enforced by the Federal Trade Commission.[38]

ii. Limitations of Existing Transparency Reporting

Independent researchers have identified several limitations of the data disclosed in transparency reports, including a lack of granularity required for independent analysis[39] and a lack of context needed to understand what types of speech are removed, and their broader patterns and impacts.[40] Others still have noted that exclusively technical statistics may not be "necessarily useful in reducing the overall opacity of the system, where key processes, protocols, and procedures remain secret."[41] More contextual data is

---

[35] Aline Blankertz and Julian Jaursch, *What the European DSA and DMA Proposals Mean for Online Platforms*, BROOKINGS INST.: TECH STREAM (2021), https://www.brookings.edu/techstream/what-the-european-dsa-and-dma-proposals-mean-for-online-platforms/.

[36] *Id.*

[37] *Id.*

[38] For a comprehensive review of US proposals see Zie Bedell and John Major, *What's Next for Section 230? A Roundup of Proposals*, LAWFARE (2020), https://www.lawfareblog.com/whats-next-section-230-roundup-proposals, *see also* Aleksandra Kuczerawy, *Safeguards for Freedom of Expression in the Era of Online Gatekeeping*, 2017 AUTEURS & MEDIA 3, 17–8 (2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3247682.

[39] Amélie Heldt, *Reading Between the Lines and the Numbers: An Analysis of the First NetzDG Reports*, 8 INTERNET POL'Y REV. 2 (2019), https://policyreview.info/articles/analysis/reading-between-lines-and-numbers-analysis-first-netzdg-reports.

[40] Chris Tenove and Heidi Tworek, *Processes, People, and Public Accountability: How to Understand and Address Harmful Communication Online*, RESEARCH REPORT - CANADIAN COMMISSION ON DEMOCRATIC EXPRESSION 14 (2020), https://www.mediatechdemocracy.com/work/processes-people-and-public-accountability-how-to-understand-and-address-harmful-communication-online.

[41] Robert Gorwa, *supra* note 21 at 302.

therefore crucial to understand and mitigate against excessive removal of lawful or dissenting speech, especially by historically marginalized groups online.[42] This too, includes activists whose speech may be removed because of platforms' failure to understand the context in which it is operating.[43] As Assistant Professor of Law, Rebecca Hamilton notes, this is partially due to platforms' failure to properly invest in "the localized cultural competence needed to fairly enforce their own standards in the markets they entered."[44] Simply reporting on, or inferring a measure of success from a quantified sum of content removals, without this kind of contextual data, would thus belie efforts to inform oversight and determine conditions for enforcement of regulatory frameworks.

B. Drafting and Legislative Design

i. Brief Overview of Intermediary Liability

Broadly speaking, there are three models of intermediary liability: strict liability, conditional liability, and broad immunity.[45] Countries such as China and Thailand operate under strict liability regimes, wherein platforms are held responsible for third-party content.[46] Meanwhile, conditional liability regimes offer platforms immunity on the condition that they adhere to prescribed procedures.[47] Notice-and-takedown regimes like the U.S. *Digital Copyright Millennium Act* (1998) fall under this category, requiring that platforms remove content upon receiving a notice of infringement.[48]Under broad immunity models, intermediaries are not held liable for third-party content.[49] Section 230 of the U.S.

---

[42] Rebecca J. Hamilton, *Governing the Global Public Sphere*, HARV. INTL'L L. J. (*forthcoming* 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3426544.

[43] Rebecca J. Hamilton, *De-platforming Following Capitol Insurrection Highlights Global Inequities Behind Content Moderation*, Just Security (2021), https://www.justsecurity.org/74258/de-platforming-following-capitol-insurrection-highlights-global-inequities-behind-content-moderation/.

[44] *Id.*

[45] Article 19, *Internet Intermediaries: Dilemma of Liability* 7 (2013), https://www.article19.org/data/files/Intermediaries_ENGLISH.pdf.

[46] *Id.*

[47] *Id.*

[48] *Id.*

[49] Article 19, *supra* note 45.

Communications Decency Act is likely the most notable example thereof, which provides that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[50]

Evidence suggests that Canada will be pursuing some form of notice-and-take down regime.[51] It remains to be seen how this new approach will comply with Article 19.17.2 of the United States-Mexico-Canada Agreement, under which Canadian law cannot "treat a supplier or user of an interactive computer service as an information content provider in determining liability for harms related to information stored, processed, transmitted, distributed, or made available by the service, except to the extent the supplier or user has, in whole or in part, created or developed the information."[52]

Within the context of notice-and-action regimes, different types of speech require different considerations.[53] Some types of speech may require prior judicial determination, while others may not. The non-consensual distribution of intimate images, for example, is easy to assess because such material is considered low-value expression[54] and removal thereof is not likely to harm expressive freedom.[55] A simple complaint that an image was distributed without consent should be sufficient grounds for removal.[56] On the other hand, hate speech and incitement of terrorism bear more ambiguity. Courts may be better-suited to assess

---

[50] *Id.*; *see also* Communications Decency Act § 230(c), 47 U.S.C. (1996).

[51] Bill Curry, *Heritage Minister Says Takedown Rules Coming, Welcomes Calls for New Social-Media Regulator*, THE GLOBE AND MAIL (Jan. 27, 2021), https://www.theglobeandmail.com/politics/article-heritage-minister-says-takedown-rules-coming-welcomes-calls-for-new/.

[52] Vivek Krishnamurthy et al., *CDA 230 Goes North American? Examining the Impacts of the USMCA's Intermediary Liability Provisions in Canada and the United States*, The Samuelson-Glushko CANADIAN INTERNET POL'Y AND PUB. INT. CLINIC AND CYBER CLINIC AT HARV. LAW SCHOOL'S BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3645462.

[53] Emily Laidlaw, *Notice-and-Notice-Plus: A Canadian Perspective Beyond the Liability and Immunity Divide*, OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIAB. 455–466 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3311659.

[54] *Id.*

[55] *Id.* at 458.

[56] *Id.*

difficult questions of what constitutes terrorism propaganda.[57] Similarly, courts may be better-suited to make determinations on whether something is merely offensive or whether it can be classified as hate speech.[58] Others may argue that requiring judicial determination is costly, inefficient, and a barrier to justice.[59] We do not weigh in on this thorny issue.

Under notice-and-action regimes, different types of speech may also require different configurations of notice requirements and counter-notice mechanisms. However, whatever the regime may look like—be it notice-and-notice or notice-and-takedown—it is worth considering integrating due process into its architecture.[60] A greater level of specificity not only about different types of speech but also regarding the characteristics of platforms and intermediaries to which the hate speech legislation should apply —and how—will evade limitations of a 'one-size-fits-all-approach' to content moderation processes for smaller or newer start-ups, especially those that reach "popularity-by-surprise."[61] Given the complexity of take-down legislation, we have taken a circumscribed approach. With a view to protecting expressive freedom and due process, we suggest that Canada's new take-down legislation be clear on two fronts: (a) in its definitions of unlawful speech that must be removed, and (b) in its notice requirements. Ambiguous definitions of what may constitute unlawful speech and a lack of guidance on adequate notice encourages intermediaries to err on the side of censorship to avoid liability and open the door to abuses of process. Furthermore, we suggest that Canada consider integrating a counter-notice mechanism into its framework in order to stem over-removal and to protect due process.

ii. Defining Unlawful Speech

---

[57] *Id*. at 459–462.

[58] *Id*. at 462–465.

[59] *Id*. at 453.

[60] *Id.* at 456.

[61] *See* Ysabel Gerrard, *Too Good to be True: The Challenges of Regulating Social Media Startups* in *Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates*, 9 INTERNET POL'Y REV.4 (2020), https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy#footnoteref2_apdze36.

Legal clarity is an essential principle of the rule of law.[62] It is thus important that the Canadian legislative framework regulating online speech be "precise, clear, and accessible."[63] As the Manila Principles note, "[i]mposing liability on internet intermediaries without providing clear and accessible guidance as to the precise type of content that is not lawful and the precise requirements of a legally sufficient notice encourages intermediaries to over-remove content."[64]

Canada can learn from cases of legal ambiguity in other jurisdictions that have taken on the challenge of defining unlawful speech in take-down legislation. For example, concerns over imprecision animate some of the criticisms of NetzDG.[65] Critics observe that the German law, which requires platforms to remove "content that is manifestly unlawful" within twenty-four hours or risk fines up to €50M, provides platforms with too little guidance and too much discretion and is likely to result in over-compliance.[66] As mentioned above, Canada's Heritage Minister noted that the incoming take-down regime will contain clear definitions of each category of online speech. This may well have been a response to a lesson learned from Germany.

iii. Specifying Procedural Requirements

Well-defined notice requirements help prevent over-removal by providing platforms with guidance on how to respond to removal requests, thus tempering incentives to err on the side of caution to

---

[62] Jeremy Waldron, *The Rule of Law*, STAN. ENCYCLOPEDIA OF PHI. (2016), https://plato.stanford.edu/entries/rule-of-law/#FormProcSubsRequ.

[63] *The Manila Principles on Intermediary*, supra note 31 at 18; *see also* Aleksandra Kuczerawy, *supra* note 38 at 6.

[64] *Id*.

[65] Amélie Heldt, *Reading Between the Lines and the Numbers: An Analysis of the First NetzDG Reports*, 8 INTERNET POL'Y REV. 2 (2019), https://policyreview.info/articles/analysis/reading-between-lines-and-numbers-analysis-first-netzdg-reports.

[66] *Id*. at 5; *see also* House of Lords Select Committee on Communications, *The Internet: To Regulate or Not to Regulate?*, Article 19 (2018), https://www.article19.org/wp-content/uploads/2018/06/HOL-inquiry-Internet-regulation-A19-written-evidence-18-May-2018-.pdf and David Kaye, *Report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UNITED NATIONS GENERAL ASSEMBLY: HUMAN RIGHTS COUNCIL (2016), https://documents-dds-ny.un.org/doc/UNDOC/GEN/G16/095/12/PDF/G1609512.pdf?OpenElement.

avoid liability while also deterring vexatious removal requests. According to the Manila Principles, complainants must provide information including the legal basis for the removal request, the Internet identifier and a description of the content in question, an overview of potential defenses open to the recipient, and documentation of legal standing.[67]

Annemarie Bridy and Daphne Keller observe that without clearly codified notice requirements, "[i]ntermediaries do not receive enough information to meaningfully assess a legal claim, but they feel they must take prompt action or risk liability."[68] Section 512(c)(3)(A) of the DCMA, however, outlines elements of notification, which require complainants to identify matters including the copyrighted work in question and the content allegedly infringing it. Keller and Bridy likewise observe that these requirements provide intermediaries with guidance on the take-down process and reduce the likelihood of over-removal.[69]

Thorough notice requirements also restrain abuses of the take-down system. The DMCA's elements of notification require complainants to state in their notice that they have a "good faith belief that use of the material in the manner complained of is not authorized by the copyright owner, its agent, or the law."[70] In the same vein, they must attest to the accuracy of the contents of the notification under penalty of perjury.[71] Bridy and Keller note that these measures deter frivolous complaints.[72] Requirements for content notices must thus be legislated in clear detail.[73]

---

[67] *The Manila Principles on Intermediary*, supra note 31 at 31.

[68] Annemarie Bridy and Daphne Keller, S*ection 512 Study: Comments of Annemarie Bridy and Daphne Keller*, U.S. COPYRIGHT OFFICE 23 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2920871; *see also* Aleksandra Kuczerawy, *supra* note 38 at 8, who argues that a "legal framework providing safeguards for the notice and action mechanisms procedure should specify the formal requirements for a valid notice, i.e. what information must be included to put the mechanisms in motion."

[69] *Id.*

[70] Digital Millennium Copyright Act § 512(c)(3)(A)(v).

[71] *Id.* § 512(c)(3)(A)(vi).

[72] Annemarie Bridy and Daphne Keller, *supra* note 68.

[73] Emily Laidlaw, *supra* note 53 at 453; *see also* Aleksandra Kuczerawy, *From "Notice and Takedown" to "Notice and Stay Down": Risks and Safeguards for Freedom of Expression*, OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY (2020), https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780198837138.001.0001/oxfordhb-9780198837138-e-27; Aleksandra Kuczerawy, *supra* note 38 at 8–9.

Canada need look no further than its own copyright regime to learn the importance of clear notice requirements. Even seemingly detailed regulations are still "not entirely free from issues regarding interpretation and application."[74] When Canada's notice-and-notice regimes neglected to codify restrictions on the contents of notices, complainants took to intimidating respondents by incorporating settlement offers.[75] As a result, the federal government amended the *Copyright Act* to prohibit the inclusion of settlement offers and demands for payment or personal information within infringement notices.[76] Laidlaw endorses the DMCA's content requirements, and notes that Canada's misstep can serve as a cautionary tale for future notice-and-action regimes.[77]

In the context of a notice-and-notice regime for defamation, Laidlaw further notes that including the legal basis for one's complaint can be required by law. This would fall in line with the content requirements outlined in the Manila Principles. Platforms could model their complaint forms on the legal grounds set out in law, putting the onus on complainants to make a case by demonstrating the elements of their claims.[78] These forms may assist them in making a determination, and would not only provide guidance to platforms, but would temper abusive requests.[79]

Legislators, however, should also be weary of introducing procedural hurdles that are too onerous. Instances of a non-consensual distribution of intimate images or child pornography may not be the appropriate context to include intimidating legalese or demand complex legal justifications. Murkier categories such as hate speech may benefit from forms detailing the legal elements of a claim. As mentioned, different categories of unlawful speech will inevitably invite their own unique considerations.[80] We hope that the Canadian government will consider these difficult questions while developing its framework.

---

[74] Aleksandra Kuczerawy, *supra* note 68 at 534.
[75] *Id.*
[76] Copyright Act § 41.25(3).
[77] Emily Laidlaw, *supra* note 53 at 453.
[78] *Id.*
[79] *Id.*
[80] *See, e.g.*, Aleksandra Kuczerawy, *supra* note 38 at 9–10.

C. Clear Notice and Counter-Notice Requirements

Bridy and Keller argue that, when combined with clear notice requirements, counter-notice mechanisms may help minimize the risk of over-removals in content moderation practices.[81] These counter-notices, which challenge the legitimacy of a notice of complaint, have been widely suggested in effort to reform traditional notice requirements, and to mandate platforms to moderate in good faith.[82] The possibility of a counter-notification mechanism allows parties to respond to a complaint and put forward a defense for their use of the content.[83] Counter-notifications must usually meet specified requirements and time frames. They are resolved by the hosting providers, who can effectively put the content back online. In the DMCA, for example, content is reinstated after a counter-notice in ten to fourteen days, unless the service provider receives no-tice that the rights holder took the case to court.[84] In other words, counter-notice mechanisms allow parties subject to content moderation to defend their allegedly unlawful post(s) from removal.

Counter-notice mechanisms are an important tool in "tailoring" intermediary liability laws.[85] Although research has shown that counter-notice mechanisms are rarely used in practice, their "symbolic acknowledgment" of users' expressive rights emphasize platform companies' curatorial obligations of due process.[86] Bridy and Keller note that the safeguarding of these expressive rights is—or at least should be—of primary importance to both users and platform corporations who, historically, have had little involvement in the negotiation of best practices for mitigating online harms, including in content removal practices.[87] Guidance by

---

[81] Annemarie Bridy and Daphne Keller, *supra* note 68.

[82] Emily Laidlaw and Hilary Young, *Internet Intermediary Liability in Defamation*, 56 OSGOODE HALL L. J. 1 (2019), https://digitalcommons.osgoode.yorku.ca/cgi/viewcontent.cgi?article=3389&context=ohlj, at 129.

[83] Emily Laidlaw and Hilary Young, *Internet Intermediary Liability in Defamation: Proposals for Statutory Reform*, LAW COMMISSION OF ONTARIO 104 (2017), http://www.lco-cdo.org/wp-content/uploads/2017/07/DIA-Commissioned-Paper-Laidlaw-and-Young.pdf.

[84] Aleksandra Kuczerawy, *supra* note 68 at 531.

[85] Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money*, HOOVER INSTITUTION Aegis Series Paper no. 1807 18 (2018).

[86] Aleksandra Kuczerawy, *supra* note 68 at 532.

[87] Annemarie Bridy and Daphne Keller, *supra* note 68 at 4.

the Manila Principles to uphold due process in content moderation processes is essential to the protection of users' fundamental rights in the digital sphere.[88] Both the U.S. Department of Justice and the European Commission's Digital Services Act have proposed a series of recommendations for due process to impose "more effective redress and protection against unjustified removal for legitimate content...online."[89] Requiring the implementation of a counter-notice mechanism is one way to introduce elements of due process into platform governance procedures.

IV. FURTHER CONSIDERATIONS AND CONCLUDING REMARKS

While this note focuses on a small subset of content moderation as it relates to transparency reporting regarding content removal and clarity of definitions and procedures, the field of content moderation implicates a much broader policy landscape,[90] including significant issues relating to labor conditions,[91] human

---

[88] *Id.* at 531; *see also* Emily Laidlaw, *supra* note 53 at 454; Jack Balkin's proposal of curatorial due process in Jack Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011(2018); Aleksandra Kuczerawy, *supra* note 38 at 12-4.

[89] Mark McCarthy, *The Justice Department's Good Ideas for Platforms Needn't Be Done Through Section 230 Reform*, LAWFARE (2020), https://www.lawfareblog.com/justice-departments-good-ideas-platforms-neednt-be-done-through-section-230-reform; *see also Digital Services Act: Deepening the Internal Market and Clarifying Responsibilities for Digital Services*, EUROPEAN COMMISSION (2020), https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-Internal-Market-and-clarifying-responsibilities-for-digital-services.

[90] Note that there are several other considerations pertinent to legislation dealing with unlawful speech on the Internet specifically, and content moderation within platform governance more broadly, that fall outside the scope of this paper but which require careful attention, including algorithmic transparency. *See, e.g.*, Gorwa et al., *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, BIG DATA & SOC'Y (2020), https://journals.sagepub.com/doi/full/10.1177/2053951719897945.

[91] Sarah T. Roberts, *Commercial Content Moderation is a Soft Economic and Political Tool* in *Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates*, 9 INTERNET POL'Y REV. 4 (2020), https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy#footnoteref2_apdze36.

rights and user rights[92] and geopolitics.[93] As such, some have argued that the term "content moderation" is far too limited to describe both the breadth of actions and decisions platforms take[94] and their global impact. Content policy is itself only one part of an increasingly overlapping global platform governance agenda that intersects across content moderation, data governance, competition policy, and antitrust law.[95] Interdisciplinary research aimed to inform law and policy on platform governance must be attuned to some of these broader considerations, as well as move beyond predominantly text-based analyses of hate speech, a focus on the dominant U.S. based platforms, and a lack of critical race perspectives when examining hate speech and social media.[96] This is especially pertinent given that content moderation is far from a neutral process.[97]

In calls for greater transparency, there often lie assumptions about the ability of that transparency, writ large, to bolster greater platform responsibility. The call for transparency reporting without regulation to necessitate accountability, though, has been viewed by some as a "market friendly" response to the demands for corporate oversight.[98] Rather, transparency reports must include privacy-preserving contextual and granulated data so as to provide access to

---

[92] Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L. J. 1353 (2018), https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2018/07/Regulating-Online-Content-Moderation.pdf.

[93] Tarleton Gillespie et al., *Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates*, 9 INTERNET POL'Y REV. 4 (2020), https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy#footnoteref2_apdze36.

[94] Sarah Myers West, *Thinking Beyond Content in the Debate About Moderation* in *Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates*, 9 INTERNET POL'Y REV. 4 (2020), https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy#footnoteref2_apdze36.

[95] Taylor Owen, *The Case for Platform Governance*, CTR. FOR INT'L GOVERNANCE Innovation Paper No. 231 (2019), https://www.cigionline.org/publications/case-platform-governance.

[96] Ariadna Matamoros-Fernández and John Farkas, *Racism, Hate Speech, and Social Media: A Systematic Review and Critique*, 22 TELEVISION AND NEW MEDIA 2 (2021), https://journals.sagepub.com/doi/full/10.1177/1527476420982230.

[97] Rebecca J. Hamilton, *supra* note 43.

[98] Suzor et al., *What Do We Mean When We Talk About Transparency? Towards Meaningful Transparency in Commercial Content Moderation*, 13 INT'L J. COMMC'N (2019), https://ijoc.org/index.php/ijoc/article/view/9736.

researchers seeking to understand broader patterns and structures of online harm, but also for oversight bodies or external regulators seeking to review and enforce upcoming legislation. Platforms may —and often do—resist this type of reporting, though, especially with concern for privacy and the protection of proprietary technology, investments, and trade secrets.[99]

In conclusion, Canada's announcement seems like an early step toward greater online oversight and accountability, but it remains to be seen how it will safeguard transparency, due process, and freedom of expression and information while ensuring public and civil society participation. Any attempt to protect online speakers from both harmful/unlawful speech and oppressive content moderation must simultaneously accommodate for content moderation regimes that uphold and protect the inherently democratic ideals of online spaces.[100] This surely difficult task requires a thoughtful and well-informed approach.

---

[99] Mike Ananay and Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability*, NEW MEDIA AND SOC'Y 7 (2016), http://ananny.org/papers/anannyCrawford_seeingWithoutKnowing_2016.pdf.

[100] Kyle Langvardt, *supra* note 92 at 1362.