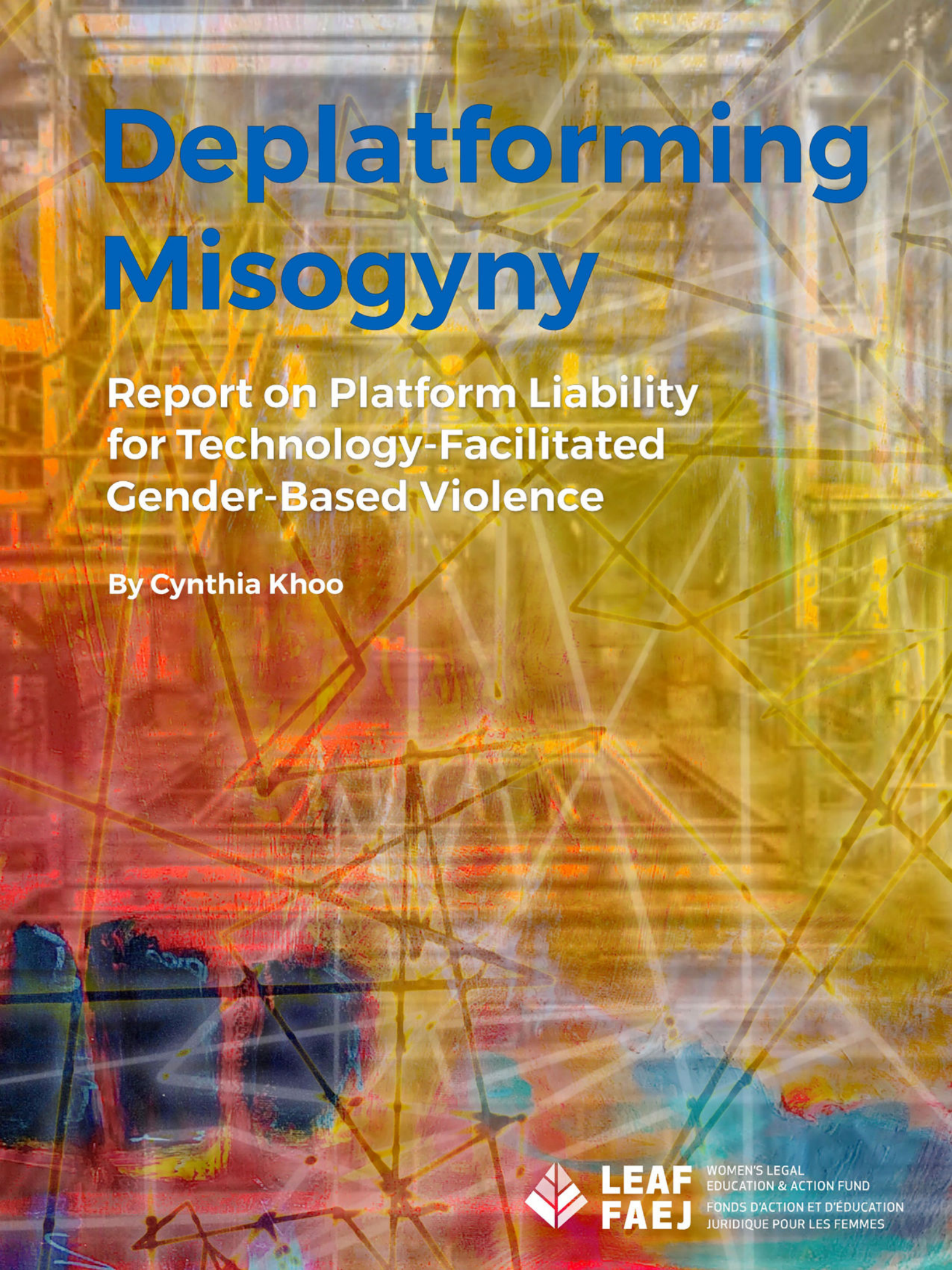

Deplatforming Misogyny

Report on Platform Liability for Technology-Facilitated Gender-Based Violence

By Cynthia Khoo

LEAF FAEJ
WOMEN'S LEGAL EDUCATION & ACTION FUND
FONDS D'ACTION ET D'ÉDUCATION JURIDIQUE POUR LES FEMMES

Published by
Women's Legal Education and Action Fund (LEAF)
180 Dundas Street West, Suite 1420
Toronto, Ontario, Canada M5G 1C7
www.leaf.ca

LEAF is a national, charitable, non-profit organization, founded in 1985. LEAF works to advance the substantive equality rights of women and girls in Canada through litigation, law reform and public education using the *Canadian Charter of Rights and Freedoms*.

This publication was created as part of LEAF's Technology-Facilitated Violence (TFV) Project. The TFV Project brings together feminist lawyers and academics to conduct research and produce publications imagining legal responses to TFV against women and gender-diverse people that are informed by equality principles. The project also supports and informs LEAF's law reform efforts and potential upcoming interventions concerning TFV.

Acknowledgements

Deep gratitude and appreciation go to the many people whose efforts and support made this publication possible.

This report was researched and written by **Cynthia Khoo**, a technology and human rights lawyer and researcher. Cynthia holds an LL.M. (Concentration in Law and Technology) from the University of Ottawa, where she worked on cases as junior counsel at the Samuelson-Glushko Canadian Internet Policy and Public Interest Clinic (CIPPIC). Her paper on platform liability for emergent systemic harm to historically marginalized groups received the inaugural Ian R. Kerr Robotnik Memorial Award for the Best Paper by an Emerging Scholar at We Robot 2020. She has managed a sole practice law firm, Tekhnos Law, and obtained her J.D. from the University of Victoria. Her work and expertise span across key areas of technology and human rights law and policy, including privacy and surveillance, equality and freedom from discrimination, online censorship and freedom of expression, intermediary liability, algorithmic decision-making, and technology-facilitated violence.

The report was overseen and coordinated by **Rosel Kim**, Staff Lawyer at LEAF; **Pam Hrick**, Executive Director and General Counsel at LEAF; and **Megan Stephens**, former Executive Director and General Counsel at LEAF.

Sincere thanks are further extended to **Pam Hrick** and **Rosel Kim** for contributing substantive revisions, editing, and copyediting of the report.

This report benefited significantly from the insights and expertise of the following reviewers (in alphabetical order by last name): **Moira Aikenhead**, **Jane Bailey**, **Karen Bellehumeur**, **Nicole Biros-Bolton**, **Suzie Dunn**, **Lex Gill**, **Nicola Henry**, **Pam Hrick**, **Tamir Israel**, **Jo-Ann Kolmes**, **Rosel Kim**, **Emily Laidlaw**, **Brenda McPhail**, **Jill Presser**, **Molly Reynolds**, **Megan Stephens**, and **Jennifer Tomaszewski**.

Thank you as well to **Donald Jackson, Julie Mouris, Kat Owens, Cee Strauss, Rosel Kim,** and **Pam Hrick**, for much-appreciated assistance with the citations for this report.

Special thanks to the **LEAF TFV Advisory Committee**: Moira Aikenhead, Jane Bailey, Karen Bellehumeur, Suzie Dunn, Gillian Hnatiw, Pam Hrick, Nathalie Léger, Raine Liliefeldt, Molly Reynolds, Hadiya Roderique, and Karen Segal.

Thank you to **Kristyn Watterworth**, who designed the cover page of this report.

This report was funded by the generous support of: the Canadian Bar Association's Law for the Future Fund; the Pilot Fund for Gender Equality, a collaboration between Community Foundations of Canada and the Equality Fund, with support from the Government of Canada; and the J. Armand Bombardier Foundation.

Table of Contents

Executive Summary

Digital platforms have enhanced and expanded the ways in which we interact and share information with one another. They have also simultaneously provided new mechanisms for those who might seek to engage in abusive conduct to inflict harm on targeted groups and individuals—particularly from historically marginalized and systemically oppressed communities.

This report examines the role of digital platforms in the proliferation of technology-facilitated gender-based violence, abuse, and harassment (abbreviated as 'TFGBV'). It also examines whether and how digital platforms—such as Facebook, YouTube, and Twitter—should be held accountable for TFGBV through regulation or the imposition of liability under Canadian law.

The consideration of these issues begins with a review of the substance and nature of TFGBV that commonly occurs on digital platforms, as well as examples of platform content moderation models. This is followed by an explanation of the current Canadian landscape concerning platform liability for TFGBV and a review of platform liability regimes that exist in other jurisdictions around the world. The report then grapples with some critical issues and legal complexities associated with holding platforms liable for user conduct. It concludes by making 14 recommendations for federal legal reform and complementary actions to address TFGBV in Canada through the lens of digital platform liability and accountability.

Technology-Facilitated Gender-Based Violence, Abuse, and Harassment (TFGBV)

TFGBV refers to a spectrum of activities and behaviours that involve technology as a central aspect of perpetuating violence, abuse, or harassment against (both cis and trans) women and girls. This term also captures those who hold intersecting marginalized identities, such as 2SLGBTQQIA, Black, Indigenous, and racialized women; women with disabilities; and women who are socioeconomically disadvantaged.

Activities that fall under the umbrella of TFGBV include:

- doxing;
- hate speech;
- threats and intimidation;
- trolling;
- voyeurism;
- impersonation;
- spying and monitoring through account hacking or interception of private communications;
- online mobbing;
- coordinated flagging campaigns;
- sexual exploitation resulting from online luring;
- defamation;
- non-consensual distribution of intimate images (NCDII);
- image-based abuse (including both deepfakes and shallow fakes);
- sextortion; and
- stalking.

These activities may be referred to more generally as aspects or examples of 'online violence', 'online abuse', or 'online harassment'. The terms are not necessarily interchangeable, and depend on context.

TFGBV relegates women and girls to secondary status online and in the world. They are rendered unable to freely and fully participate in society and prevented from enjoying true or equal protection of their human rights and fundamental freedoms. The most common response to facing online abuse and harassment is that women reduce their online activities, avoid certain social media platforms or conversations, withdraw from expressing their views, or self-censor if they continue to engage online. This curtails their ability to participate in the contemporary public sphere, including engaging in activism and advocacy, influencing public opinion, or mobilizing social, cultural, or political change. The current state of affairs amounts to a systemic democratic failure and must be addressed as such.

Nearly all TFGBV on digital platforms is committed through online expression, whether through speech, images, videos, or other multimedia. Whether or not a specific instance of TFGBV is illegal depends on whether it meets the definition of a pre-existing criminal offence or cause of civil action. For example, acts of TFGBV that constitute invasion of privacy, impersonation, defamation, criminal harassment, threats of violence, interception of private communications, stalking, recording or surveilling someone without consent (where they have a reasonable expectation of privacy), or NCDII are all already civil and/or criminal offences in Canada.

Instances of TFGBV that fall short of attracting legal liability might be considered 'just' speech or expression. However, expression-based TFGBV can be as or more damaging to women and girls and impact their lives in ways that go far beyond the screen. This may include, for example, 'everyday' online harassment amounting to social persecution; violent threats that fall short of the legal threshold for criminal liability; trolling; creating and disseminating non-sexualized deepfakes; and online mobbing. One question in Canadian law is whether digital platforms can be regulated to address this level of expression-based TFGBV by their users, whether through regulatory obligations or imposed liability. Answering that question is a complicated exercise that raises questions around intermediary liability, constitutional limitations, proportionality, and the particular dynamics of digital platforms and their role in society. A principled focus on the right to equality—intersectional, substantive equality—can help navigate such questions in setting a path forward to meaningfully address TFGBV.

Role of Digital Platforms in TFGBV

Online platforms such as social media networks, discussion forums, search engines, and video sharing websites have become central venues of our personal and professional lives. It thus comes as no surprise that online platforms are also central sites of TFGBV, which has often been exacerbated by the actions (or inaction) of the platforms themselves. For example:

- Facebook has allowed pages glorifying intimate-partner violence to stand, while removing images of women breastfeeding;
- Twitter has been quick to suspend users who are targets of online abuse, while frequently ignoring the activity of abusive users;
- YouTube's recommendation algorithms have turned it into an efficient right-wing radicalization machine; and
- Google Search has provided top-ranked search results that reflect racist sexual objectification of Black women and girls.

One particular type of platform that warrants specific attention in the context of TFGBV is the category of platforms that seem deliberately designed to encourage and profit from such abuse. These might be termed '**purpose-built platforms**', as opposed to 'platforms of general application' such as Facebook and Twitter, which may include (copious) TFGBV, but do not appear to exist exclusively to cater to TFGBV. Examples of 'purpose-built platforms' are 'The Dirty' and platforms dedicated to sharing NCDII. What would be a balanced and proportionate liability framework for platforms of general application would likely not suffice to address TFGBV where such expression and conduct constitute the core business model or central service or commodity of a purpose-built platform.

The total constellation of TFGBV as facilitated by digital platforms, particularly, can be termed **platformed misogyny**, or **platformed TFGBV**, based on Ariadna Matamoros-Fernández's concept of 'platformed racism'. The term is used to denote how the characteristic features of digital platforms' design choices, business models and content moderation policies—including their embedded cultural values and politics—combine with the power of platform governance to shape the 'platformed' systemic oppression in question, in a way that makes it distinct from non-platformed manifestations.

Digital platforms share several common features that contribute to such companies' particular role in amplifying, promoting, escalating, and entrenching TFGBV. These features include:

- platform companies' advertising-driven business models, which maximize user engagement in a way that favours more outrageous and sensationalized content;
- companies' prioritization of business growth above all else;
- the sheer ease, efficiency, and affordability of automating and multiplying instances of abuse against a particular group or individual;
- the ability for abusive users to remain anonymous and remote, taking advantage of 'safety in numbers' in online mobs or coordinated attacks; and
- the ability of users to game content moderation features and other platform affordances to abusive ends.

As a result of these platform dynamics, gender-based violence, abuse, and harassment is no longer constrained by physical boundaries. The ubiquity of the Internet means that TFGBV can become omnipresent and relentless, infiltrating a victim's most intimate physical spaces, such as their home or bedroom. Users engaging in TFGBV can also leverage their own and targeted individuals' online social networks to further the abuse, by recruiting others to knowingly or unwittingly share abusive material, and by contaminating the targeted individuals' own online spaces and communities. The online permanence of abusive material—which is exceedingly difficult to completely eradicate once shared online—also ensures continued revictimization, resulting in lasting psychological and other damage.

Platform Content Moderation Policies and Practices to Address TFGBV

When it comes to addressing TFGBV, platform content moderation measures have been deficient in both design and application. This has exacerbated harms to users most targeted by TFGBV, including historically marginalized groups. For example, community standards on digital platforms have included exceptions to rules prohibiting hateful or harmful speech. This has created major loopholes for demonstrably hateful or harmful content to proliferate. Flagging and reporting mechanisms rely on

users using them accurately and in good faith, but have often been gamed to further the abuse such mechanisms are meant to address.

Human reviewers are generally underpaid third-party contractors working in traumatizing conditions who have only seconds to determine whether a given post should be left up, taken down, or escalated. Automated content moderation is rife with further errors, which have resulted in the removal of content, including posts that constitute parody and satire; innocuous images mistaken for nudity; and content related to 2SLGBTQQIA issues or sex education. Algorithmic tweaking and downranking, fact-checking, and labelling have been applied weakly, inconsistently, and highly selectively, for the most part.

In addition, platform companies have consistently demonstrated significant degrees of selective attentiveness, contradiction, and hypocrisy in both the development and application of their content moderation policies and practices. Those targeted by TFGBV or otherwise familiar with the issue have continually reported that major platform companies ignore individual requests for help and largely neglect the broader issue of TFGBV. At the same time, they continue to support and build features that contribute to optimizing their platforms for abuse. Experts have also identified overarching systemic issues with digital platforms' content moderation policies and practices, including:

- selective reliance on the rhetoric and strength of the United States' cultural norms around 'freedom of expression' to justify inaction;
- undue reactivity and sensitivity to public opinion and political influence in content moderation decisions; and
- conflicts of interest that result in platform companies prioritizing business growth and maintaining good relations with the political right over effectively addressing TFGBV.

Canadian Legal Landscape: Platform Liability for TFGBV

There are a number of Canadian laws that could theoretically create platform liability for TFGBV by a platform's user, given the right circumstances. However, many of these have yet to be tested in court. Canada has laws that do the following:

- establish a general intermediary liability regime (e.g., section 22 of the *Act to establish a legal framework for information technology* in Quebec);
- establish platform liability or legal obligations for non-TFGBV user content (e.g., direct liability for 'enabling' copyright infringement and the notice-and-notice regime under the *Copyright Act*, or what is effectively a notice-and-takedown regime that the courts have developed in defamation law);
- address TFGBV in some form but are silent on the role of platforms (e.g., *Criminal Code* offences for NCDII and hate propaganda); and
- address neither TFGBV nor intermediary liability specifically, but are laws of general application that could apply to platform companies as organizations, provided the factual circumstances met the relevant legal test (e.g., statutory human rights law, criminal corporate negligence, or product liability).

There thus appears to be a gap in Canadian law, in that there is no *specific* form of legal liability for platforms with respect to TFGBV. However, some common principles emerge from current intermediary liability law and jurisprudence, which can inform how the law should be extended to address platform liability for TFGBV.

As a starting point, courts have generally been reluctant to hold online intermediaries liable for user expression or conduct, without something more to justify holding one party liable for another party's misconduct. This is particularly true where the intermediary is a 'mere conduit' and simply plays an infrastructural role of connecting third parties to one another. However, Canadian defamation law will hold a platform accountable for a user's speech if the platform had specific knowledge about it but took no action to address it. Canadian copyright law places a legal obligation on platforms to assist potentially injured parties, but will not hold a platform liable for user copyright infringement unless the platform's involvement rises beyond a certain level according to a six-factor test for being an 'enabler'. Overall, the degree of liability rises the more the platform is involved and the more that is at stake for the injured party, up to direct liability where the platform has essentially abandoned its 'intermediary' role in producing content that constitutes a civil or criminal offence.

Even if a platform company is not a party to a legal proceeding and is not liable for the harmful content in question, it may still be required to take certain steps to address the content, including:

- complying with court orders or statutory obligations, such as forwarding a notice to the author;
- removing, deindexing, or disabling access to content; or
- releasing information to help identify an anonymous user engaging in abuse.

These obligations are rooted in ensuring access to justice and practical remedies for victims, in a way that recognizes the realities of the Internet. Providing platforms with explicit protection from liability acknowledges the particular role of platforms in the Internet ecosystem and with respect to specific harms—namely, a dominant and facilitative role which justifies accountability and responsibility for assisting in the remedy, but does not generally warrant imposing liability for the wrongdoing itself.

There are some laws of general application in Canada that could potentially ground platform liability for TFGBV on a systemic or institutional level, based on the platform company's design choices and business decisions. Examples include human rights statutes, in situations where disproportionately exposing historically marginalized groups to TFGBV constitutes discrimination in the provision of goods, services, or facilities. The laws of commercial host liability and product liability may also apply to digital platforms in certain circumstances, based on underlying reasoning of moral hazards and incentives to act against the customer (or user), or the public interest. Additional laws that may apply include criminal corporate negligence or being party to an offence as an organization, though these would only apply to existing criminal offences, such as hate speech, intimidation, threats of violence, or criminal harassment.

Platform Liability Models: Jurisdictional Scan

Several lessons and conclusions may be drawn from the jurisdictional scan of platform liability models for harmful behaviour by users contained in this report.

First, the manner in which section 230 of the *Communications Decency Act* (CDA 230) has been applied in the United States to provide broad immunity to platforms for user content does not strike an

appropriate balance among the considerations relevant to the context of TFGBV, nor does it necessarily reflect what CDA 230 was initially intended to achieve. The breadth of platform protection under CDA 230 derives largely from a long line of overly expansive interpretation by American courts. Danielle Citron and Benjamin Wittes suggest that the core benefits of CDA 230 could have been achieved without causing such significant damage to the lives of women and girls (particularly those who hold multiple intersecting identities) and their ability to fully participate in online spaces.

Second, the experiences of the *Allow States and Victims to Fight Online Sex Trafficking Act* and *Stop Enabling Sex Traffickers Act* (FOSTA-SESTA) in the United States demonstrate that it is imperative to listen to the vulnerable and marginalized populations who will be the most impacted by any proposed legislation. Lawmakers must consider the advice and insights of directly impacted communities in assessing whether a new law will disproportionately harm vulnerable and marginalized Internet users, including by driving them off of the Internet, rather than meaningfully address TFGBV itself. Experts in TFGBV and platform accountability have emphasized that any proposed platform liability regimes, as well as platforms' own policies and practices in content moderation, must be victim/survivor-centred and trauma-aware.

Third, Germany's experience with Netzwerkdurchsetzungsgesetz (NetzDG) suggests that relying on industry self-regulation is insufficient to meaningfully address TFGBV, for countries other than the United States where major American platform companies operate. The law's primary result appeared to be galvanizing greater enforcement of platforms' own community standards, whereas researchers found that a mass spike in wrongful takedowns did not happen, though the law continues to raise concerns. There was broad consensus that it was exceedingly difficult to determine whether the law effectively achieved its goal of reducing hate speech and other objectionable content, due to lack of meaningful data—highlighting that implementing a platform liability law may be of limited use without setting up a way to effectively evaluate its impact.

Fourth, jurisdictions such as the United Kingdom and European Union, through the *Online Harms White Paper* (and related documents) and proposed *Digital Services Act* (DSA) respectively, have begun to incorporate explicit recognition of the harms that digital platforms cause *systemically*. This is reflected in both regimes' tiered approaches, which place greater obligations on platforms beyond a certain size and influence; on the UK providing for 'super-complaints' to address systemic issues; and on the DSA's requirement that Very Large Online Platforms regularly assess and respond to systemic risk flowing from use of their services. Laws that address TFGBV and similar content on a systemic, not individualized, level are particularly important given the systemic nature of TFGBV as a pillar of structural discrimination and systemic oppression.

Fifth, digital rights advocates and those primarily concerned with generic freedom of expression (as opposed to the freedom of expression of those who are driven offline or forced to self-censor by TFGBV) have consistently raised valid concerns about platform liability regimes generally. Such concerns primarily focus on transparency and oversight; due process and appeal mechanisms; definitional clarity; and safeguards to mitigate wrongful takedowns or overbroad legislation. These concerns should be taken into account and given due weight, with measures to address them incorporated into any legal and policy reforms. Simultaneously, reforms must still apply an intersectional feminist analysis and focus on upholding substantive equality.

Constitutional and Critical Analysis of Platform Liability for TFGBV

Canadian constitutional and human rights law has repeatedly recognized the necessity and justifiability of limiting free expression in order to uphold equality rights and protect historically marginalized groups. The rights to equality and freedom from discrimination are as fundamental as freedom of expression, and equally protected under the *Canadian Charter of Rights and Freedoms*. This legal understanding must govern legislative reforms to address TFGBV, including through platform regulation and platform liability.

Multiple decisions from the Supreme Court of Canada, including *R v Taylor, R v Keegstra,* and *Saskatchewan Human Rights Commission v Whatcott*, in addition to *Lemire v Canada (Human Rights Commission)* at the Federal Court of Appeal, have affirmed the constitutionality of criminal and statutory human rights laws prohibiting hate speech, including hate speech published and distributed online. Restricting expression-based abuse that directly targets and silences marginalized communities, or results in their members self-censoring, directly promotes and protects both freedom of expression *and* equality.

The platform liability context can be distinguished from circumstances that gave rise to much of the leading hate speech jurisprudence. This is due to the position of the intermediary, which is typically at least one step removed from the actual speaker or publisher at the centre of most case law in this area. The all-important layer of users whose expression is facilitated by online platforms must not be ignored in the equation, and precedents cannot necessarily be applied directly from speaker (or publisher) to platform. Establishing a platform liability regime will require considering issues such as:

- the risks of overbroad removal of legitimate, beneficial, or legal content;
- issues arising from potential privatized governance of user speech and public discourse; and
- how a platform liability framework would account for the wide range of platform companies, which vary widely by size, nature, purpose, audience, business model, and content, among other relevant factors.

Still, the reasoning and principles supporting the constitutionality of Canadian hate speech prohibitions remain highly relevant in the context of TFGBV. Law and context combine to justify legal reforms that would impose some degree of legal obligation or liability on digital platforms for TFGBV by a user. The most effective legal reforms would account for the distinct role of digital platforms in the Internet ecosystem, and as differentiated from the direct perpetrator of TFGBV, while simultaneously recognizing that digital platforms do play a facilitative role—and sometimes more—in the devastating and widespread perpetuation of TFGBV.

Guiding Priorities and Recommendations for Federal Action

This report provides 14 recommendations for federal action, including legislative reform. These recommendations are based on six guiding priorities that emerged from the research and analysis conducted in this report and should govern efforts to address TFGBV in Canadian law. These priorities are:

- recognizing a need for legal reform to address TFGBV, including through platform regulation;

- recognizing that Canadian constitutional law justifies imposing proportionate limits on freedom of expression in order to uphold and protect the rights to equality and freedom from discrimination, and also to give full effect to the core values underlying freedom of expression;
- guaranteeing that legal reforms that address TFGBV build in victim/survivor-centered, trauma-informed, and intersectional feminist perspectives;
- ensuring expedient, practical, and accessible remedies for those targeted by TFGBV;
- providing due process mechanisms to users who wish to contest platforms' content moderation decisions (whether a decision to leave up or take down content); and
- requiring transparency from platform companies regarding their content moderation policies and decisions, as well as the outcomes of such policies and decisions concerning TFGBV.

Recommendations for Federal Action

A. Centering Human Rights, Substantive Equality, and Intersectionality

1. Apply a principled human rights-based approach to platform regulation and platform liability, including giving full effect to the rights to equality and freedom from discrimination.
2. Ensure that legislation addressing TFGBV integrates substantive equality considerations and guards against exploitation by members of dominant social groups to silence expression by members of historically marginalized groups.
3. When pursuing legislative or other means of addressing TFGBV, consult substantively with and take into account the perspectives and lived experience of victims, survivors, and those broadly impacted by TFGBV.

B. Legislative Reforms

4. Establish a centralized expert regulator for TFGBV specifically, with a dual mandate: a) to provide legal remedies and support to individuals impacted by TFGBV on digital platforms, including regulatory and enforcement powers; and b) to develop research on TFGBV and provide training and education to the public, relevant stakeholders, and professionals.
5. Enact one or more versions of the current 'enabler' provision in subsections 27(2.3) and 27(2.4) of the *Copyright Act*, adapted to specifically address different forms of TFGBV, including 'purpose-built' platforms.
6. Enact a law that allows for victims/survivors of TFGBV to obtain immediate removal of certain clearly defined kinds of content from a platform *without* a court order, such as NCDII.
7. Ensure that legislation to address TFGBV focuses solely on TFGBV (including intersectional considerations)—do not dilute, compromise, or jeopardize the constitutionality of such legislation by 'bundling' TFGBV with other issues that the government may wish to also address through platform regulation.

C. Legal Obligations for Platform Companies

8. Require platform companies to provide to users *and non-users* clearly visible, easily accessible, plain-language complaint and abuse reporting mechanisms to expediently address and remedy instances of TFGBV.

9. For 'purpose-built', 'enabling', or otherwise TFGBV-dedicated platforms, and where a clearly delineated threshold of harm is met, provide that an order to remove specific content on one platform will automatically apply to any of that platform's parent, subsidiary, or sibling platform companies where the same content also appears.

10. Require platform companies to undergo independent audits (which could be conducted by the new TFGBV agency) and publish comprehensive annual transparency reports.

11. When determining legal obligations for digital platforms, account for the fact that platforms vary dramatically in size, nature, purpose, business model (including non-profit), extent of intermediary role, and user base.

D. Research, Education, and Training

12. Fund, make widely available, and mandate (where appropriate) education resources and training programs in TFGBV, which include information on how to support those who are subjected to TFGBV.

13. Fund frontline support workers and community-based organizations working to end, and supporting victims/survivors of, gender-based violence, abuse, and harassment, specifically to enhance their internal expertise, resources, and capacity to support those impacted by TFGBV (which often accompanies gender-based violence and abuse).

14. Fund further empirical, interdisciplinary, and law and policy research by TFGBV scholars, other TFGBV experts, and community-based organizations on TFGBV and the impacts of emerging technologies on those subjected to TFGBV.

1. Introduction and Overview

Technology-facilitated gender-based violence, abuse and harassment (collectively referred to as "TFGBV") encompasses a broad spectrum of harms perpetuated against women and girls, often on the basis that they are women or girls, or perceived as such. This includes both cis and trans women and people who present as femme or feminine online (or who are attributed this gender presentation), in addition to non-binary or otherwise gender-diverse individuals. TFGBV also targets women and girls based on intersecting marginalized identities, such as if they are members of the 2SLGBTQQIA community, are Black, Indigenous, or otherwise racialized, have a disability, or live in poverty.

TFGBV is unrelenting and can have devastating impacts on individuals who are targeted and victimized. These impacts include broader implications for the equality rights of women, girls, and members of other historically marginalized communities, and their ability to participate fully in society, meaningfully exercise their human rights, and benefit from equal protection of the law—beyond the bare minimum of living with baseline physical and psychological safety. Digital platforms, such as Facebook, Reddit, Google Search, Twitch, and Patreon, host and facilitate the vast majority of engagement in and access to social, political, and commercial online interactions, thus correspondingly host and facilitate extraordinary levels of online violence, abuse, and harassment aimed at women and girls. This report provides an examination and legal and policy analysis of the role that digital platforms play in TFGBV and how such platforms might be held accountable or liable for such harms in Canadian law.

1.1. Research Scope and Methodology

The overarching objective of this report is to provide a legal and policy analysis regarding the role of digital platforms in facilitating or exacerbating violence, abuse, and harassment against women and girls, with a view to assessing potential ways to hold digital platforms liable or accountable for such harms, in Canada. The research questions guiding this report are:

1. What current Canadian laws and regulations apply to online platforms in the context of technology-facilitated gender-based violence, abuse and harassment ("TFGBV"), and to what extent could such laws and regulations be used to hold online platforms accountable for TFGBV, and/or limit their ability to profit from TFGBV?
2. What international strategies have been developed to regulate online platform companies that enable, profit from or facilitate TFGBV on their platforms?
3. What does existing legal or academic research suggest about the best practices and solutions for holding online platforms accountable for their role in the proliferation of TFGBV?
4. What practices or solutions might be most effective in the Canadian context, and what barriers might exist to implementing those solutions?

Due to the sheer breadth of activities and behaviours involved in TFGBV, only some of which are currently explicitly illegal or criminalized in Canadian law, this report necessarily answers the above research questions across multiple areas of law and policy, and simultaneously confines itself to

addressing a specific subset of TFGBV, and is by no means exhaustive. Specifically, the analysis provided cuts across equality and non-discrimination law, human rights law, hate speech and freedom of expression law, privacy law, and intermediary liability law, as well as implicates criminal law and copyright law. Given the global nature of the Internet and the fact that other jurisdictions have already enacted new laws addressing digital platforms and harmful content specifically, including content that constitutes TFGBV, this report also provides a multijurisdictional perspective, looking to relevant law and policy in the United States, United Kingdom, European Union, Germany, Australia, and New Zealand to inform analysis and recommendations for reform in Canada.

The methodology engaged in for this report comprises desk research and legal research and analysis. Specifically, it examines current Canadian laws and laws of other jurisdictions for potential application or extension to different forms of TFGBV perpetuated through online platforms. The methodology also involved researching relevant academic literature, empirical studies, legal scholarship, policy reports, and civil society initiatives and materials, including those of digital rights organizations and frontline workers supporting victims and survivors of violence against women and girls, regarding the impacts of implemented laws and policies, or proposing legal and policy reforms. The legal analysis provided is interdisciplinary, informed by adjacent fields such as science and technology studies, technosociology, and research regarding the impacts and nature of violence and abuse against women and girls, including intimate partner and technology-facilitated violence, abuse, and harassment.

This report includes a large focus on speech-based or expression-based abuse constituting TFGBV (**speech-based TFGBV** or **expression-based TFGBV**)**,** e.g., verbal, written, or multimedia-based abuse, online harassment, hate speech, and threats. The types of TFGBV set out below receive less substantive treatment throughout the report, unless they are the focus of a specific law, though they will be referenced occasionally due to unavoidable overlap and the interrelated nature of different types of TFGBV. This is for reasons of time and scope, in addition to these areas of TFGBV having been extensively studied and written about by others in Canada and elsewhere (as indicated in footnotes):

- TFGBV that involves invasion of privacy and violating digital security (e.g., spying, hacking, covert surveillance, disclosure of personal data without consent);[1]

[1] See e.g., Rahul Chatterjee, et al, "The Spyware Used in Intimate Partner Violence" (Paper delivered at the 39th IEEE Symposium on Security and Privacy, San Francisco, 21 May 2018); Danielle Keats Citron, "Spying Inc" (2015) 72 Washington and Lee Law Review 1243; Cynthia Khoo, Kate Robertson & Ronald Deibert, "Installing Fear: A Canadian Legal and Policy Analysis of Using, Developing, and Selling Smartphone Spyware and Stalkerware Applications" (June 2019) at 124-46, online (pdf): *Citizen Lab* <https://citizenlab.ca/docs/stalkerware-legal.pdf>; Christopher Parsons, et al, "The Predator in Your Pocket: A Multidisciplinary Assessment of the Stalkerware Application Industry" (June 2019), online (pdf): Citizen Lab <https://citizenlab.ca/docs/stalkerware-holistic.pdf>; Molly Dragiewicz, et al, "Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms," (2018) 18:4 Feminist Media Studies 609; Diana Freed et al, "'A Stalker's Paradise': How Intimate Partner Abusers Exploit Technology" (Paper delivered at CHI 2018, Montreal, 21 April 2018); Adam Molnar & Diarmaid Harkin, "The Consumer Spyware Industry An Australian-based analysis of the threats of consumer spyware" (August 2019), online (pdf): *Australian Communications Consumer Action Network* <https://accan.org.au/files/Grants/2017%20successful%20projects/Deakin%20-%20Consumer%20Spyware%20Industry%20%20-%2030Jul19%20WEB.pdf>; Karen Levy, "Intimate Surveillance" (2015) 51 Idaho Law Review 679; and Diarmaid Harkin, Adam Molnar & Erica Vowles, "The commodification of mobile phone surveillance: An analysis of the consumer spyware industry" (2020) 16:1 Crime Media Culture 33; and Heather Doulas & Mark Burdon, "Legal Responses to Non-Consensual Smartphone Recordings in the Context of Domestic and Family Violence," (2018) 41:1 University of New South Wales Law Journal 157.

- TFGBV that involves sexualized harassment and abuse (e.g., non-consensual distribution of intimate images, sextortion, luring, child sexual abuse material);[2]
- TFGBV that involves harming someone's reputation or compromising their social media presence (e.g., defamation, impersonation);[3] and
- TFGBV that involves direct censorship and silencing, as opposed to indirect silencing through other kinds of TFGBV (e.g., falsely reporting accounts for violations of platforms' terms and conditions, shadowbanning, wrongful automated takedowns, discriminatory content moderation policies, biased algorithmic rankings).[4]

This report also focuses exclusively on digital platforms *in the role of online intermediary*. This means that the analysis does not include examining the liability of the direct wrongdoer who is using the platform (such as the actual stalker or actual harasser), nor will it include situations where the digital platform is straightforwardly the direct perpetrator of an offence. Rather, the focus is on circumstances where a digital platform's users are directly responsible for the violence, abuse, or harassment, and this report investigates the legal and policy issues that arise where the proposition is holding the digital platform liable for the contents and actions of its users. This also means that other kinds of online intermediaries that are not digital platforms—as defined in Section 3.3.1 ("What Are Digital Platforms?")—are not the focus of this report, including Internet service providers (ISPs), mobile wireless service providers, and cloud computing providers.

1.2. Report Outline

Part 2 discusses in detail the characteristics and impacts of TFGBV as enacted through online *speech,* or *expression*. This section details common terminology and behaviours associated with TFGBV and examines the specific experiences of women, girls, and intersecting marginalized identities online, including the impacts of TFGBV and its consequences for impacted individuals on a personal level, as well as its broader systemic, political and democratic repercussions for women's human rights and

---

[2] See e.g., Emily Laidlaw & Hilary Young, "Creating a Revenge Porn Tort for Canada" (2020) Supreme Court Law Review 147; Andrea Slane & Ganaele Langlois, "Debunking the Myth of 'Not My Bad': Sexual Images, Consent, and Online Host Responsibilities in Canada" (2018) 30:1 Canadian Journal of Women and the Law 42; Nicola Henry & Asher Flynn, "Image-Based Sexual Abuse: Online Distribution Channels and Illicit Communities of Support" (2019) 25:16 Violence Against Women 1932; Suzie Dunn & Alessia Petricone-Westwood, "More than 'Revenge Porn': Civil Remedies for the Non-consensual Distribution of Intimate Images" (Paper delivered at the 38th Annual Civil Litigation Conference, Mont Tremblant, QC, 16 November 2018); Danielle Keats Citron & Mary Anne Franks, "Criminalizing Revenge Porn" (2014) 49 Wake Forest Law Review 345; and Moira Aikenhead, "Non-Consensual Disclosure of Intimate Images as a Crime of Gender-Based Violence" (2018) 30 Canadian Journal of Women and the Law 117.

[3] See e.g., Law Commission of Ontario, "Defamation Law in the Internet Age: Final Report (March 2020), online (pdf): <https://www.lco-cdo.org/wp-content/uploads/2020/03/Defamation-Final-Report-Eng-FINAL-1.pdf>; Emily B Laidlaw & Hilary Young, "Internet Intermediary Liability in Defamation" (2018) 56:1 Osgoode Hall Law Journal 112; Suzie Dunn, "Identity Manipulation: Responding to advances in artificial intelligence and robotics" (Paper delivered at We Robot 2020, Ottawa, 2 April 2020) [unpublished]; and Jane Bailey & Valerie Steeves, "Defamation Law in the Age of the Internet: Young People's Perspectives" (June 2017), online (pdf): *Law Commission of Ontario* <http://www.lco-cdo.org/wp-content/uploads/2017/07/DIA-Commissioned-Paper-eQuality.pdf>.

[4] See e.g., Danielle Blunt et al, "Posting Into the Void" (2020), online (pdf): *Hacking//Hustling* <https://hackinghustling.org/wp-content/uploads/2020/09/Posting-Into-the-Void.pdf>; and "News and Analysis", online: OnlineCensorship.org <https://onlinecensorship.org/news-and-analysis>.

freedoms. Part 2 provides examples of platformed TFGBV in Canadian civil and criminal case law and explains how TFGBV particularly targets women who have a public presence or high visibility online.

Part 3 explains how digital platforms' business models, design decisions, and technological affordances optimize them for abusive speech and behaviour by users; and presents characteristics unique to *platformed* misogyny that arise from the platform-mediated nature of technology-facilitated violence, abuse, and harassment. Part 3 then canvasses several key types of content moderation policies and practices that digital platforms rely on to address technology-facilitated gender-based violence, abuse, and harassment among their users, including brief assessments and critiques on the effectiveness of such policies and practices.

Part 4 provides a systematic review of federal and provincial laws in Canada that involve either platform liability for user expression or conduct, or address some form of TFGBV. This includes federal copyright and criminal law, provincial laws regarding non-consensual distribution of intimate images, an intermediary liability regime in Quebec, and common law intermediary liability principles. Part 4 also reviews two platform liability frameworks that have been substantively developed and proposed, in the specific contexts of NCDII and defamation. This part concludes with a review of a number of laws that address neither intermediary liability nor TFGBV, but could provide grounds to hold platform companies liable for TFGBV on a systemic or institutional level. These laws include statutory human rights law, corporate negligence, commercial host liability (analogously), and product liability.

Part 5 provides a jurisdictional scan of different platform liability regimes and proposed legislative reforms around the world that either were put forward specifically to address TFGBV or could be applied to that end. Part 5 reviews key platform liability legislation and policy proposals in the United States, Germany, the United Kingdom, the European Union, Australia, and New Zealand. The overview of each jurisdiction's activities includes critiques and assessments of their effectiveness and impacts.

Part 6 highlights critical issues in efforts to hold digital platforms liable for technology-facilitated violence, abuse, and harassment enacted by their users. Specifically, the first half of this section discusses factors relevant to a constitutional analysis of a law purporting to regulate user expression through an intermediary platform, with a focus on the role of the right to equality in assessing the proportionality of a limitation on freedom of expression. The second half of Part 6 discusses issues associated with the unique role and position of digital platforms in society, and potential legal implications, such as wrongful removal of legitimate, beneficial, or lawful content, and the potential establishing of privatized regulation of public discourse. Part 6 concludes by briefly touching on several additional challenges that may arise in the course of legal reform to address platformed TFGBV.

Part 7 provides a suite of recommendations that emerged from the research and analysis presented in Parts 1 through 6. The recommendations focus on federal law reform and are thus aimed at the federal government for implementation. They are grouped into the following categories: centering human rights and substantive equality; specific legislative reforms to enact a TFGBV-specialized regulator; legal obligations to place on digital platforms; and funding for research, training, and education. The recommendations are animated by six overarching priorities, with the first and foremost being to ensure that any reforms centre a principled human rights-based approach and emphasize substantive equality, while applying an intersectional lens. Such reforms must also centre the experiences and needs of victims/survivors and those who have been negatively impacted by TFGBV.

2. Technology-Facilitated Gender-Based Violence, Abuse, and Harassment

Technology-facilitated gender-based violence, abuse, and harassment (collectively referred to as TFGBV) encompasses a formidable variety of activities and behaviours.[5] Conduct and expression constituting TFGBV has been widely documented in the media and academic literature; in empirical research and legal scholarship;[6] among community-based organizations;[7] frontline support workers, and activists advocating for gender equality and human rights; and through women's personal stories shared online and otherwise. TFGBV impacts women, girls, and gender-diverse individuals throughout nearly all spheres of private and public life in Canada and around the world.

This part of the report will provide, first, an introduction to what kinds of specific activities and behaviours constitute TFGBV, and their impacts on women, girls, and gender-diverse individuals (in Section 2.1), and second, a discussion of TFGBV as specifically perpetrated through online expression on digital platforms (in Section 2.2).

2.1. Introduction to TFGBV

The overview of TFGBV below will proceed as follows. Section 2.1.1 will provide a narrative glossary of specific TFGBV-related terminology used to describe common activities that constitute TFGBV. Section 2.1.2 will briefly discuss TFGBV in the context of intimate partner and dating violence. Section 2.1.3 will examine the impacts of TFGBV on women, girls, and gender-diverse individuals. The effects of TFGBV includes intersectional impacts on those who belong to more than one historically marginalized group, and are thus subjected to uniquely intersecting forms of systemic oppressions, which is also reflected in the TFGBV they experience. Section 2.1.4 will conclude with a discussion of considerations regarding terminology used throughout this report.

---

[5] See e.g., "ICT [information and communications technology] may be used directly as a tool for making digital threats and inciting gender-based violence, including threats of physical and/or sexual violence, rape, killing, unwanted and harassing online communications, or even the encouragement of others to harm women physically. It may also involve the dissemination of reputation-harming lies, electronic sabotage in the form of spam and malignant viruses, impersonation of the victim online and the sending of abusive emails or spam, blog posts, tweets or other online communications in the victim's name. ICT-facilitated violence against women may also be committed in the work place or in the form of so-called 'honour-based' violence or of domestic violence by intimate partners." Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 31 (footnotes omitted). See also Suzie Dunn, "Technology-Facilitated Gender-Based Violence: An Overview" (December 2020), online (pdf): *Centre for International Governance Innovation* <https://www.cigionline.org/sites/default/files/documents/SaferInternet_Paper%20no%201_0.pdf>.

[6] See e.g., "Research Publications", online: *eQuality Project* <http://www.equalityproject.ca/research/research-publications/>.

[7] See e.g., "Technology Safety", online: *BC Society of Transition Houses* <https://bcsth.ca/projects/technology-safety/>.

2.1.1. What Is TFGBV?

Technology-facilitated gender-based violence, abuse, and harassment is part of the continuum of violence, abuse, and harassment that women and girls face in the world regardless of technology, "whether it be physical abuse, or sexual assault […] [gender-based] violence is wielded as a tool to control and have power over women, to maintain men's dominance over women as a class, and to reinforce patriarchal norms, roles and structures."[8] To be clear, this report includes trans women when referring to women, and TFGBV, as implied in the term, also impacts trans men and nonbinary individuals. TFGBV is rooted in, arises from, and is exacerbated by misogyny, sexist norms, and rape culture, all of which existed long before the Internet. However, TFGBV, in turn, accelerates, amplifies, aggravates, and perpetuates the enactment of and harm from these same values, norms, and institutions, in a vicious cycle of technosocial systemic oppression.[9] In many cases, abuse and harassment that begins online directly leads to violence and abuse in the physical world, such as when women are stalked, followed, or attacked, or are threatened to the extent of needing to move homes or move schools.[10] In fact, one of the central components of TFGBV is that it collapses what is now recognized as a false dichotomy between 'online' and 'offline' worlds—both constitute 'real life' and are increasingly interwoven with and inseparable from each other.[11] The UN Special Rapporteur on violence against women, its causes and consequences, Dubrava Šimonovic has noted:

> It is therefore important to acknowledge that the Internet is being used in a broader environment of widespread and systemic structural discrimination and gender-based violence against women and girls, which frame their access to and use of the Internet and other ICT [information and communications technology]. Emerging forms of ICT have facilitated new types of gender-based violence and gender inequality in access to technologies, which hinder women's and girls' full enjoyment of their human rights and their ability to achieve gender equality.[12]

Similarly, the House of Commons Standing Committee on the Status of Women in Canada has recognized that TFGBV "while enabled by ICTs and social media, are rooted in larger social and cultural problems—including sexism and misogyny—that contribute to violence against young women and girls in the offline world."[13]

---

[8] Jessica West, "Cyber-Violence Against Women" (May 2014) at 2, online (pdf): *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf>.

[9] See e.g., *ibid* at 16, Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2008) 18:4 Feminist Media Studies 609 at 611; Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 14 and 20.

[10] Jessica West, "Cyber-Violence Against Women" (May 2014) at 16, online (pdf): *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf>.

[11] Anastasia Powell & Nicola Henry, *Sexual Violence in a Digital Age* (London: Palgrave Macmillan UK, 2017) ch 3 at 49.

[12] Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 14.

[13] House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 32 (Chair: Marilyn Gladu).

TFGBV can include a wide range of specific behaviours that occur on digital platforms. TFGBV on digital platforms, or platformed TFGBV, is often collectively or generically referred to as online harassment or online abuse, which can also include online sexual harassment or sexualized online violence, abuse, or harassment. In addition to encompassing the specific activities described in the rest of this section, **online abuse** additionally involves verbally and emotionally abusing someone online, such as insulting and harassing them, their work, or their personality traits and capabilities, including telling that person she should commit suicide or deserves to be sexually assaulted. **Online harassment** describes persistently engaging with someone online in a way that is unwanted, often (but not necessarily) with a view to causing distress or inconvenience to that person. This is particularly the case where the harassment is frequent or voluminous, whether it comes from one person ongoingly or from an ongoing stream of harassers acting on their own accord or under a coordinated campaign deliberately targeting the victim.[14] The term **online violence** may also encompass some forms or situations of online harassment or online abuse, including the specific types of activities described below.

Another overarching component of TFGBV is that the violence, abuse, and harassment that women receive online is often sexualized, whether as **online sexual harassment** or other forms of sexualized interactions that are perpetrated without consent across digital platforms. Such TFGBV may include reference to the targeted person's sexuality or sexual activity, sexualized insults and harassment, or shaming the person for their sexuality or for engaging in sexual activity ('slut-shaming').[15] Due to the broader historical and ongoing sociocultural context, TFGBV that involves sexualized content or interactions, in addition to sexual violence, abuse, and harassment that occurs in person, is a particularly gendered phenomenon that is routinely weaponized against women and girls, in both virtual and physical environments.

Many behaviours that fall under TFGBV or platformed TFGBV are forms of **speech/expression-based abuse** or **speech/expression-based TFGBV**—violent, abusive, or harassing conduct enacted through written, audio, image- or video-based, or otherwise multimedia-based expression online. This report will use 'speech-based' and 'expression-based' interchangeably, as 'expression' is more accurately encompassing and is the relevant term in Canadian law, while recognizing that 'speech' may be the more relevant or common term in the United States, among the major platforms themselves which are the focus of this report, and for others in the platform regulation and TFGBV field.

Some perpetrators may post statements or other content that conveys such misogynistic or harmful attitudes towards women, girls, and other marginalized identities, that they meet the legal definition for **hate speech**. Speech-based TFGBV also includes sending **threats** to targeted individuals, including **rape threats**, **death threats**, or threats to harm the targeted person's family and friends. **Trolling** is another form of platformed TFGBV, which occurs when users post messages, images, videos, or otherwise online content, or create online campaigns such as through hashtags on Twitter, "for the

---

[14] See e.g., Suzie Dunn, "Technology-Facilitated Gender-Based Violence: An Overview" (2020) at 7, online (pdf): *Centre for International Governance Innovation* <https://www.cigionline.org/sites/default/files/documents/SaferInternet_Paper%20no%201_0.pdf>.

[15] "Online sexual harassment refers to any form of online unwanted verbal or nonverbal conduct of a sexual nature with the purpose or effect of violating the dignity of a person, in particular by creating an intimidating, hostile, degrading, humiliating or offensive environment." Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 40.

purpose of annoying, provoking or inciting violence against women and girls. Many 'trolls' are anonymous and use false accounts to generate hate speech".[16]

TFGBV can also include a number of abusive behaviours that violate women's right to privacy, including **sexual privacy**.[17] The most commonly known example of this kind of TFGBV is **non-consensual distribution of intimate images (NCDII)**: circulating intimate or sexual images or recordings of someone without their consent, such as where the person is nude, partially clothed, or engaged in sexual activity, often "with the purpose of shaming, stigmatizing or harming the victim".[18] NCDII may also be known as **image-based sexual exploitation** or fall into the broader category of **image-based abuse**.[19] If someone attempts to sexually extort another person by capturing sexual or intimate images or recordings of them and threatening to distribute them without consent *unless* the targeted person pays the perpetrator, follows their orders, or commits sexual acts with or for them, the abuse is often referred to as **sextortion**.[20] Such images or recordings may have been obtained or created with the knowledge and consent of the targeted person—which does not constitute additional consent to distribution. They may also have been captured through other illegal acts such as technology-facilitated **voyeurism**, a criminal offence that involves surreptitiously observing or recording someone while they are in a situation that gives rise to a reasonable expectation of privacy.[21] This includes spying on someone engaged in sexual activity or in an intimate setting (e.g., their bedroom) by illicitly accessing their webcam or their phone camera, without their consent or knowledge.[22]

---

[16] Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 37.

[17] Danielle Citron, "Sexual Privacy" (2019) 128 Yale Law Journal 1870.

[18] Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 41. NCDII has in the past been referred to as 'revenge porn'. However, this is an inaccurate and harmful term as it embeds the incorrect notions that (a) that the victim has done something 'wrong' for which the perpetrator is seeking 'revenge'; and (b) that the content is in fact pornography, i.e., sexually explicit imagery or video created for consumption and sexual arousal, rather than what it is at core: an act of misogynistic violence, power, and control. Sophie Gallagher, "'Revenge Porn' Is Not The Right Term To Describe Our Experiences, Say Victims", *Huffington Post* (3 August 2020), online: <https://www.huffingtonpost.co.uk/entry/why-are-we-still-calling-it-revenge-porn-victims-explain-change-in-the-laws-needed_uk_5d3594c2e4b020cd99465a99?>.

[19] Josh Taylor, "Don't call it 'revenge porn', victims' groups say", *Crikey* (15 January 2016), online: https://www.crikey.com.au/2016/01/15/dont-call-it-revenge-porn-victims-groups-say/>.

[20] Jessica West, "Cyber-Violence Against Women" (May 2014) at 10, online (pdf): *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf> (footnotes omitted); see also House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 34 (Chair: Marilyn Gladu).

[21] *R v Jarvis*, 2019 SCC 10 at para 1; *Criminal Code*, RSC 1985, c C-46, s 162(1). In its intervener factum in *Jarvis*, LEAF argued that what constitutes a reasonable expectation of privacy must be determined contextually and that women and girls cannot be considered to have abandoned all privacy rights when they enter a public or semi-public space where they may be observed. "LEAF celebrates Supreme Court of Canada ruling in R. v. Jarvis" (February 2019), online: *LEAF* <https://www.leaf.ca/news/leaf-celebrates-supreme-court-of-canada-ruling-in-r-v-jarvis/>. The Court recognized and incorporated this understanding of the right to privacy in its decision: *R v Jarvis*, 2019 SCC 10 at paras 60-61.

[22] "Voyeurism is a heavily gendered crime, with the majority of complainants being women and children and the majority of perpetrators being men. It can occur in private spaces, such as a changeroom or bedroom where a camera might be hidden; but it can also happen in public spaces, when someone photographs an unsuspecting person from a drone or uses a cellphone to secretly record videos of a woman walking." Kristen Thomasen & Suzie Dunn, "The Supreme Court's ruling on a voyeurism case contributes to a broader conversation about surveillance and privacy in public and semi-public spaces.",

Another form of platformed TFGBV that engages women's privacy is **doxing**, which involves publicly disclosing someone's personal information online, such as their full name, home address, and social insurance number.[23] Doxing is particularly concerning for individuals who are, for example, in or escaping from situations of intimate partner violence, or who use pseudonyms due to living in repressive regimes or to avoid harmful discrimination for aspects of their identity, such as being transgender or a sex worker.[24] Someone's personal information and sensitive data may be obtained illicitly through **interception of private communications** (a criminal offence), which can involve surreptitiously hacking into a person's devices or online accounts and obtaining their personal data, including the contents and metadata of text messages, social media activity, browsing history, call logs, photos and videos, and other forms of private information.[25] Some mobile apps, known as **spyware** or **stalkerware**, are designed and marketed for the purpose of enabling their customers to systematically **spy on, monitor, and track** intimate partners or former partners through their mobile phones, after covertly installing the software. Ostensibly "non-malicious" apps with similar features, usually advertised for child or employee monitoring, are also routinely repurposed into spyware and stalkerware. Such apps facilitate spying on and intimately monitoring someone's private communications and private online activities, such as reading their emails and private social media messages, monitoring their text messages and phone calls, tracking their real-time location, or checking their browser history. [26]

TFGBV on digital platforms may also take the form of attempts to ruin the targeted individual's public image or reputation among their friends, family, peer group, coworkers, professional community, and/or other social networks and communities to which they belong. These activities may fall under the broad, albeit not all-encompassing, category of **defamation**: lying about or misrepresenting an

---

Policy Options (25 February 2019), online: <https://policyoptions.irpp.org/magazines/february-2019/court-ruling-voyeurism-broad-social-impact/>. In *R v Trinchi*, 2019 ONCA 356, the court found the defendant to have committed voyeurism when he captured and saved screenshots of his long-distance intimate partner, without her knowledge or consent, during consensual Skype videocall sessions where she appeared nude.

[23] House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 34 (Chair: Marilyn Gladu) ("[Doxing] has commonly been used against women, at times because they opposed sexism or turned down sexual advances online"). See also: "'Doxing' refers to the publication of private information, such as contact details, on the Internet with malicious intent, usually with the insinuation that the victim is soliciting sex (researching and broadcasting personally identifiable information about an individual without consent, sometimes with the intention of exposing the woman to the "real" world for harassment and/or other purposes). It includes situations where personal information and data retrieved by a perpetrator is made public with malicious intent, clearly violating the right to privacy." Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 36.

[24] See e.g., "Facebook's 'real name' policy hurts real people and creates a new digital divide", *Guardian* (3 June 2015), online: <https://www.theguardian.com/commentisfree/2015/jun/03/facebook-real-name-policy-hurts-people-creates-new-digital-divide>.

[25] Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 45.

[26] See generally Cynthia Khoo, Kate Robertson & Ronald Deibert, "Installing Fear: A Canadian Legal and Policy Analysis of Using, Developing and Selling Smartphone Spyware and Stalkerware Applications" (June 2019), online (pdf): *Citizen Lab* <https://citizenlab.ca/docs/stalkerware-legal.pdf> and Christopher Parsons et al, "Predator in Your Pocket: A Multidisciplinary Assessment of the Stalkerware Application Industry" (June 2019), online (pdf): *Citizen Lab* <https://citizenlab.ca/2019/06/the-predator-in-your-pocket-a-multidisciplinary-assessment-of-the-stalkerware-application-industry/>.

individual online to ruin their reputation and relationships, including referencing their sexuality or sexual activity. Perpetrators of TFGBV may also engage in **impersonation** of the targeted individual, whether through hacking into and taking over their social media accounts, or creating fake social media accounts purporting to be the victim.[27] (Note that this does not include similar activity that would constitute satire or parody of powerful male public figures.) TFGBV in the form of **identity manipulation**[28] has further increased with the rise of **image manipulation** achieved through **deepfakes**, which is the use of artificial intelligence to produce videos that include false but realistic images of an individual saying something they did not say or doing something they did not do.[29] Approximately 96% of deepfakes online today involve manipulating a pornographic video to replace the actress's face with the face of an ex-partner, celebrity, or another real woman, creating what looks like real pornography featuring that person, without their consent.[30] Image manipulation does not require deepfakes, however, which are created through artificial intelligence and machine learning algorithms. Videos, images, or audio recordings manipulated without the use of artificial intelligence (such as through Photoshop edits or basic video editing software) may be known as **cheap fakes** or **shallow fakes**.[31] Image manipulation, used against a specific person, may also be considered a form of **image-based abuse**.

Several types of platformed TFGBV involve coordinated or collective action on the part of those engaging in the abusive behaviour, sometimes exploiting platform features or unwitting actors such as other Internet users or law enforcement. **Online mobbing** or **swarming**[32] is what occurs when large numbers of people simultaneously engage in online harassment or online abuse against a single individual. These events may involve a small group of actors who planned and coordinated the mobbing, with other individuals joining in either knowingly or being misled into piling on without awareness of the full context. **Coordinated flagging** involves gaming a platform's mechanisms for reporting abuse, and comprises organized activity where a large group of individuals 'flag' or report someone's post for removal or account suspension, claiming it is a violation of the platform's community standards or terms of use, as a way to silence the target or cause them harm or inconvenience.[33] **Brigading** is a tactic used to manipulate social media algorithms that determine what

---

[27] Suzie Dunn, "Technology-Facilitated Gender-Based Violence: An Overview" (December 2020), at 15, online (pdf): *Centre for International Governance Innovation* <https://www.cigionline.org/sites/default/files/documents/SaferInternet_Paper%20no%201_0.pdf>.

[28] Suzie Dunn, "Identity Manipulation: Responding to advances in artificial intelligence and robotics" (Paper delivered at We Robot 2020, Ottawa, 2 April 2020) [unpublished].

[29] Danielle K Citron & Robert Chesney, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" (2019) 107 California Law Review 1753.

[30] Suzie Dunn, "Identity Manipulation: Responding to advances in artificial intelligence and robotics" (Paper delivered at We Robot 2020, Ottawa, 2 April 2020), at 10 [unpublished].

[31] Britt Paris & Joan Donovan, "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence" (2019), online (pdf): Data & Society <https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf>.

[32] Ben Collins & Brandy Zadrozny, "Twitter Bans 7,000 QAnon accounts, limits 150,000 others as part of broad crackdown", *NBC News* (21 July 2020), online: <https://www.nbcnews.com/news/amp/ncna1234541>.

[33] See e.g. "As Fiore-Silfvast […] describes, a group of bloggers angered by the presence of pro-Muslim content on YouTube began an effort called 'Operation Smackdown.' Launched in 2007 and active as recently as 2011, the group coordinated their supporters to flag specific YouTube videos under the category of 'promotes terrorism' (a submenu under 'violent repulsive content' [...]). They offered step-by-step instructions on how to flag content, set up playlists on YouTube of the videos they wanted to target, and added a Twitter feed announcing a video to be targeted that day. Participating bloggers would celebrate the number of targeted videos that YouTube removed, and would lambast YouTube and Google for allowing others

content is promoted across users' feeds and what content is suppressed by appearing lower and less likely to be viewed. Users engage in brigading to "amplify harassment by […] boosting harmful content in ways that make it seem more relevant to the algorithm, which can place it higher in search results or make it more likely to be delivered to audiences as a trending topic".[34] The reverse can also occur, where abusers may orchestrate mass 'downvoting' of posts by specific women, to prevent their words from reaching a wider audience.[35] **Swatting**, named after police Special Weapons and Tactics (SWAT) teams, involves "calling 911 and lying about someone doing something really bad, like holding a hostage, to get dispatchers to send police officers—and particularly a SWAT team—to a victim's location."[36] People have been killed by police as a result of swatting,[37] and the practice is even more dire when placed within the context of police brutality, shootings, and excessive use of force with respect to members of Black, Indigenous, and other racialized communities.[38] Swatting as a form of TFGBV can involve threatening to swat a woman or girl unless she complies with a request, such as sending nude photos,[39] or swatting a woman due to dislike for her politics or other views expressed online.

TFGBV may also include sexual assault or sexual exploitation that is facilitated or aggravated by the use of digital platforms in relation to the assault or exploitation. For example, **technology-aggravated sexual assault** describes sexual assault with an online component, such as men or teenage boys filming themselves raping a woman or girl and then posting the video on social media.[40] When this occurs, technology and social media are used "both during the sexual assault to record or take pictures of the

---

to remain." Kate Crawford & Tarleton Gillespie, "What is a flag for? Social media reporting tools and the vocabulary of complaint" (2014) 18:3 New Media & Society 410 at 421 (footnotes omitted).

[34] *House of Commons FEWO Committee Meeting*, 42nd Parl, 1st Sess, No 37 (5 December 2016) at 1535 (Matthew Johnson).

[35] "Manipulating algorithms in this way can also be used to essentially silence victims of [TFGBV], especially in platforms that allow for downvoting content as well as upvoting." *Ibid*.

[36] German Lopez, "Swatting, the horrible 'prank' that's hit gamers, Justin Bieber, and many more, explained", *Vox* (11 March 2016), online: <https://www.vox.com/2016/3/11/11196282/what-is-swatting-video>.

[37] Michael Brice-Saddler, Avi Selk & Eli Rosenberg, "Prankster sentenced to 20 years for fake 911 call that led police to kill an innocent man", *Washington Post* (29 March 2019), online: <https://www.washingtonpost.com/nation/2019/03/29/prankster-sentenced-years-fake-call-that-led-police-kill-an-innocent-man/>.

[38] "Swatting, in which a 911 caller falsely reports a life-threatening crime so heavily armed tactical units will swarm an innocent person's house, 'has disproportionately targeted communities of color, the LGBTQ community, and religious communities,' [...]" Abigail Hauslohner, Maria Sacchetti & Shayna Jacobs, "Incidents of calling police on black people lead states to consider new laws", *Philadelphia Inquirer* (28 May 2020), online: <https://www.inquirer.com/news/nation-world/states-legislation-racist-calls-new-york-new-jersey-oregon-washington-20200528.html>.

[39] German Lopez, "Swatting, the horrible 'prank' that's hit gamers, Justin Bieber, and many more, explained", *Vox* (11 March 2016), online: <https://www.vox.com/2016/3/11/11196282/what-is-swatting-video>.

[40] "There have been a number of other stories of boys sexually assaulting unconscious girls and recording the assault on cell phones, through picture and video. We can list them off: in November 2011, 15 year old Rehtaeh Parsons was gang-raped by 4 classmates who took pictures and distributed them to her classmates in Nova Scotia; in January 2012, 14 year old Daisy Coleman was raped by a senior on the high school football team, Matthew Barnett, in Maryville, Missouri, while another boy filmed it; also in 2012, 15 year old Audrie Pott was sexually assaulted by three boys who took pictures and distributed them to their peers in Saratoga, California; Savannah Dietrich of Louisville, Kentucky, was 16 years old when she was sexually assaulted by two boys who also took pictures that she didn't find out about until a month later; and in June 2013, a 21 year old woman went on a date and was gang-raped by 4 football players who captured the assault on cell phone cameras, in a dorm room at Vanderbilt University in Nashville, Tennessee." Jessica West, "Cyber-Violence Against Women" (May 2014) at 5-6, online (pdf): *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf> (footnotes omitted).

assault and after the assault, as a way to revictimize, humiliate and intimidate survivors" as well as "used by community members to bombard [victims/survivors] with threats and abuse to try to keep them from reporting and to shame them" and by the perpetrators "to undermine and discredit the girls' stories after the assault".[41] **Luring for sexual exploitation** also occurs on digital platforms, where predators may find and groom young women or girls through social media and various chat platforms, or post false advertisements online, in order to lure them into 'offline' forms of sexual exploitation (e.g., sex trafficking and child sexual abuse).[42]

TFGBV is a gendered phenomenon that disproportionately impacts women and girls, reflecting and perpetuating their inequality in society beyond and prior to the existence of the Internet.[43] Research by the eQuality Project found that out of 114 Canadian criminal law decisions in 2017 that involved technology-facilitated violence, 90 identified the victim as a woman or girl, and 106 involved a male defendant.[44] A 2018 survey on gender-based violence and unwanted sexual behaviour in Canada found that women were more likely than men to have "experienced an unwanted behaviour that made them feel unsafe or uncomfortable in a virtual space in the past 12 months", and to have been "pressured to send, share, or post sexually suggestive or explicit images or messages".[45] In addition, young women were "twice as likely as their male counterparts to say someone on a dating site or app has called them an offensive name (44% vs. 23%) or threatened to physically harm them (19% vs. 9%)."[46] Research in other jurisdictions, such as Australia and the European Union, similarly reflect the disproportionate impact of TFGBV on women and girls, both in the frequency and intensity of the abuse as well as in the

---

[41] *Ibid* at 6.

[42] See e.g., "Another form of [TFGBV] is the luring and online exploitation of minors by adults. In these cases, adults share existing or self-produced sexual images of children (also referred to as child pornography) or communicate with children over the Internet for the purpose of committing a sexual offence or trafficking." House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 34 (Chair: Marilyn Gladu); see also Dubravka Šimonović , *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 32.

[43] "The consequences of and harm caused by different manifestations of online violence are specifically gendered, given that women and girls suffer from particular stigma in the context of structural inequality, discrimination and patriarchy. Women subjected to online violence are often further victimized through harmful and negative gender stereotypes, which are prohibited by international human rights law." Dubravka Šimonović , *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 25; Jane Bailey, Valerie Steeves & Suzanne Dunn, *Submission to The Special Rapporteur on Violence Against Women Re: Regulating Online Violence and Harassment Against Women* (September 2017) at para 13e, online: *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2017/12/Bailey-Steeves-Dunn-Submission-27-Sep-2017.pdf>.

[44] Jane Bailey, Valerie Steeves & Suzanne Dunn, *Submission to The Special Rapporteur on Violence Against Women Re: Regulating Online Violence and Harassment Against Women* (September 2017) at para 7, online: *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2017/12/Bailey-Steeves-Dunn-Submission-27-Sep-2017.pdf>. One additional case involved a co-accused man and woman. The authors note, "Given the under-reporting of sexual violence, as well as the fact that many criminal law decisions are not made available online, this body of cases almost certainly represents only a fraction of instances of TFVAWG in Canada."

[45] Statistics Canada, "Gender-based violence and unwanted sexual behaviour in Canada, 2018: Initial findings from the Survey of Safety in Public and Private Spaces", by Adam Cotter & Laura Savage, in *Juristat*m Catalogue No 85-002-X (Ottawa: Statistics Canada, 2019).

[46] Monica Anderson & Emily A Vogels, "Young women often face sexual harassment online – including on dating sites and apps", *Pew Research Center* (6 March 2020), online: <https://www.pewresearch.org/fact-tank/2020/03/06/young-women-often-face-sexual-harassment-online-including-on-dating-sites-and-apps/>.

severity of social, mental, emotional, economic, and democratic repercussions that result from such abuse.[47] However, the House of Commons Standing Committee on the Status of Women explains in the context of violence against women and girls generally:

> Measuring violence against young women and girls in Canada is challenging as such violence is widely underreported for a number of reasons. Firstly, the law enforcement and justice systems have not earned the trust and confidence of survivors of gender-based violence because of long-standing failures and inaction in many past cases of gender-based violence. Furthermore, because of a pervasive culture of victim blaming, victims may internalize feelings of shame and self-blame and may avoid reporting for fear [of] re-victimization. As well, in situations where girls are victims of violence, they may be too young to be capable of making a report.[48]

Thus, statistics likely reflect only a fraction of the true extent of the problem.

2.1.2. TFGBV in Intimate Partner and Dating Violence

TFGBV regularly occurs within the context of dating and intimate partner violence, abuse, and harassment.[49] According to a 2017 national survey of transition houses and women's shelters across Canada, respondents reported 18 forms of technology-enabled abuse among those who sought help at their organizations, including (rounded to nearest whole number): sending threats and intimidating messages (93%); tracking the person's location through their phone, GPS, or another location service (66%); impersonating the person through their email or online profiles (62%); hacking into social media, email, or utilities accounts (62%); monitoring online activities and exfiltrating data (43%); tracking or monitoring the woman through devices that the abuser gave to their children as gifts (28%); and installing spyware or keyloggers (21%).[50] In fact, the phenomenon known as 'Gamergate'—what has arguably become a byword for online violence against women and, more broadly, for sexism and

---

[47] See e.g., Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2008) 18:4 Feminist Media Studies 609 at 612; Molly Dragiewicz et al, "Domestic violence and communication technology: Survivor experiences of intrusion, surveillance, and identity crime" (July 2019) at 9, online (pdf): *Australian Communications Consumer Action Network* <https://accan.org.au/files/Grants/20190823%20Domestic%20violence%20and%20communication%20technology%20victim%20experiences%20of%20intrusion%20surveillance%20and%20identity%20theft.pdf>; Jessica West, "Cyber-Violence Against Women" (May 2014) at 4, online: *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf>; and House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 32 (Chair: Marilyn Gladu).

[48] House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 3 (Chair: Marilyn Gladu).

[49] "A recent survey on online harassment in the United States found that the most common perpetrators of digital abuse and stalking are current and former partners." Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2008) 18:4 Feminist Media Studies 609 at 613 (inline citations omitted).

[50] Women's Shelters Canada, "Shelter Voices" (June 2017) at 3, online: *Women's Shelters Canada* <https://endvaw.ca/wp-content/uploads/2017/06/shelterVoices_ENG_2017_WEB.pdf>.

misogyny in the gaming and technology sectors[51]—was instigated in the first instance by an angry male ex-partner of one of the targeted women.[52]

Dragiewicz et al use the term **technology-facilitated coercive control (TFCC)** to refer to TFGBV in the intimate partner context, which "encompass[es] the technological and relational aspects of abuse in the specific context of coercive and controlling intimate relationships."[53] In TFCC, perpetrators use social media and other digital platforms and communications technologies to intimidate, isolate, and control their partners or former partners, including leveraging their own social networks to target the victim/survivor, while threatening, co-opting, and undermining the victim/survivor's own social networks as a means of further control and isolation.[54] Like TFGBV generally, TFCC "sits within the broader context of patriarchal gender inequality, which includes sexist and heterosexist social norms, gendered structural inequality, and the traditionally male-dominated digital media industry".[55] Legal and policy approaches to platform accountability for TFGBV must thus also take into account that it includes TFCC between intimate partners in private, semi-public, and public channels.

2.1.3. Impacts of TFGBV and Intersectionality

The impacts of TFGBV on women and girls, including those who are also members of other historically marginalized groups, are substantial and far-reaching. Jane Bailey and Valerie Steeves write:

> [TFGBV] can lead to fear, social withdrawal, physical and psychological illness, physical danger and harm, serious consequences relating to reputation that affect targets'

---

[51] "What started that night would eventually be called Gamergate. Its catalyst was a blog post written by an ex-boyfriend, accusing [Quinn] of sexual promiscuity. Within days, nude photographs of her were circulating on the internet alongside commentary and speculation about her weight, her looks, her genitalia. […] Ms. Quinn was not the only person to be chased out of her home by Gamergate. Days later, the feminist media critic Anita Sarkeesian left her home after receiving a series of explicit death and rape threats that included her home address and her parents' home address. Two months later, someone posted the home address of the game developer Brianna Wu on 8chan." Sarah Jeong, "When the Internet Chases You From Your Home", *New York Times* (15 August 2019), online: <https://www.nytimes.com/interactive/2019/08/15/opinion/gamergate-zoe-quinn.html>.

[52] "On August 15, 2014, an angry 20-something ex-boyfriend published a 9,425-word screed and set in motion a series of vile events that changed the way we fight online. The post, which exhaustively documented the last weeks of his breakup with the video game designer Zoë Quinn, was annotated and punctuated with screenshots of their private digital correspondence — emails, Facebook messages and texts detailing fights and rehashing sexual histories. It was a manic, all-caps rant made to go viral. And it did. The ex-boyfriend's claims were picked up by users on Reddit and 4chan and the abuse began." Charlie Warzel, "How an Online Mob Created a Playbook for a Culture War", *New York Times* (15 August 2019), online: <https://www.nytimes.com/interactive/2019/08/15/opinion/what-is-gamergate.html>.

[53] Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2008) 18:4 Feminist Media Studies 609 at 610.

[54] *Ibid*; and Molly Dragiewicz et al, "Technology-facilitated coercive control" in Walter S DeKeseredy, Callie Marie Rennison & Amanda K Hall-Sanchez, eds, *The Routledge International Handbook of Violence Studies* (London: Routledge) 244; Molly Dragiewicz et al, "Domestic violence and communication technology: Survivor experiences of intrusion, surveillance, and identity crime" (July 2019) at 9, online: *Australian Communications Consumer Action Network* <https://accan.org.au/files/Grants/20190823%20Domestic%20violence%20and%20communication%20technology%20victim%20experiences%20of%20intrusion%20surveillance%20and%20identity%20theft.pdf>; and Jessica West, Cyber-Violence Against Women (May 2014), online: *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf>.

[55] Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2008) 18:4 Feminist Media Studies 609 at 610.

> social, employment and family lives, and, in limited circumstances can be a contributing factor to self-harming behaviours and suicide. As a result, TFVAWG [technology-facilitated violence against women and girls] affects women's and girls' physical, sexual and psychological integrity, equality, privacy, and autonomy in ways that undermine their right to full public participation. It therefore triggers international obligations relating to both violence and discrimination against women and girls.[56]

For individuals subjected to it, TFGBV results in consequences such as social ostracization and isolation,[57] physical illness, and emotional and psychological trauma, including "damaged self-esteem, a loss of self-worth, feelings of sadness and anger, anxiety, fear for personal safety, social withdrawal, and depression. In the most serious of cases, TFGBV can lead to women and girls dying by suicide"[58]—as it did in the case of Rehtaeh Parsons in Nova Scotia and Amanda Todd in British Columbia.[59] Online violence, abuse, and harassment can also devastate women's financial well-being (due to, e.g., "costs related to legal support, online protection services, missed wages, and professional consequences")[60] and harm their employment or opportunities for career advancement.[61] For example, the Canadian

---

[56] Jane Bailey, Valerie Steeves & Suzie Dunn, "Submission to the Special Rapporteur on Violence Against Women Re: Regulating Online Violence and Harassment Against Women" (27 September 2017) at 3, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2017/12/Bailey-Steeves-Dunn-Submission-27-Sep-2017.pdf>.

[57] "The social consequences for women can be very severe, particularly if their entire community is involved with the [TFGBV]. In the case of Daisy Coleman, her brother and herself were bullied at school, she was suspended from her cheer leading squad, her mother lost her job, her family was forced to move back to Albany and their home in Maryville was burned down." Jessica West, "Cyber-Violence Against Women" (May 2014) at 16-17, online (pdf): *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf> (footnotes omitted); and "A common social impact of [TFGBV] is isolation from friends and family. In our volunteer focus group, one of the volunteers spoke about a caller whose ex-partner was posting things about them on Facebook and how as a result of the things they were saying, some of her friends and family, including her sister, stopped speaking to her. They believed whatever her attacker had posted on Facebook. In our survey, this is one of the most commonly reported social impacts with 28% of women responding that they experienced isolation from friends and family as a result of [TFGBV]." *Ibid* at 17 (footnotes omitted).

[58] House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 36 (Chair: Marilyn Gladu).

[59] "There have been a number of other stories of boys sexually assaulting unconscious girls and recording the assault on cell phones, through picture and video. We can list them off: in November 2011, 15 year old Rehtaeh Parsons was gang-raped by 4 classmates who took pictures and distributed them to her classmates in Nova Scotia; […] In each of these examples, technology was used both during the sexual assault to record or take pictures of the assault and after the assault, as a way to revictimize, humiliate and intimidate survivors." Jessica West, "Cyber-Violence Against Women" (May 2014) at 6-7, online: *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf>; "Weeks after posting haunting Youtube video on her years of torment at classmates' hands, 15-year-old B.C. girl commits suicide", *Canadian Press* (12 October 2012), online: <https://nationalpost.com/news/canada/amanda-todd-suicide-2012>.

[60] Ronald J Deibert et al, "Submission of the Citizen Lab (Munk School of Global Affairs, University of Toronto) to the United Nations Special Rapporteur on violence against women, its causes and consequences, Ms. Dubravka Šimonović" (2 November 2017) at 2, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2017/11/Final-UNSRVAG-CitizenLab.pdf>.

[61] House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 38 (Chair: Marilyn Gladu); "Economic harm can be done when the explicit image of a victim of [TFGBV] covers several pages of search engine results, making it difficult for the victim to find employment, or even preventing the victim from even attempting to find employment because of the shame and fear of potential employers discovering the images." Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 27.

Judicial Council subjected former Associate Chief Justice of the Manitoba Court of Queen's Bench, Lori Douglas, to a rare public inquiry to assess her fitness to remain a judge because her husband had posted nude photos of her online without her consent—essentially being professionally censured for having become a victim of NCDII.[62]

On a broader societal, political, and democratic level, TFGBV relegates women and girls to secondary status online and in the world. They are rendered unable to freely and fully participate in society and prevented from enjoying true or equal protection of their human rights and fundamental freedoms, including the right to freedom of expression. The most common response to facing online abuse and harassment is that women reduce their online activities, avoid certain social media platforms or conversations, withdraw from expressing their views, or self-censor if they continue to engage online.[63] In other words, women are "driven off of the Internet". [64] This curtails their ability to participate in the contemporary public sphere, including engaging in activism and advocacy, influencing public opinion, or mobilizing social, cultural, or political change. The connection between online abuse and harassment targeting women, and their public participation in democratic society and politics, is made clear in the particularly virulent and voluminous abuse that is routinely hurled at female journalists,[65] female politicians,[66] female activists and human rights defenders,[67] and feminists.[68] The UN Special Rapporteur on violence against women, its causes and consequences, Dubrava Šimonovic, elaborates:

> Women human rights defenders, journalists and politicians are directly targeted, threatened, harassed or even killed for their work. They receive online threats, generally

---

[62]Glenn Kauth, "Behind the headlines", *Canadian Lawyer Magazine* (4 January 2016), online: <https://www.canadianlawyermag.com/news/general/behind-the-headlines/270024>; "Lori Douglas, celeb Jennifer Lawrence both nude photo victims: lawyer", *CBC News* (27 October 2014), online: <https://www.cbc.ca/news/canada/manitoba/lori-douglas-celeb-jennifer-lawrence-both-nude-photo-victims-lawyer-1.2814282>.

[63] Jessica West, "Cyber-Violence Against Women" (May 2014) at 17, online (pdf): *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf> (footnotes omitted); "The women in our focus group for women who access services at BWSS also felt this impact the most. Having experienced abusive relationships in the past, all of the women avoided using social media and online platforms in order to keep themselves safe." *Ibid* at 17 (footnotes omitted).

[64] Raine Liliefeldt, "How cyberviolence is threatening and silencing women", *Policy Options* (14 June 2018), online: <https://policyoptions.irpp.org/magazines/june-2018/how-cyberviolence-is-threatening-and-silencing-women/>.

[65] See e.g.., Becky Gardiner et al, "The dark side of Guardian comments", *Guardian* (12 April 2016), online: <https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments>; and Becky Gardiner, ed, *New Challenges to Freedom of Expression: Countering Online Abuse of Female Journalists* (Vienna: OSCE Office of the Representative on Freedom of the Media, 2016).

[66] See e.g., Ashley Burke, "Relentless online abuse of female MPs raises concern for safety of staff", *CBC News* (5 November 2019), online: <https://www.cbc.ca/news/politics/mps-staff-online-hate-security-measures-1.5347221>.

[67] See e.g., Ronald J Deibert et al, "Submission of the Citizen Lab (Munk School of Global Affairs, University of Toronto) to the United Nations Special Rapporteur on violence against women, its causes and consequences, Ms. Dubravka Šimonović" (2 November 2017) at 16, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2017/11/Final-UNSRVAG-CitizenLab.pdf>.

[68] Michelle Goldberg, "Feminist writers are so besieged by online abuse that some have begun to retire", *Washington Post* (20 February 2015), online: <https://www.washingtonpost.com/opinions/online-feminists-increasingly-ask-are-the-psychic-costs-too-much-to-bear/2015/02/19/3dc4ca6c-b7dd-11e4-a200-c008a01a6692_story.html>; "Toxic Twitter - Triggers of Violence and Abuse Against Women on Twitter" (March 2018), online: *Amnesty International* <https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-2/>.

> of a misogynistic nature, often sexualized and specifically gendered. The violent nature of these threats often leads to self-censorship. Some resort to the use of pseudonyms, while others maintain low online profiles, an approach that can have a detrimental impact on their professional lives and reputations. Others decide to suspend, deactivate or permanently delete their online accounts, or to leave the profession entirely. Ultimately, the online abuse against women journalists and women in the media are a direct attack on women's visibility and full participation in public life. […] Online violence against women not only violates a woman's right to live free from violence and to participate online but also undermines democratic exercise and good governance, and as such creates a democratic deficit. [69]

The harmful impacts of TFGBV are further layered and experienced in unique and additionally devastating ways by women and girls with other intersecting identities that also face systemic discrimination. In developing the concept of 'intersectionality', in the context of systemic oppression of Black women, Kimberlè Crenshaw wrote:

> Any particular disadvantage or disability is sometimes compounded by yet another disadvantage emanating from or reflecting the dynamics of a separate system of subordination. An analysis sensitive to structural intersections explores the lives of those at the bottom of multiple hierarchies to determine how the dynamics of each hierarchy exacerbates and compounds the consequences of another.[70]

Remaining "sensitive to structural intersections'' is necessary to understanding TFGBV and evaluating potential legal responses to it. TFGBV impacts, in ways particular to their respective experiences, women, girls, and gender-diverse individuals who belong to more than one historically marginalized group.[71] People who engage in TFGBV target those who belong to the 2SLGBTQQIA community;[72] who

---

[69] Dubravka Šimonović , *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 29.

[70] Kimberlè Williams Crenshaw, "Beyond Racism and Misogyny: Black Feminism and 2 Live Crew" in Mari J Matsuda et al, eds, *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (New York: Routledge, 193) 111 at 114.

[71] See e.g., House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 59-69 (Chair: Marilyn Gladu); Ronald J Deibert et al, "Submission of the Citizen Lab (Munk School of Global Affairs, University of Toronto) to the United Nations Special Rapporteur on violence against women, its causes and consequences, Ms. Dubravka Šimonović" (2 November 2017) at 2, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2017/11/Final-UNSRVAG-CitizenLab.pdf>; and Jessica West, "Cyber-Violence Against Women" (May 2014) at 17, online (pdf): *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf>. For example, according to Status of Women Canada, in 2014, Indigenous women were more than three times as likely as non-Indigenous women to report being a victim of intimate partner violence (despite disproportionate under-reporting among Indigenous communities overall); women with a disability were nearly twice as likely to have been sexually assaulted in the 12 months prior; and lesbian and bisexual women were 3.5 times more likely to report intimate partner violence than heterosexual women: Statistics Canada, *Police-reported sexual assaults in Canada, 2009 to 2014: A statistical profile*, by Cristine Rotenberg, Catalogue No 85-002-X (Ottawa: Statistics Canada, 3 October 2017).

[72] "LGBT youth were nearly three times as likely as non-LGBT youth to say they had been bullied or harassed online (42% vs. 15%) and twice as likely to say they had been bullied via text message (27% vs. 13%). [...] One in four LGBT youth (26%) said they had been bullied online specifically because of their sexual orientation or gender expression in the past year, and one in five (18%) said they had experienced bullying and harassment for these reasons via text message." GLSEN, CiPHR & CCRC, "Out Online: The Experiences of Lesbian, Gay, Bisexual and Transgender Youth on the Internet" (2013) at x, online (pdf): *Gay,*

are Black,[73] Indigenous, or otherwise racialized; who have a disability;[74] who are socioeconomically disadvantaged; who are immigrants or refugees; and/or who practice a marginalized religion, for example[75]—including individuals whose identities overlap multiple intersections among those groups. Black women, in particular, have long been at the forefront of combating TFGBV, due to having been some of the earliest targets of online abuse and coordinated trolling campaigns.[76]

In addition, researchers who investigated the scale of transphobia across social media platforms found 1.5 million transphobic comments out of an analyzed 10 million posts over 3.5 years, which "range[d] in severity from transphobic attitudes through to genocide and violence."[77] Abigail Curlew describes a forum board that gave rise to a troll community whose members "are notorious for their constant attacks on women, specifically trans women, plus sized women, and women with disabilities".[78] Curlew further states, "For those of us who experience marginalization on a daily basis, the presence of far-right trolls on the Internet can transform our participation in digital spaces into a paranoia-fueled nightmare."[79]

Regarding online harassment and abusive speech based on disability, Philippa Hall describes how the impact on targeted individuals is amplified and particularly damaging due to their "greater practical reliance" on digital communications. The Internet played a transformative role in the lives of women with disabilities and thus TFGBV, as facilitated by the same tools and platforms relied upon, "increase[s] their vulnerability as hate speech targets" correspondingly.[80]

Lesbian & Straight Education Network <https://www.glsen.org/sites/default/files/2020-01/Out_Online_Full_Report_2013.pdf>.

[73] See e.g., Maeve Duncan, "1 in 4 black Americans have faced online harassment because of their race or ethnicity", (25 July 2017), online: *Pew Research Center* <https://www.pewresearch.org/fact-tank/2017/07/25/1-in-4-black-americans-have-faced-online-harassment-because-of-their-race-or-ethnicity/>.

[74] See e.g., Philippa Hall, "Disability Hate Speech: Interrogating the Online/Offline Distinction" in Karen Lumsden & Emily Harmer, eds, *Online Othering: Exploring Digital Violence and Discrimination on the Web* (London: Palgrave Macmillan, Cham, 2019) 309.

[75] See e.g., Dubravka Šimonović , *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 28.

[76] Rachelle Hampton, "The Black Feminists Who Saw the Alt-Right Threat Coming", *Slate* (23 April 2019), online: <https://slate.com/technology/2019/04/black-feminists-alt-right-twitter-gamergate.html>; and Seyi Akiwowo, "Amnesty's Latest Research Into Online Abuse Finally Confirms What Black Women Have Known For Over A Decade", *Huffington Post* (19 December 2018), online: <https://www.huffingtonpost.co.uk/entry/amnesty-online-abuse-women-twitter_uk_5c1a0a2fe4b02d2cae8ea0c1>.

[77] "Exposed: The Scale of Transphobia Online", online: *Brandwatch* <https://www.brandwatch.com/reports/transphobia/>.

[78] Abigail Curlew, "Doxxing, vigilantes, and transmisogyny" (3 May 2019), online: *Medium* <https://medium.com/@digitaljusticelab/doxxing-vigilantes-and-transmisogyny-c2b8a6abb2b2>.

[79] Abigail Curlew, "'People will try to find where you live': On doxxing and the social media surveillance monster" (22 March 2019), online: *Medium* <https://medium.com/@abigail.curlew/people-will-try-to-find-where-you-live-on-doxxing-and-the-social-media-surveillance-monster-b05bab5438fd>.

[80] Philippa Hall, "Disability Hate Speech: Interrogating the Online/Offline Distinction" in Karen Lumsden & Emily Harmer, eds, *Online Othering: Exploring Digital Violence and Discrimination on the Web* (London: Palgrave Macmillan, Cham, 2019) 309 at 313.

Women and girls who live at the intersection of multiple systemic oppressions are also disproportionately subjected to TFCC by intimate or former partners.[81] Further, they are more likely to experience other forms of violence against women, such as sexual abuse, sexual assault, or sexual harassment; intimate partner violence; and workplace harassment. For example, Inuit women in Nunavut face criminal harassment, sexual assault, and "indecent or harassing communication" at 3.6 times, 7.2 times, and 8.9 times the national average, respectively.[82] This disproportionate impact is reflected in online abuse targeting them as well.[83] In fact, the volume and extremity of hate-based and discriminatory speech targeting Indigenous peoples, including Indigenous women, in the comments section of articles by the Canadian Broadcasting Corporation (CBC), Canada's national news service, was of such an extent that it compelled the CBC to shut down their comments feature entirely for articles concerning Indigenous peoples.[84]

2.1.4. Note on Terminology: Call It TFGBV

The types of activities and behaviours comprising TFGBV are varyingly referred to by a litany of other terms, such as 'cybermisogyny', 'cyberviolence', 'technology-assisted violence and abuse', or 'ICT-facilitated violence against women' (where ICT stands for information and communications technology).[85] Similarly, many pre-existing forms of violence and abuse against women, such as stalking, harassment, or defamation, have been described as 'cyberstalking', 'cyber harassment', or 'cyber defamation', respectively, if the act has occurred online or been facilitated or aggravated by the integration of technology or online platforms.[86] This report uses 'technology-facilitated gender-based violence, abuse, and harassment' (TFGBV), and has replaced terms such as 'cyberviolence' with 'TFGBV' throughout, for the following reasons.

First, indicating that violence, abuse, and harassment is facilitated by technology includes a central component of the harm (technology), while not minimizing the fact that it is still violence, abuse, or

---

[81] Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2008) 18:4 Feminist Media Studies 609 at 612.

[82] See e.g., Kent Driscoll, "Targeted: Inuit women face harassment - online and off", *APTN News* (20 February 2020), online: <https://aptnnews.ca/2020/02/20/targeted-inuit-women-face-harassment-online-and-off/>.

[83] *Ibid*.

[84] "Uncivil dialogue: Commenting and stories about indigenous people", *CBC News* (30 November 2015), online: <https://www.cbc.ca/newsblogs/community/editorsblog/2015/11/uncivil-dialogue-commenting-and-stories-about-indigenous-people.html>.

[85] See e.g., Karlie E Sonard et al, "'They'll Always Find a Way to Get to You': Technology Use in Adolescent Romantic Relationships and Its Role in Dating Violence and Abuse" (2017) 32:14 Journal of Interpersonal Violence 2083; House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 32 (Chair: Marilyn Gladu); Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 17; Raine Liliefeldt, "How cyberviolence is threatening and silencing women", *Policy Options* (14 June 2018), online: <https://policyoptions.irpp.org/magazines/june-2018/how-cyberviolence-is-threatening-and-silencing-women/>.

[86] See e.g., Sara Baker, "#WhatareyoudoingaboutVAW Campaign: Social Media Accountability" (12 September 2014), online: *GenderIT.org* <https://genderit.org/feminist-talk/whatareyoudoingaboutvaw-campaign-social-media-accountability>; R Stuart Geiger, "Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space" (2016) 19:6 Information, Communication & Society 787 at 791; "Defamation laws (cyber-libel) and the Internet", online: *LegalLine* <https://legalline.ca/legal-answers/defamation-laws-cyber-libel-and-the-internet/>.

harassment, and should be treated with the same seriousness that is, in theory, accorded to such acts committed in the physical world without technology. The Citizen Lab has written,

> Attempting to draw clear boundary lines between "online" and "offline" conduct in this context is often difficult and frequently unhelpful. In some cases, online behaviour may amplify, facilitate, or exacerbate traditional categories of problematic conduct. In other cases, technology allows for entirely new forms of violence, abuse, or harassment to take place. As a result, the language of "technology-facilitated" violence may be more inclusive or appropriate in some cases.[87]

Bailey, Steeves, and Dunn also emphasize that terminology should depart from artificial distinctions between "separate online and offline spheres because this increasingly does not comport with the lived realities of most people and certainly does not comport with the lives of young Canadians."[88]

Second, avoiding the term 'cyberbullying', in particular, "ensure[s] that the structural underpinnings of violence facilitated against women and girls through digital technologies is not obscured or minimized by a term often associated with disagreements among individual school children."[89] Research has shown that the term holds no meaning for those it is most associated with—youth[90]—and that those impacted "hate the term cyber-bullying. They felt that the term cyber-bullying has really done them a disservice. What they say is, 'Call it what it is; it's violence. Call it what it is; it's misogyny and racism.'"[91]

Third, terms that add 'cyber-' as a prefix to pre-existing forms of misogyny, violence, abuse, and harassment are increasingly unhelpful in that too often they contribute to minimization and dismissal of the very non-cyber consequences of such acts. Indicating that abuse is technology-facilitated appropriately focuses on the manner and mechanism of how the abuse was committed, without inherently imposing presumptions that the nature or extent of the harm itself has changed. This is especially important given that in many cases, technology's ability to efficiently amplify and automate abuse causes greater harm than if the same act were not technology-facilitated. For example, 'cyberstalking' is stalking, but facilitated through technology such as spyware or GPS location tracking. The manner of the offence does not take away from what the offence is at core, but is important to note

---

[87] Ronald J Deibert et al, "Submission of the Citizen Lab (Munk School of Global Affairs, University of Toronto) to the United Nations Special Rapporteur on violence against women, its causes and consequences, Ms. Dubravka Šimonović" (2 November 2017) at 3, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2017/11/Final-UNSRVAG-CitizenLab.pdf>.

[88] Jane Bailey, Valerie Steeves & Suzie Dunn, "Submission to the Special Rapporteur on Violence Against Women Re: Regulating Online Violence and Harassment Against Women" (27 September 2017) at 3, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2017/12/Bailey-Steeves-Dunn-Submission-27-Sep-2017.pdf>. They continue, "Further, conceiving of separate spheres too often leads to unrealistic and paternalistic advice that targeted women and girls should just 'go offline' to avoid violence, which in many ways emulates the all-too-familiar victim-blaming approaches to VAWG that shift responsibility away from perpetrators and the community, toward survivors themselves."

[89] *Ibid* at 3.

[90] "[W]e were careful in our questionnaire not to use the term 'cyberbullying' because we have a strong sense, from American research, that this is not a term that has meaning for youth. It's something they see adults using. They perceive it as something that either younger kids do or other kids do, but not themselves." Canada, Parliament, House of Commons, *Standing Committee on the Status of Women*, 42nd Parl, 1st Sess, No 20 (16 June 2016) at 1705 (Matthew Johnson).

[91] House of Commons, *Standing Committee on the Status of Women,* 42nd Parl, 1st Sess, No 21 (21 September 2016) at 1615 (Valerie Steeves).

given the extent to which technology often exacerbates and compounds harms from the violence and abuse inherent in the act, regardless of how it is facilitated.

It is on the basis of the above research and reasoning that this report recommends the permanent retirement of the term 'cyberbullying'. We have begun that process in this report, wherever possible (such as where it does not change the substantive meaning of a sentence and where the quotation is not excerpted legislation). At the very least, the term has outlived any possible usefulness given the availability of more specific and informed language. More pertinently, it minimizes and detracts from the violence, abuse, and harassment which is targeted at women or other historically marginalized groups, on the basis of gender, race, sexual orientation, or other marginalized identities, through technological means.

With respect to using "gender-based violence, abuse, and harassment", instead of "violence against women and girls", Bivens explains the dilemma surrounding these terms:

> For some, "violence against women" evokes the deep-seated racism, ableism, heterosexism, and cissexism that taint early iterations of the women's movement. For others, "gender-based violence" can be problematic because it has been employed by some as a way of neutralizing the differences between men's and women's experiences of sexual violence.[92]

This report has opted to use "technology-facilitated gender-based violence, abuse, and harassment" as the central term, but will also refer throughout to violence against women and girls, to reflect its gendered nature. Similarly, while the shortened acronym TFGBV is used for ease of reference, it is intended to represent all of "violence, abuse, and harassment" (not just "violence"), and the report will use each of these terms separately and together throughout. This is to encompass the full spectrum of harmful and damaging behaviours towards women and girls online, while recognizing that the boundaries between violence, abuse, and harassment can be contested.

Bivens also recommends that organizations may dispense with umbrella terms and "narrowly focus on what they are doing at that particular moment" within a particular project, which would allow for "focus on the intersections arising out of a particular situation while resisting the impulse to include everything within one label, thus obscuring the specific ways in which power operates."[93] In recognition of that risk, this report focuses primarily on technology-facilitated violence, abuse, and harassment perpetrated against women and girls (including trans women), while maintaining an intersectional outlook, such as noting intersecting impacts of race or disability, for instance. The underlying common thread is technology-facilitated violence rooted in patriarchy and misogyny. However, given that digital platform liability is still an emerging issue in Canada, many of the issues discussed and the analyses provided in this report may to some extent be transferrable to other marginalized groups who are targeted for violence, abuse, and harassment through online platforms (pending further research with a specific focus on other marginalized identities, which also should be conducted).

---

[92] House of Commons, *Standing Committee on the Status of Women,* 42nd Parl, 1st Sess, No 21 (21 September 2016) at 1545 (Rena Bivens).

[93] House of Commons, *Standing Committee on the Status of Women,* 42nd Parl, 1st Sess, No 21 (21 September 2016) at 1545 (Rena Bivens).

2.2. Speech-Based TFGBV on Digital Platforms

Speech (or expression)-based violence, abuse, and harassment make up the broadest category of TFGBV, in part due to the category's nebulous and catch-all nature. The harmful expression and behaviours examined include online verbal and multimedia-based abuse and harassment, online mobbing with verbal and multimedia-based abuse, hate speech, threats (including rape threats and death threats), trolling, intimidation, and coordinated online attack campaigns. Such speech, or expression can amount to substantive behaviours that are simply enacted through words (or images or other multimedia), often with intent to induce tangible negative impacts on the recipient.[94] The terms 'speech' and 'expression' are generally used interchangeably throughout this report, but both are intended to refer to the same spectrum and kind of content. This is to invoke and respond to relevant discussions in the Canadian context, where 'expression' is the familiar term, and relevant discussions in other jurisdictions such as the United States, where 'speech' may be the more prevalent term.

Although TFGBV may exist independently of the types of platforms that are the focus of this report, the nature of the platforms themselves uniquely characterizes and plays a central role in the TFGBV that proliferates across them.[95] Writing about racist abuse on social media, Ariadna Matamoros-Fernández coined the term 'platformed racism' to capture how the particular dynamic of digital platforms—including such companies' embedded cultural values and politics—combines with the dynamics of racism to result in "a new form of racism articulated via social media".[96]

In the context of TFGBV, Matamoros-Fernández's concept of 'platformed racism' might be adapted to capture similar dynamics where platform-facilitated TFGBV is concerned, such as in the term 'platformed misogyny' or 'platformed TFGBV' (while noting that TFGBV extends beyond misogyny to include systemic oppression of other and intersecting historically marginalized groups). This is the context informing the discussion and analysis of TFGBV throughout this report, as perpetuated by and across digital platforms in particular.

The remainder of this section provides an overview of what constitutes speech (or expression)-based TFGBV, using illustrative examples from Canadian case law and the relevant literature. Section 2.2.1 will review examples of expression-based TFGBV that appear in Canadian criminal and civil case law. Section 2.2.2 will discuss how TFGBV particularly targets women who have a strong online presence, are highly visible (or attributed visibility) in some way, are publicly vocal about advocating for gender equality or other issues of social justice, or whose professions place them in the public sphere—all elements which implicate their right to freedom of expression and its underlying values, which such women are attacked for exercising.

---

[94] For a more detailed discussion of harassing, abusive, or violent expression amounting to substantive action, see Section 3.2.1 ("Platformed TFGBV Weaponizes Expression to Harm Women").

[95] This is discussed in depth in Part 3 ("Role of Digital Platforms in TFGBV").

[96] Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930 at 931.

***Content Warning:* Please note that some of the case details and quotations included in the body and primarily the footnotes of Sections 2.2.1, 2.2.2, 3.1.3, 3.3.1, and 3.4.2 may be disturbing, such as excerpts of violent threats sent to women or girls. We have indicated (Content Warning) prior to some of the more graphic quotations.**

2.2.1. Expression-Based TFGBV in Canadian Case Law

Canadian civil and criminal law contain numerous cases exemplifying one or more of the above situations. It is important to note that these examples are taken only from situations that have escalated to the point of being reported to the police, that have been taken seriously enough by law enforcement and the justice system to prosecute, and that have been disposed of through a judicial determination. As such, they represent only a fraction of the reality of platformed TFGBV perpetrated against women and girls.

Speech-based TFGBV through digital platforms, which have been documented in Canadian case law, includes behaviours such as:

- sending onslaughts of messages to women through social media platforms such as Facebook or other communications channels (including Twitter, email, voice and text messages, and in at least one case, Google Review)[97], despite their requests to stop and despite their taking actions such as blocking or deleting the harasser;[98]
- escalating volume, frequency, violence, and vitriol in communications in the absence of a response or if given a negative response;[99] and
- creating and sending to the victim's coworkers a website with intimate details.[100]

Threats to women have also extended to their family members.[101] In at least one documented case, an individual posted to Facebook "a threat to cause death or bodily harm to All Women".[102]

---

[97] *R v JR*, [2018] OJ No 6409 (QL) (via Benjamin Perrin, *Social Media Crime in Canada: Annotated Criminal Code, R.S.C., 1985, c. C-46*, 2nd ed (Vancouver: Peter A Allard School of Law Allard Research Commons, 2019) at 55).

[98] *R v Broydell*, 2018 CanLII 1161 (NL PC); *R v Donatucci*, 2009 ONCJ 734.

[99] *Ibid*.

[100] *R v Fox*, 2017 BCSC 2361, summarized in eQuality Project, "Technology-Facilitated Violence: Criminal Harassment Case Law" (3 July 2020), at 23, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2020/07/TFVAW-Criminal-Harassment-3-July-2020.pdf>. (Details sent to the targeted person's coworkers included: "private photos of her, her friends, and family members, including her child from another partnership; her full name, address, and contact information; allegations that she was a white supremacist, a sociopath, and an unfit mother, among other things; disparaging comments about her and people she associated with; private email communications between the two; and a blog purportedly written in the perspective of Ms. C [the victim] describing her as a terrible person.")

[101] (*Content Warning*) "Tell your mom that if she doesn't fucking straighten out, she will be fucking drug (sic) behind a truck." *R v Lauck*, [2018] AJ No 1312, at para 54 via Benjamin Perrin, *Social Media Crime in Canada: Annotated Criminal Code, RSC 1985, c. C-46*, 2nd ed (Vancouver: Peter A Allard School of Law Allard Research Commons, 2019) at 54-55.

[102] *R c Hunt*, 2012 QCCA 4688.

Women are additionally affected by expression that attacks them based on intersecting marginalized identities. For example, one case involved someone posting threats and hateful language with respect to Muslim individuals.[103]

In *R v Fox*, the perpetrator "sent [the targeted individual] hundreds, if not thousands, of emails to her and people she knew with the intention of degrading, humiliating and tormenting her. The emails included comments such as 'I will destroy you—slowly and incrementally [...] Every moment of my life is focused on the single goal'."[104] Other cases involved setting up a fake Facebook profile to catfish[105] a former spouse to fraudulently extract information from her;[106] and sending a classmate unwanted Facebook messages describing violent sexual fantasies involving her.[107]

Often the abusive expression is combined with other forms of violence, abuse, or harassment that exceed the boundaries of speech-based TFGBV to overlap with other substantive harms such as invasion of privacy, violation of sexual boundaries, impersonation, defamation, and putting the victim in physical danger in addition to psychological distress. This has included harassing the victim in person or through sending physical mail and unwanted objects; distributing nude photos of the victim to the public and to the woman's coworkers, family, and friends;[108] creating fake social media profiles impersonating the victim and making false claims (e.g., that the victim was spreading HIV);[109] impersonating the victim to set up sexual encounters with male strangers online and sending them to the victim's apartment for sex, without her knowledge;[110] recording the victim engaged in sexual activity without her knowledge or consent, and distributing the video on Facebook and through email to her

---

[103] *R c Rioux*, 2016 QCCQ 6762.

[104] *Ibid*, summarized in eQuality Project, "Technology-Facilitated Violence: Criminal Harassment Case Law" (3 July 2020), at 23, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2020/07/TFVAW-Criminal-Harassment-3-July-2020.pdf>.

[105] To 'catfish' an individual is to lure that individual into an online relationship, while pretending to be another person (including using a different name and photo). Catfishing may be done for its own sake or as a basis to engage in further illicit activities, such as fraudulently obtaining money from the target of the catfish.

[106] *R v Smith*, 2014 ONCA 324, summarized in eQuality Project, "Technology-Facilitated Violence: Criminal Harassment Case Law" (3 July 2020), at 67, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2020/07/TFVAW-Criminal-Harassment-3-July-2020.pdf>.

[107] (*Content Warning*) "Much of the Facebook conversation was about the defendant's desire to inflict pain. On several occasions he asked the complainant to submit to his wishes – but, he also stated he did not consider this necessary. He said: "I wanna cut ur stomach open and stick my dick in it" and "break ur legs and jerk on ur face". After the complainant told him to "calm down" the defendant responded as follows[:] "I don't care if u want it to happen or not". Later he said, "I want to cut u....so why can't I". … It is significant that the parties had only known each other for three days." *R v DD*, 2013 ONCJ 134 at paras 18-19. The Court went on to state, at para 21, "The Facebook conversation reflects his need to cause bodily harm as a source of sexual gratification. He described the violent nature of the acts contemplated and sought the complainant's submission to his desire. He also said he did not care if she consented. I have no doubt these words were meant to be taken seriously and that they intimidated the complainant. Indeed, I am confident he derived pleasure from the threats themselves."

[108] See e.g., *R v Wenc*, 2009 ABCA 328; *R v SB et al*, 2014 BCPC 279; and *R v Korbut*, 2012 ONCJ 691.

[109] *Ibid*.

[110] *Ibid*; see also *R v Korbut*, 2012 ONCJ 691, summarized in eQuality Project, "Technology-Facilitated Violence: Criminal Harassment Case Law" (3 July 2020), at 68, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2020/07/TFVAW-Criminal-Harassment-3-July-2020.pdf> ("To punish her for entering a new relationship against his wishes, he then posted his ex-partner's personal information on a dating site, prompting phone calls and visits from strange men at night. He also left a pornographic video at her new partner's residence, calling it a 'Valentine's gift.'")

friends and family;[111] hiding an Internet-connected webcam in her bedroom;[112] doxing and SWATting her;[113] sextortion;[114] taking over the victim's own social media accounts;[115] covert surveillance and tracking and monitoring the victim through her digital devices;[116] sexual luring and child sex exploitation;[117] and non-technological forms of violence and abuse, such as stalking, in-person harassment, assault, and sexual assault.[118]

In one case with a particularly reprehensible and wide-ranging suite of abusive behaviours, involving at least twenty-five known victims, "Mr. B used a variety of tactics to harass, threaten, and harm his victims, many of whom were female video gamers he encountered online [often through the livestreaming platform Twitch.tv]. For example, he remotely interfered with his victims' internet service, made fraudulent 9-1-1 calls to victims' homes ['SWATting'], made fraudulent bomb, kidnapping and death threats to the police, and disclosed victims' credit card information online"[119]—after first threatening to disclose such information unless the victim agreed to "show her butt on the internet".[120] The judge in the case found that the defendant had done to numerous women all of the activities listed in the following passage, and more:

> When Victim #23 was live streaming her play of a video game you sent her a follower request that she denied. You retaliated by posting her personal information and that of her parents on Twitter. You attacked the functionality of her computer several times and you phoned the home of Victims #24 and #25 multiple times between 8:00 PM and midnight every day for a period of almost two months, asking to speak to her and indicating that she owed you. You also called Victim #23's cell phone 5-20 times a day, particularly while she was broadcasting on Twitch.tv. You told her that you were sorry for everything that you had done but that if she refused to talk to you, her family would be in danger and that you would swat her. You also overwhelmed her Twitch and

---

[111] *R v PD*, 2011 ONCJ 133.

[112] *R v Corby*, 2016 BCCA 76.

[113] *R v BLA*, 2015 BCPC 20.

[114] *R v MR,* 2017 ONCJ 943, summarized in eQuality Project, "Technology-Facilitated Violence: Criminal Harassment Case Law" (3 July 2020), at 56-57, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2020/07/TFVAW-Criminal-Harassment-3-July-2020.pdf> ("During their relationship, the victim had sent multiple nude images to Mr. R. Following the breakdown of the relationship, he threatened to share the intimate images with her parents if she did not agree to follow his side of the story in relation to the incident that he wanted to keep secret. […] Later, several friends and family members of the victim received an anonymous email with several intimate images of the victim and copies of her Facebook messages about a previous boyfriend. A second anonymous email with an intimate image was sent to her father. Her school supervisor also received an anonymous email accusing the victim of fraud.")

[115] *R v MR*, 2017 ONCJ 943.

[116] *R v Smith*, 2014 ONCA 324.

[117] *R v Adams*, 2016 ABQB 648.

[118] *R v CL*, 2014 NSPC 79.

[119] *R v BLA*, 2015 BCPC 203, summarized in eQuality Project, "Technology-Facilitated Violence: Criminal Harassment Case Law" (3 July 2020), at 31-32, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2020/07/TFVAW-Criminal-Harassment-3-July-2020.pdf>.

[120] *Ibid* at para 9.

> Twitter accounts and her phone with a huge number of simultaneously sent messages using a bot, [a] software application that runs automated tasks over the internet.[121]

Women are also abused and harassed on digital platforms in response to and in direct relation to their professional activities and political views. As Sundén & Paasonen note, "Online misogyny violently targets non-men, non-white, and non-straight subjects who make noise and embody difference on public online platforms. Public figures like politicians and journalists inhabit particularly vulnerable positions, as do authors, artists, and musicians who stand up for feminism and anti-racism."[122]

For example, in one case, the defendant was charged with criminally harassing, via Twitter, two women who were feminist activists in Toronto politics, Stephanie Guthrie and Heather Reilly.[123] Although the defendant was acquitted of the criminal charges, the Court did find that he had either knowingly or recklessly engaged in harassment of the two women through repeated communications and after they had requested he stop. Notably, Guthrie "quit [Twitter] after men's rights activists deluged her account with abuse"[124] following the acquittal, in essence punishing her for having come forward to the police. In another case, a man was convicted of criminal harassment and uttering threats for tweeting "a stream of violent threats" which were found to be "overtly threatening both physically and sexually in nature" to Michelle Rempel, a Conservative Member of Parliament who is also outspoken on Twitter.[125]

Examples of expression-based and other types of TFGBV also abound throughout civil law cases in Canada. In *AB v Bragg*, a fifteen-year-old girl "was targeted by a fake Facebook profile that an unknown person created about her.[126] The fake profile not only used a variation on AB's name, but also included a photo of her and purported to discuss her allegedly preferred sexual acts, as well as her appearance and weight."[127] In *Jane Doe 72511 v NM*, which established the tort of public disclosure of private facts in Ontario, the defendant had posted a sexually explicit video of her on an online pornography platform, without her consent.[128] After admitting to having posted the video "as revenge for her calling the police and having him charged with assault", the defendant later threatened to post nude photos of the plaintiff if she proceeded with legal action against him.[129] In *R v MM*, "a group of six high-school boys had pressured girls for nude images and then saved them to a Dropbox folder they all had access to,

---

[121] *Ibid* at para 38.

[122] Jenny Sundén & Susanna Paasonen, "Shameless hags and tolerance whores: feminist resistance and the affective circuits of online hate" (2018) 18:4 Feminist Media Studies 643 at 650.

[123] *R v Elliott*, 2016 ONCJ 35.

[124] Shane Dingman, "BuzzFeed writer's harassment just the latest example of why Twitter is broken for women", *Globe and Mail* (22 February 2016), online: <https://www.theglobeandmail.com/technology/harassment-of-buzzfeeds-koul-shows-how-twitter-is-broken-for-women/article28845004>.

[125] Ashley Csanady, "The Twitter trial you never heard about: Toronto man found guilty of harassing Michelle Rempel", *National Post* (29 January 2016), online: <https://nationalpost.com/news/politics/the-twitter-trial-you-never-heard-about-toronto-man-found-guilty-of-harassing-michelle-rempel>.

[126] *AB v Bragg*, 2012 SCC 46.

[127] Jane Bailey, "'Sexualized Online Bullying' Through an Equality Lens: Missed Opportunity in *AB v Bragg*" (2014) 59:3 McGill Law Journal 711.

[128] *Jane Doe 72511 v NM*, 2018 ONSC 6607 at paras 24-25, 28-29.

[129] *Ibid* at paras 28-29.

unbeknownst to the girls."[130] Lastly, although it does not involve TFGBV, the January 2021 decision *Caplan v Atas* is notable for having established the tort of online harassment in Ontario, which will likely see significant application to situations of TFGBV in the future.[131]

2.2.2. TFGBV Targets Women with High Visibility or Public Presence

TFGBV particularly impacts women who are outspoken online, have a greater public presence, vocally advocate for gender equality or other social justice issues, or are highly visible (or are attributed visibility). This is significant because laws addressing TFGBV through platform regulation often face opposition on the basis of concerns for freedom of expression. Where women with a public presence are concerned, however, they are targeted precisely because they are exercising freedom of expression and furthering its underlying core values: individual fulfillment through self-expression, contributing to truth-seeking by imparting views and information from the perspectives of historically marginalized groups, and participating in politics and democracy.[132]

Curlew explains the relationship between having an active online presence and exposure to TFGBV:

> This form of strategic harassment [monitoring, doxing, and online abuse] is made possible through the visibility afforded to us by daily participation in social media platforms across the Internet. [...] [F]or those who engage in any kind of knowledge or creative based work, the Internet is our home, a way to contribute our work to the wider public discourse. For journalists, activists, artists, and academics, our IRL [in-real-life] identity and our digital identities are becoming increasingly wed together.[133]

Public visibility that is required for professional purposes, for instance, or simply for self-expression as an aspect of human flourishing, becomes unwanted visibility that results in heightened vulnerability to abuse, particularly for women who belong to multiple historically marginalized groups. As Curlew states, "We live in the era of visibility, where we knowingly make ourselves visible to an unknowable audience. And the [data and informational] exhaust from this visibility can be decontextualized and used by trolls and vigilantes to discredit or embarrass their victims."[134]

In addition to activists, artists, and academics, female politicians and journalists in Canada and globally are targeted by high levels of TFGBV.[135] According to a 2016 Inter-Parliamentary Union study, 82 percent of female parliamentarians in 39 countries across five global regions have

---

[130] Suzie Dunn & Alessia Petricone-Westwood, "'More than 'Revenge Porn': Civil Remedies for the Non-consensual Distribution of Intimate Images" (Paper delivered at the 38th Annual Civil Litigation Conference, Mont Tremblant, QC, 16 November 2018) at 6.

[131] *Caplan v Atas*, 2021 ONSC 670 at para 168.

[132] The relationship between TFGBV and the freedom of expression of individuals impacted by TFGBV is discussed in depth in Section 6.1.3 ("TFGBV Is Low-Value Expression Far from the Core of Section 2(b)").

[133] Abigail Curlew, "'People will try to find where you live': On doxxing and the social media surveillance monster" (22 March 2019), online: *Medium* <https://medium.com/@abigail.curlew/people-will-try-to-find-where-you-live-on-doxxing-and-the-social-media-surveillance-monster-b05bab5438fd>.

[134] Abigail Curlew, "Doxxing, vigilantes, and transmisogyny" (3 May 2019), online: *Medium* <https://medium.com/@digitaljusticelab/doxxing-vigilantes-and-transmisogyny-c2b8a6abb2b2>.

[135] See e.g., Ashley Burke, "Relentless abuse of female MPs raises concern for safety of staff", *CBC News* (5 November 2019), online: <https://www.cbc.ca/news/politics/mps-staff-online-hate-security-measures-1.5347221>; "Online Misogyny in

> experienced some form of psychological violence (remarks, gestures and images of a sexist or humiliating sexual nature made against them or threats and/or mobbing) while serving their terms. They cited social media as the main channel through which such psychological violence is perpetrated; nearly half of those surveyed (44 per cent) reported having received death, rape, assault or abduction threats towards them or their families. Sixty-five per cent had been subjected to sexist remarks, primarily by male colleagues in parliament and from opposing parties as well as their own.[136]

In 2018, the *Guardian* published a study finding that their female journalists received a "significantly higher proportion of blocked comments" on their articles (where comments were blocked by moderators due to violating their commenting policy, including abusive or dismissive comments); and 57% of abusive comments targeted at female journalists "focused on their body, private life, or sexuality", over three times more than similar comments aimed at male journalists (17%).[137] The *Guardian* also found that "women are more likely to experience abuse when they are perceived to be intruding on 'male' spaces" in the subject matter of their articles.[138] Thus, the same spaces that are most in need of diversifying perspectives—such as the journalism industry as a whole, alongside male-dominated topic areas journalism may cover—are also those that are more hostile to anyone who may contribute such perspectives.

Although the *Guardian* is legally a publisher (as opposed to a platform), similar dynamics follow female journalists whenever they express themselves online on digital platforms. Twitter is one of the most prominent and popular platforms for journalists in Canada, and as such, often the site of abusive speech aimed at female journalists. For example, after posting a call for submissions of longform pieces that expressly encouraged marginalized writers ("not white and not male") to submit their writing, *Buzzfeed* editor and writer Scaachi Koul faced an "onslaught of violent threats" that lasted for days, until she was forced to delete her Twitter account—thus cutting off a major channel to her professional network and a valuable avenue of leads, key information, and sources. At one point, Koul noted, *(Content Warning)* she started "to get tweets from white internet men saying that my (white, male) boss should rape and/or murder me as professional discipline."[139] When Milo Yiannopoulous, a prominent far right "online provocateur with 168,000 followers on Twitter, and a hero among the Gamergate and men's rights activist movements", brought his followers' attention to the situation with Koul—thereby ensuring a further barrage of invective targeting her—one user responded, "Toronto is infested Milo,

---

Canadian Politics" (January 2019) online: *Project Someone* <https://projectsomeone.ca/wp-content/uploads/2019/06/ONLINE-MISOGYNY_Feb2019.pdf>; Wendy Kaur, "Prominent Women in Public Office Say That Systemic Sexism Needs a Political Shakedown", *Elle Canada* (5 January 2021), online: <https://www.ellecanada.com/culture/society/prominent-women-in-public-office-say-that-systemic-sexism-needs-a-political-shakedown>.

[136] "Facts and Figures: Ending violence against women" (March 2021), online: *UN Women* <https://www.unwomen.org/en/what-we-do/ending-violence-against-women/facts-and-figures>, citing Inter-Parliamentary Union, "Sexism, harassment and violence against women parliamentarians" (October 2016) online (pdf): *United Nations IPU Archive*: <http://archive.ipu.org/pdf/publications/issuesbrief-e.pdf>.

[137] Becky Gardiner, "'It's a terrible way to go to work:' what 70 million readers' comments on the Guardian revealed about hostility to women and minorities online" (2018) 18:4 Online Misogyny 592 at 598 and 601.

[138] *Ibid* at 600.

[139] Michelle da Silva, "Why were online threats to Rogers taken seriously but not those directed at women?", *Now* (14 March 2016), online: <https://nowtoronto.com/police-take-online-threats-to-rogers-seriously-but-not-those>.

come help us exterminate (not a death threat) these feminist bugs from our great city."[140] In the end, though she has since returned at time of writing, "Koul, a woman of colour who writes critically about racism and sexism, was forced off social media for giving an ear to those who often go unheard."[141]

Misogynistic and sexist expression on digital platforms also affects female students in professional programs, further impacting, at an early stage, their ability to succeed and thrive—particularly in male-dominated fields already known for being exclusionary long before and apart from the Internet. For example, a group of male students from Dalhousie University's Faculty of Dentistry, class of 2015, formed a Facebook group in which *(Content Warning)* "group members voted on which female classmates they'd most like to 'hate fuck,' and joked about drugging women with chloroform and nitrous oxide"[142]—as future medical professionals who would have the power to render patients unconscious during surgical procedures. At least thirteen men were suspended for their involvement in the group—a significant majority of the 19 men in the 38-student 2015 dentistry cohort.[143] As one media outlet reported (*Content Warning*), "a consistent presence of misogynistic, often violent imagery has defined the group's comments. […] Jokes about getting a woman unconscious before sex. How the men's penises were tools 'used to wean and convert lesbians and virgins.' There's plenty more that hasn't been reported. Their [female] classmates were mentioned and pictured often,"[144] and the men "named and rated [them] according to the women's body size and appearance".[145]

Feminist blogger, lawyer, and author Jill Filipovic has detailed similar abusive behaviours in the legal profession, where male lawyers and law students did not wait for women to even begin their careers before misogynistically sabotaging them on a public message board called AutoAdmit, in 2006.[146] The

---

[140] Shane Dingman, "BuzzFeed writer's harassment just the latest example of why Twitter is broken for women", *The Globe and Mail* (22 February 2016), online: <https://www.theglobeandmail.com/technology/harassment-of-buzzfeeds-koul-shows-how-twitter-is-broken-for-women/article28845004>.

[141] Davide Mastracci, "BuzzFeed's search for marginalized writers is progressive, not racist", *Review of Journalism* (21 February 2016), online: <http://rrj.ca/buzzfeeds-search-for-marginalized-writers-is-progressive-not-racist/>.

[142] Hilary Beaumont, "Anonymous Is Threatening to Out Members of a Misogynistic Facebook Group of Canadian Dentistry Students", *Vice* (1 June 2015) online: <https://www.vice.com/en_ca/article/7b7yq4/anonymous-threatens-to-out-members-of-misogynist-dalhousie-dental-student-facebook-group-274>.

[143] Sarah Hampson, "How the dentistry school scandal has let loose a torrent of anger at Dalhousie", *Globe and Mail* (6 March 2015), online: <(https://www.theglobeandmail.com/news/national/education/how-the-dentistry-school-scandal-has-let-loose-a-torrent-of-anger-at-dalhousie/article23344495/>.

[144] Jacob Boon, "'What are they going to do...kick every guy out of fourth year?'", *Coast* (18 December 2014) online: <https://www.thecoast.ca/halifax/what-are-they-going-to-dokick-every-guy-out-of-fourth-year/Content?oid=4488208>. (*Content Warning*) The full version of the latter post describes a penis as "[t]he tool used to wean and convert lesbians and virgins into useful, productive members of society", to positive responses in the comments: Kyle Shaw, "The six Facebook posts that Dal suspended 13 students for", *Coast* (23 January 2015), online: <https://www.thecoast.ca/RealityBites/archives/2015/01/23/the-six-facebook-posts-that-dal-suspended-13-students-for>.

[145] Judy Haiven, "Op-ed: Dalhousie Class of DDS Gentlemen 2015 Graduates this Friday!" *Halifax Media Co-op* (25 May 2015), online: <http://halifax.mediacoop.ca/story/dalhousie-class-dds-gentlemen-2015-graduates-frida/33587>.

[146] (*Content Warning*) ""I found comments reading, "that nose ring is fucking money, rape her immediately;" "I know that girl. Shes a feminazi. She's the person always writing in washington square news [sic] about how men should be killed. She's an incoming IL. Uber-left wing, crazy bitch about sums it up;" "She would be a good hate flick;" and "I really want to kick her in the box for some reason." Someone posted my full name, email address, and AOL screen name. Several pictures of me were posted, and commenters weighed in on my appearance, complete with more remarks about sexual violence." Jill Filipovic, "Blogging While Female: How Internet Misogyny Parallels 'Real-World' Harassment" (2007) 19 Yale Journal of Law and Feminism 295 at 296.

forum also featured a contest to judge who were the "Most Appealing" female law students. Filipovic's requests to take the contest down was refused and her email posted on the AutoAdmit board. Hypocritically, the contest was eventually taken down out of concern for the privacy of one of the male readers engaging in the TFGBV against the female law students.[147]

In the arena of pop culture, feminist media critic Anita Sarkeesian, founder of the non-profit organization Feminist Frequency, became the target of a sustained and particularly brutal online mobbing campaign in response to her launching a series of videos about sexism in video games. Danielle Citron describes:

> Posters tried to hijack her fundraising effort. A campaign was organized to mass-report her Kickstarter project as fraud to get it canceled. Posters tried to shut down her Twitter and Facebook profiles by reporting them as "terrorism," "hate speech," and spam. Her e-mail and social media accounts were hacked. After her Wikipedia page was continually vandalized with explicit sexual images and sexist commentary, Wikipedia reverted the text and locked it down so no further edits could be made.
>
> The […] mob engaged in tactics designed to terrify her. Hundreds of tweets threatened rape. Anonymous e-mails said she should watch her back because they were coming for her. Someone created an online game whose goal was to batter an image of her face. Users of the game were invited to "beat the bitch up" and punch a digital version of her face until it appeared bloodied and bruised. Images of her being ejaculated on and raped spread all over the web. As of March 2014, the attacks had not stopped. Her website Feminist Frequency continues to be hit with denial-of-service attacks.[148]

The creator of the violently misogynistic video game to batter Sarkeesian was a 25-year-old Canadian man living in Sault Ste. Marie, Ontario.[149] In fact, he was publicly identified by Stephanie Guthrie, which played a role in the online abuse campaign that eventually targeted her. Thus, women are not only targeted online for writing, speaking, engaging in politics, advocating for substantive equality, or making their way in professional or male-dominated spaces, but also for attempts to hold the perpetrators of this very abuse accountable.

---

[147] "One of the contest creators "angered other AutoAdmit regulars when he posted the full name of an AutoAdmit and 'Most Appealing' contest reader who happens to be a male attorney at a major New York firm. … Few posters seemed to mind that the pictures and personal information of female law students had also been posted." *Ibid* at 297.

[148] Danielle Keats Citron, *Hate Crimes in Cyberspace* (Cambridge: Harvard University Press, 2014) at 153-54 (footnotes omitted).

[149] "Campaign against misogyny in video game turns ugly", *CTV News* (11 July 2012), online: <https://www.ctvnews.ca/sci-tech/campaign-against-misogyny-in-video-games-turns-ugly-1.874849>.

3. Role of Digital Platforms in TFGBV

As digital platform companies have flourished and grown into the multinational household names that they are today, so too did tremendous harm proliferate on those same platforms. Major social media platforms such as Facebook, Twitter, and YouTube hosted, and continue to host, online abuse and hate speech against women, racialized communities—including Black and Indigenous individuals—2SLGBTQQIA groups and individuals, people with disabilities, immigrants and refugees, and other historically marginalized populations, including those protected under Canadian equality and non-discrimination laws.[150] Even where the most popular dominant platforms have banned or suspended users for engaging in abusive expression or behaviour, other companies such as Gab[151] and Parler[152] have emerged to provide a platform to such users, styling themselves as champions of 'free speech'.

Google Search and Facebook have facilitated additional forms of discrimination and bias against historically marginalized groups, such as advertising arrest record look-up services in response to someone searching a 'Black-sounding' name, and through enabling users and businesses to engage in housing and employment discrimination through targeted ads, respectively.[153] Facebook has become a byword for privacy violations and electoral interference,[154] while social media websites and group-chat-friendly, frictionless-forwarding messaging apps such as WhatsApp are now notorious breeding grounds of disinformation, particularly in the realms of politics and public health.[155]

---

[150] See e.g., Lizzy Davies, "Facebook refuses to take down rape joke pages", *Guardian* (30 September 2011), online: <https://www.theguardian.com/technology/2011/sep/30/facebook-refuses-pull-rape-jokepages>; "Toxic Twitter – A Toxic Place for Women" (March 2018), online: *Amnesty International*, <https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/>; Frances Ryan, "Online abuse of disabled people is getting worse – when will it be taken seriously?", *Guardian* (10 May 2019), online: <https://www.theguardian.com/commentisfree/2019/may/10/online-abuse-disabled-people-social-media>; Casey Newton, "How white supremacists are thriving on YouTube", *Verge* (19 September 2018), online: <https://www.theverge.com/2018/9/19/17876892/youtube-extremism-report-rebecca-lewis-data-society>.

[151] Travis M Andrews, "Gab, the social network that has welcomed Qanon and extremist figures, explained", *Washington Post* (11 January 2021), online: <https://www.washingtonpost.com/technology/2021/01/11/gab-social-network/>

[152] Jack Nicas & Davey Alba, "How Parler, a Chosen App of Trump Fans, Became a Test of Free Speech", New York Times (10 January 2021), online: <https://www.nytimes.com/2021/01/10/technology/parler-app-trump-free-speech.html>

[153] See e.g. Safiya Noble, *Algorithms of Oppression* (New York: NYU Press, 2018); Sarah Dobson, "Facebook targeting discriminatory job ads in Canada" (15 January 2020), online: *HR Reporter* <https://www.hrreporter.com/focus-areas/recruitment-and-staffing/facebook-targeting-discriminatory-job-ads-in-canada/325023>; Tracy Jan & Elizabeth Dwoskin, "Facebook agrees to overhaul targeted advertising system for job, housing and loan ads after discrimination complaints", *Washington Post* (19 March 2019), online: <https://www.washingtonpost.com/business/economy/facebook-agrees-to-dismantle-targeted-advertising-system-for-job-housing-and-loan-ads-after-discrimination-complaints/2019/03/19/7dc9b5fa-4983-11e9-b79a-961983b7e0cd_story.html>; and Ava Kofman & Ariana Tobin, "Facebook Agreed Not to Let Its Ads Discriminate. But They Still Can" (19 December 2019), online: *Mother Jones* <https://www.motherjones.com/politics/2019/12/facebook-agreed-not-to-let-its-ads-discriminate-but-they-still-can/>.

[154] See e.g., Canada, Parliament, House of Commons, Standing Committee on Access to Information, Privacy and Ethics, *Democracy under Threat: Risks and Solutions in the Era of Disinformation and Data Monopoly*, 42nd Parl, 1st Sess (December 2018) (Chair: Bob Zimmer).

[155] See e.g., Laura Ryckewaert, "MPs join fight to stamp out COVID-19 disinformation that's 'spreading faster than the virus'", *Hill Times* (8 April 2020), online: <https://www.hilltimes.com/2020/04/08/mps-join-fight-to-stamp-out-covid-19-disinformation-spreading-faster-than-the-virus/242558>; Samantha Bradshaw & Philip N Howard, "The Global Disinformation Order 2019: Global Inventory of Organised Social Media Manipulation" (2019), online (pdf): *Oxford Internet*

Pornography sharing platforms such as Pornhub, owned by the Canadian company MindGeek,[156] routinely host sexually abusive materials, such as intimate images or video recordings of sexual activities, which were taken and/or distributed without consent, including where the sexual activity itself was not consensual (i.e., sexual assault or rape).[157] Such non-consensual distribution of intimate images (NCDII) also routinely occurs across social media platforms such as Reddit, Snapchat, and Facebook, and is often shared across multiple platforms simultaneously.[158]

At the same time, users from historically marginalized groups and communities who use digital platforms to speak out against and call attention to injustices— including injustices facilitated by or that occur on the platforms themselves—have disproportionately experienced online censorship by those platforms. This has taken the form of wrongful takedowns of content, seemingly selective application of content moderation policies, and suspended or banned accounts and pages.[159] Users

---

Institute <https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf>; Mike Isaac & Kevin Roose, "Disinformation and fake news spreads over WhatsApp ahead of Brazil's presidential election", *Independent* (21 October 2018), online: <https://www.independent.co.uk/news/world/americas/brazil-election-2018-whatsap-fake-news-presidential-disinformation-a8593741.html>; and Priyanjana Bengani, "India had its first 'WhatsApp election.' We have a million messages from it", *Columbia Journalism Review* (16 October 2019), online: <https://www.cjr.org/tow_center/india-whatsapp-analysis-election-security.php>.

[156] Christopher Reynolds, "More than 70 MPs, senators call for criminal investigation into Pornhub's Canadian owners", National Post (15 March 2021), online: <https://nationalpost.com/news/canada/more-than-70-lawmakers-call-for-criminal-investigation-of-mindgeek>

[157] See e.g., Samantha Cole & Emanuel Maiberg, "Pornhub Doesn't Care", *Vice* (6 February 2020), online: <https://www.vice.com/en_ca/article/9393zp/how-pornhub-moderation-works-girls-do-porn>; Laila Mickelwait, "Time to shut Pornhub down", *Washington Examiner* (9 February 2020), online: <https://www.washingtonexaminer.com/opinion/time-to-shut-pornhub-down>; Bonnie Allen, "Revenge porn and sext crimes: Canada sees more than 5,000 police cases as law marks 5 years", *CBC News* (24 December 2019), online: <https://www.cbc.ca/news/canada/saskatchewan/revenge-porn-and-sext-crimes-canada-sees-more-than-5-000-police-cases-as-law-marks-5-years-1.5405118>; Olivia Solon, "Inside Facebook's efforts to stop revenge porn before it spreads", *NBC News* (18 November 2019), online: <https://www.nbcnews.com/tech/social-media/inside-facebook-s-efforts-stop-revenge-porn-it-spreads-n1083631>.

[158] Nicola Henry, Asher Flynn & Anastasia Powell, "Responding to 'revenge pornography': Prevalence, nature and impacts" (March 2019) at 35, online (pdf): Australian Institute of Criminology <https://www.aic.gov.au/sites/default/files/2020-05/CRG_08_15-16-FinalReport.pdf>

[159] Sam Levin, "Facebook temporarily blocks Black Lives Matter activist after he posts racist email", *Guardian* (12 September 2016), online: <https://www.theguardian.com/technology/2016/sep/12/facebook-blocks-shaun-king-black-lives-matter>; Tracey Jan & Elizabeth Dwoskin, "A white man called her kids the n-word. Facebook stopped her from sharing it", *Washington Post* (31 July 2017), online: <https://www.washingtonpost.com/business/economy/forfacebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html?utm_term=.451805b729db>; Kate Crawford & Tarleton Gillespie, "What is a flag for? Social media reporting tools and the vocabulary of complaint" (2016) 18:3 New Media & Society 410 at 411 ("So began a public controversy in which Facebook was accused of hypocrisy and homophobia, with critics noting that gay kisses were being flagged and removed while straight kisses went unremarked."); Julia Angwin & Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children", *ProPublica* (28 June 2017), online: <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>; Dottie Lux & Lil Miss Hot Mess, "Facebook's Hate Speech Policies Censor Marginalized Users", *Wired* (14 August 2017), online: <https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalizedusers/>; Sarah Myers West, "Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms" (2017) 5:3 Media and Communication 28 ("Shortly afterward, Heather Bays, a maternity photographer, had her Instagram account shut down after receiving a negative comment on a photo of her breastfeeding her daughter."); and Nellie Bowles & Cara Buckley, "Rose McGowan's Twitter Account Locked After Posts About Weinstein", *New York Times* (12 October 2017), online: <https://www.nytimes.com/2017/10/12/arts/rose-mcgowan-twitter-weinstein.html>.

who are already marginalized and made vulnerable by existing structures of power are silenced on and by platforms even as these companies cite commitments to freedom of expression to justify their leniency towards the kinds of abusive content and perpetrators of abuse being spoken out against.[160]

The above-mentioned harmful consequences, patterns, and dynamics have been occurring alongside the outsized growth and power that have come to be associated with digital platforms.[161] No longer are they scrappy upstart websites or obscure corners of the Internet; to much of the world's population, including in countries such as Myanmar, Malaysia, the Philippines, and Indonesia, digital platforms such as Facebook *are* the Internet.[162] Even where that is not the case, such as in Canada, digital platforms now occupy a powerful position with significant influence over politics, public policy, public discourse, human behaviour, and cultural and sociopolitical norms equivalent to or greater than that of some governments (but, saliently, without the corresponding responsibilities and obligations normally attached to public service in democratic societies). Kate Klonick has labelled online platforms the "new governors", due in part to how their internal and unilaterally determined content policies regularly displace formal laws and legal systems of governance over online expression in practice[163]—with consequences that reach far beyond speech itself, as demonstrated in the examples above.

The formidable power of online platforms to shape day-to-day behaviour, cultural norms, and sociopolitical forces domestically and internationally, combined with the high stakes for historically marginalized and vulnerable communities, human rights, and democratic values and institutions, have coalesced into a rising tide of criticism aimed at both online platforms and the laws that protect them.[164]

---

[160] See e.g., Kari Paul, "Zuckerberg defends Facebook as bastion of 'free expression' in speech", *Guardian* (17 October 2019), online: <https://www.theguardian.com/technology/2019/oct/17/mark-zuckerberg-facebook-free-expression-speech>; and Cecilia Kang & Kate Conger, "Inside Twitter's Struggle Over What Gets Banned", *New York Times* (10 August 2018), online: <https://www.nytimes.com/2018/08/10/technology/twitter-free-speech-infowars.html>.

[161] "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. [...] 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." Joris van Hoboken & Daphne Keller, "Design Principles for Intermediary Liability Laws" (8 October 2019), A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression at 3, online (pdf): *Institute for Information Law* <https://www.ivir.nl/publicaties/download/Intermediary_liability_Oct_2019.pdf>.

[162] See e.g., Brandon Paladino, "Democracy Disconnected: Social Media's Caustic Influence On Southeast Asia's Fragile Republics" (July 2018) at 6, online (pdf): *Brookings Institution* <https://www.brookings.edu/wp-content/uploads/2018/07/FP_20180725_se_asia_social_media.pdf>.

[163] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harvard Law Review 1598.

[164] See e.g., "Ottawa should stand up to Big Tech on privacy and democracy", *Toronto Star* (25 February 2019), online: <https://www.thestar.com/opinion/editorials/2019/02/25/ottawa-should-stand-up-to-big-tech-on-privacy-and-democracy.html>; "Poll demonstrates support for strong social media regulations to prevent online hate and racism" (25 January 2021), online: *Canadian Race Relations Foundation* <https://www.crrf-fcrr.ca/en/news-a-events/media-releases/item/27349-poll-demonstrates-support-for-strong-social-media-regulations-to-prevent-online-hate-and-racism>; Matt Laslo, "The Fight Over Section 230—and the Internet as We Know It", *Wired* (13 August 2019), online: <https://www.wired.com/story/fight-over-section-230-internet-as-we-know-it/>; "Internet firms face a global techlash", *Economist* (12 August 2017), online: <https://www.economist.com/international/2017/08/10/internet-firms-face-a-global-techlash>; Kari Paul, "A brutal year: how the 'techlash' caught up with Facebook, Google and Amazon", *Guardian* (28 December 2019), online: <https://www.theguardian.com/technology/2019/dec/28/tech-industry-year-in-review-facebook-google-amazon>; Gallup, Inc., "Techlash? America's Growing Concern with Major Technology Companies" (2020), online

There is increasing recognition that laws initially written for 'mere conduits' and 'passive hosts' such as Internet service providers do not adequately take into account digital platforms' sociocultural and sociopolitical roles, function, power, control, and influence which has since come to characterize them.[165] As a result, gender equality advocates, human rights experts, TFGBV-focused scholars and lawyers, community-based organizations combating violence against women and girls, lawmakers, and various governments, in addition to other stakeholders and decision-makers, have begun drawing the conclusion that traditional intermediary liability frameworks may be standing in the way of effectively addressing the most pressing problems associated with digital platforms today.[166] At the same time, platform companies' own efforts to address TFGBV have proven ineffective, insufficient, or counter-productive, such that self-regulation, on its own, cannot be relied upon as a solution.[167]

Part 3 of the report will delve further into how digital platforms specifically facilitate, perpetuate, and have attempted to address TFGBV, with a focus on expression-based TFGBV. Section 3.1 provides an overarching introduction, including defining what a 'platform' is for the purposes of this report, how digital platforms are central sites of TFGBV, and the fundamental role of such platforms' designs and business models in exacerbating TFGBV. Section 3.2 details unique characteristics of expression-based TFGBV facilitated by digital platforms, including its function as a weaponization of speech against women and intersecting marginalized identities; its networked and distributed nature; its socialization and gamification, and the propensity of platformed TFGBV to normalize and escalate violence against women, girls, and other marginalized identifies. Section 3.3 examines key content moderation features that digital platforms have implemented, including their effectiveness in responding to TFGBV. Section 3.4 presents three overarching critiques of platforms' approach to TFGBV, including: inconsistent use of 'free speech' rhetoric as a self-serving defence; responding reactively and arbitrarily on a 'damage control' basis driven by public outcry or threat of regulation; and being unable to resist inherent conflicts of interest where digital platforms' own business models and political ties incentivize them against making any substantive progress towards mitigating TFGBV on their platforms.

(pdf): *Knight Foundation* <https://knightfoundation.org/wp-content/uploads/2020/03/Gallup-Knight-Report-Techlash-Americas-Growing-Concern-with-Major-Tech-Companies-Final.pdf>;

[165] See e.g., Natalia Homchick, "Reaching Through the "Ghost Doxer:" An Argument for Imposing Secondary Liability on Online Intermediaries" (2019) 76:3 Washington and Lee Law Review 1307 at 1316-17; and Danielle Keats Citron & Benjamin Wittes, "The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity" (2017) University of Maryland Francis King Carey School of Law Legal Studies Research Paper at 20.

[166] See e.g., Jane Bailey & Valerie Steeves, "Submission to the Standing Committee on Justice & Human Rights Regarding Online Hate" (9 May 2019), online (pdf): *Canada House of Commons Standing Committee on Justice and Human Rights* <https://www.ourcommons.ca/Content/Committee/421/JUST/Brief/BR10520601/br-external/BaileyJane-e.pdf>; Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (21 August 2019), online: *Knight First Amendment Institute* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment>; and Angela Chen, "The legal crusader fighting cyber stalkers, trolls, and revenge porn", *MIT Technology Review* (26 August 2019), online: <https://www.technologyreview.com/2019/08/26/133255/carrie-goldberg-nobodys-victim-revenge-porn-sexual-privacy-section-230-cyber-crimes/>.

[167] For more details, see Section 3.4 ("Critiques of Platform Approaches to Speech-Based TFGBV").

3.1. How Digital Platforms Facilitate TFGBV

This section of the report will, first, introduce and define digital platforms for the purposes of this report; second, discuss how digital platforms are central sites of TFGBV; and third, examine how platforms' user features and business models make them optimized for TFGBV to proliferate.

3.1.1. What Are Digital Platforms?

Digital platforms, or online platforms, are Internet-based services accessed through websites or apps which are characterized by facilitating activity among and between the platform's users. Gillespie describes online platforms as "online services that a) host, organize, and circulate users' shared content or social interactions for them, b) without having produced or commissioned (the bulk of) that content, c) built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising, and [/or] profit."[168] Not all digital platforms are for-profit—for example, Wikipedia, an online encyclopedia written entirely by users on a volunteer basis and home to vigorous debate among such users, is a non-profit platform.[169] McNamee and Fernández Perez further provide a classification system that categorizes platforms based on the type of relationships between the platform, individual users, and third-party businesses (such as advertisers or retailers) and on which of these relationships are commercially transactional in nature.[170]

Examples of digital platforms include the following:

- social media platforms (e.g., Facebook, Twitter, Instagram, Reddit, TikTok);
- personal writing or blogging websites (e.g., WordPress, Medium);
- video-sharing and livestreaming platforms (e.g., YouTube, Vimeo, TikTok, Twitch);
- online dating apps (e.g., OkCupid, Tinder, Grindr);
- massively multiplayer online games (MMOG or MMO)—including their most popular subset, massively multiplayer online roleplaying games (MMORPG)—(e.g., World of Warcraft, EVE Online, Overwatch, League of Legends);
- pornography or sexual service platforms (e.g., Pornhub, XVideos, Chaturbate);
- search engines (e.g., Google Search, Bing, DuckDuckGo);
- review websites (e.g., Yelp, Travel Advisor);
- online commerce websites (e.g., Etsy, Amazon, eBay);
- payment processors and crowdfunding platforms (e.g., Patreon, PayPal, GoFundMe, Kickstarter); and

---

[168] Tarleton Gillespie, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media* (New Haven: Yale University Press, 2018) at 18.

[169] Wikimedia Foundation, "About" (2021), online: *Wikimedia Foundation* <https://wikimediafoundation.org/about/>.

[170] Joe McNamee & Maryant Fernández Pérez, "Fundamental Rights and Digital Platforms in the European Union: a Suggested Way Forward" in Luca Belli & Nicolo Zingales, eds, *Platform Regulations How Platforms are Regulated and How They Regulate Us* (Rio de Janeiro: FGV Direito Rio, 2017) 100 at 100.

- "gig economy" platforms that connect users to other users offering services such as ridesharing, short-term vacation rentals, food and grocery delivery, or completion of tasks (e.g., Uber, Lyft, Airbnb, Instacart, TaskRabbit).

While online platforms might be generally categorized for ease of reference, the boundaries of such categories are fluid and many platforms would fall into multiple categories simultaneously. For instance, Reddit involves social media, personal writing, video-sharing, and pornography all on the same platform, while Snapchat includes both private messaging and public posting.

Private messaging apps and workspace collaboration apps, such as WhatsApp, Telegram, Slack, or Keybase, may in some ways be considered to be types of online platforms, on the basis that they can also host and facilitate user interactions in large groups. What distinguishes these services, however, is their default private nature—for the most part, apps such as WhatsApp and Telegram are used between individuals to privately message each other, and are more akin to conventional texting, or short message service (SMS). Slack and Keybase require users to create private group spaces which are only accessible by authorized users, such as employees of a single company or members of a single team. There is no "public sphere", so to speak, only an indefinite number of private rooms that are unaware of and inaccessible to each other.[171] However, users have begun manipulating WhatsApp and Telegram in ways that blur this line, such as creating chat groups containing up to hundreds or thousands of people in order to spread disinformation leading up to elections;[172] and online abuse and harassment occurs just as easily on private platforms as on public ones.[173]

The term "platform", although popularized in public and academic discourse to refer to the kinds of companies listed above, has pre-existing meanings in adjacent contexts, such as in computing and Internet infrastructure.[174] For example, the computational meaning of "platform" refers to "infrastructure that supports the design and use of particular applications, be they computer hardware, operating systems, gaming devices, mobile devices or digital disc formats."[175] A platform is something that one can build more software on top of. The Windows and Android operating systems, Apple's iOS

---

[171] For a brief discussion on the distinct "private rooms" aspect of messenger apps in the context of platform regulation and content moderation, see e.g., Chand Rajendra-Nicolucci & Ethan Zuckerman, "Chat Logic: When you want a living room, not a town square" (13 November 2020), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/chat-logic-when-you-need-a-living-room-and-not-a-town-square>.

[172] See e.g., Caio Machado & Marco Konopacki, "Computational Power: Automated Use of WhatsApp in the Elections" (26 October 2018), online: *ITS Rio* <https://feed.itsrio.org/computational-power-automated-use-of-whatsapp-in-the-elections-59f62b857033>; Ayushman Kaul & Max Rizzuto, "UK-based far-right Telegram channels amplify disinfo targeting U.S. election integrity" (17 December 2020), online: *DFRLab* <https://medium.com/dfrlab/uk-based-far-right-telegram-channels-amplified-disinfo-targeting-u-s-election-integrity-b0d007b1375b>; Jacob Gursky et al, "Encrypted Propaganda: Political Manipulation via Encrypted Messaging Apps in the United States, India, and Mexico" (2020), online (pdf): *Center for Media Engagement* <https://mediaengagement.org/wp-content/uploads/2020/10/Encrypted-Propaganda-Political-Manipulation-Via-Encrypted-Messages-Apps-in-the-United-States-India-and-Mexico.pdf>; and Munsif Vengattil, Aditya Kalra & Sankalp Phartiyal, "In India election, a $14 software tool helps overcome WhatsApp controls", Reuters (15 May 2019), online: <https://www.reuters.com/article/india-election-socialmedia-whatsapp/in-india-election-a-14-software-tool-helps-overcome-whatsapp-controls-idUSKCN1SL0PZ>.

[173] See e.g., Caroline Sinders, "No one is talking about the biggest problem with Slack" (13 June 2019), online: *Quartz* <https://qz.com/1641708/slack-doesnt-care-that-you-cant-block-a-workplace-harasser/>.

[174] See e.g., "The 9 Types of Software Platforms" (12 June 2016), online: *Platform Hunt* <https://medium.com/platform-hunt/the-8-types-of-software-platforms-473c74f4536a>.

[175] Tarleton Gillespie, "The politics of 'platforms'" (2010) 12:3 New Media & Society 347 at 349.

and MacOS, Linux, Google's Chrome OS, Internet browsers such as Firefox and Safari, and cloud service providers such as Amazon Web Services and Microsoft Azure are all types of platforms and provide online services on some level, though not always directly to individual users.

For the purposes of this report, the terms "platform", "online platform", and "digital platform" will be used interchangeably to mean the kinds of online services that host and facilitate user-generated content and user interactions, as defined by Gillespie above. This report will focus predominantly on platforms such as social media websites, video-sharing platforms, online dating apps, pornography or sexual service platforms, and search engines. This focus is due to the ostensible and documented prevalence of TFGBV on digital platforms that mostly fall into these particular categories.

One particular type of platform that warrants specific attention in the context of TFGBV is the category of platforms that seem deliberately designed to encourage and profit from such abuse. These might be termed '**purpose-built platforms**', as opposed to 'platforms of general application' such as Facebook and Twitter, which are inclusive of user-generated content that constitutes TFGBV, but do not appear to exist exclusively to cater to such content. Examples of 'purpose-built platforms' are 'The Dirty' and 'She's A Homewrecker',[176] both seemingly designed to facilitate misogynistic character assassinations of young women. The founder of 'The Dirty' built the website on the basis of "encouraging his readers to rat each other out by emailing him 'the dirt' on one another. He'd block-quote his favorite emails in blog posts, often accompanied by images of the scantily clad, inebriated, and [allegedly] unfaithful".[177] Another example may be specific self-contained and distinct forums on larger platforms, such as focusing on individual subreddits for the purpose of the relevant analysis, rather than Reddit as a whole, where it is clear that a given subreddit only exists, for instance, predominantly to promote violence against women, exchange intimate images of women without their consent, or mock, deride, and incite contempt for and violence against a systemically oppressed group in society.[178]

What would be a balanced and proportionate liability framework for platforms of general application would likely not suffice to address TFGBV where such expression and conduct constitute the core business model or central service or commodity of a purpose-built platform.

3.1.2. Digital Platforms as Central Sites of TFGBV

Online platforms such as social media networks, discussion forums, search engines, and video sharing websites have become central venues of our personal, political, and professional lives. As the House of Commons Standing Committee on the Status of Women notes, "Some social media platforms have more 'citizens' than some countries, and these companies are making significant decisions, such as categories of identity and what constitutes online safety and violence. For instance, Facebook has 'over 22 million Canadians – 1.71 billion people globally – using Facebook each month.' Twitter told the

---

[176] Helen AS Popkin, "Website exposes 'homewreckers' — but doesn't break the law", *NBC News* (8 November 2013), online: <https://www.nbcnews.com/technolog/website-exposes-homewreckers-doesnt-break-law-8c11554126>. 'She's a Homewrecker' was reported in media coverage as one of the websites used in the events leading up to *Caplan v Atas*, 2021 ONSC 670. which established the tort of online harassment in Ontario. Kashmir Hill, "A Vast Web of Vengeance", *New York Times* (30 January 2021), online: <https://www.nytimes.com/2021/01/30/technology/change-my-google-results.html>

[177] Kate Knibbs, "Cleaning Up the Dirty", *Ringer* (19 April 2017), online: <https://www.theringer.com/2017/4/19/16041942/the-dirty-nik-richie-gossip-site-relaunch-4a086aa24536>.

[178] "Controversial Reddit communities" (11 April 2021), online: *Wikipedia* <https://en.wikipedia.org/wiki/Controversial_Reddit_communities>.

Committee it had 313 million users, and that it currently sees '500 million tweets on a single-day basis.'"[179] Research has shown that youth "have embedded networked technologies seamlessly into their social lives, using social media to explore their identities, deepen their connection with friends and family and explore their interests", with girls being "more likely to use social media for communication and identity play", demonstrated in higher usage of Facebook, Twitter, Instagram, Tumblr, and Pinterest, compared to boys.[180]

It thus comes as no surprise that online platforms are also central sites of TFGBV.[181] Facebook has been criticized for allowing pages glorifying intimate partner violence to stand, while at the same time taking down images of women breastfeeding.[182] Twitter has featured in multiple criminal harassment cases and also been accused of double-standards in enforcing its policies against abusive users, while it throws the book at users who are the recipients of abuse.[183] YouTube has been deemed an engine of right-wing radicalization due to its recommendations algorithm systematically and disproportionately promoting to viewers a networked group of content creators who espouse white supremacist and misogynistic ideologies.[184] Google Search has displayed top-rank search results that sexually objectified Black girls and perpetuated negative stereotypes of Black teenagers.[185] For years, Reddit earned a reputation for being a bastion of sexism, misogyny, and online violence and abuse against women and girls.[186] These are only system-level observations, before even accounting for the day-to-day deluge of

---

[179] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 50-51.

[180] Jane Bailey & Valerie Steeves, "Big Data, Social Norms and Discrimination: Lessons from The eGirls Project" (2015), at 1-2, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2016/05/Big-Data-Social-Norms-and-Discrimination-Paper.pdf>.

[181] "Women use social networking sites more often than men, but these sites have failed to fully respond to the concerns of their women users. In more than 4,000 cases of cyberstalking reported to Halt Online Abuse since 2000, 70% of victims were female. Presently, 82% of social media violence against women reported on Take Back the Tech!'s map happened on one of the big three sites – Facebook, Twitter or YouTube – with Facebook alone accountable for half." Sara Baker, "#WhatAreYouDoingAboutVAW Campaign: Social Media Accountability" (12 September 2014), online: *GenderIT.org* <https://genderit.org/feminist-talk/whatareyoudoingaboutvaw-campaign-social-media-accountability>.

[182] Simon Van Zuylen-Wood, "'Men Are Scum': Inside Facebook's War on Hate Speech", Vanity Fair (26 February 2019), online: <https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech>.

[183] *R v Elliott*, 2016 ONCJ 35; Ashley Csanady, "The Twitter trial you never heard about: Toronto man found guilty of harassing Michelle Rempel", *National Post* (29 January 2016), online: <https://nationalpost.com/news/politics/the-twitter-trial-you-never-heard-about-toronto-man-found-guilty-of-harassing-michelle-rempel>; and Aja Romano, "Twitter's suspension of Rose McGowan epitomizes the site's most infuriating problem", *Vox* (12 October 2017), online: <https://www.vox.com/culture/2017/10/12/16464752/twitter-suspended-rose-mcgowan>.

[184] Kevin Roose, "The Making of a YouTube Radical", *New York Times* (8 June 2019), online: <https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html> and Paul Lewis, "'Fiction is outperforming reality': how YouTube's algorithm distorts truth", *Guardian* (2 February 2018), online: <https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth>; and Rebecca Lewis, "Alternative Influence: Broadcasting the Reactionary Right on YouTube" (2018), online (pdf): *Data & Society* <https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf>; but see Mathew Ingram, "The YouTube 'radicalization engine' debate continues", (9 January 2020), online: *Columbia Journalism Review* <https://www.cjr.org/the_media_today/youtube-radicalization.php>.

[185] Safiya Noble, *Algorithms of Oppression* (New York: NYU Press, 2018) at 64-109.

[186] See e.g., Adrienne Massanari, "#Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures" (2017) 19:3 new media & society 329; Lindy West, "How Reddit's Ellen Pao survived one of 'the largest trolling attacks in history'", *Guardian* (22 December 2015), online: <https://www.theguardian.com/lifeandstyle/2015/dec/22/reddit-ellen-pao-trolling-revenge-porn-ceo-internet-misogyny>;

gender-based violence, abuse, or harassment that occurs in private or semi-private interactions between individual users or within groups of friends, classmates, or communities, which do not reach the level of media coverage or public discourse, but equally contribute to making digital platforms a toxic and hostile environment for women and girls.

The specific set of features common to most digital platforms, and features of technology more broadly, uniquely transforms and intensifies the nature and extent of gender-based violence, abuse, and harassment and its impacts on those targeted. First, the technological affordances of online platforms combined with their advertising-driven, thus attention- and 'engagement'-driven, business models seem almost specifically designed to optimize and maximize violent or abusive content and behaviours. For example, one factor that contributed to YouTube's rise as a propagator of misogynistic and racist content is that the company focused on growth at the expense of user well-being—ignoring their own employees' warnings and shutting down any measures that would have impeded user activity, such as changing the algorithm to promote more credible sources in lieu of conspiracy theories.[187]

Second, the efficiency, ease, and affordability with which one can engage in, automate, perpetuate, and multiply instances of abuse against a single individual or group of individuals lowers the cost of doing so to almost zero.[188] Nearly all of today's major digital platforms are free to use and quick to create accounts on, while other kinds of technology, such as stalkerware apps with highly sophisticated surveillance capabilities, are also free or obtainable at moderate monthly subscription fees. Third, anonymity and a sense of 'safety in numbers', where one is part of an online mob or coordinated campaign of attack, highly reduce the risks of engaging in violence, abuse, or harassment online, such as the risk of being identified or getting caught. The remoteness of interacting with others online, combined with anonymity, may also lead to disinhibition: "Perpetrators may feel less empathy and find it easier to be cruel when they cannot see or be seen by their target."[189]

Fourth, perpetrators learn to exploit, game, and work around digital platforms' affordances and content moderation features in order to enact abuse, as Dragiewicz et al. explain in the case of Twitter:

> For instance, the high level of anonymity on Twitter, the ease with which a user can create multiple accounts, and the platform's historical reluctance to police the free expression of their users, have combined to make the platform a hotbed of abuse and

---

and Molly Dragiewicz et al., "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 616.

[187]Max Bergen, "YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant", *Bloomberg* (2 April 2019), online: <https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant>; see also: "[A]s Alice Marwick and Rebecca Lewis (2017) argue, the diffuse online subculture associated with the "alt-right," which includes the misogynist "manosphere" of antifeminist men's groups, is significantly empowered by the ability to exploit the affordances and algorithmic characteristics of the contemporary digital media environment, to "manipulate news frames, set agendas, and propagate ideas." Molly Dragiewicz et al., "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 616.

[188] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 33; Jessica West, "Cyber-Violence Against Women" (May 2014) at 2, online (pdf): *Battered Women's Support Services* <https://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf>.

[189] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 33.

> harassment. Malicious users have appropriated Twitter's targeted advertisement feature to abuse transgender people, since it enables them to promote abusive posts that cannot be traced back to them. Other harassment tactics include 'tweet and delete' practices, in which abusers temporarily make available the private information of their targets but remove the content before it can be flagged and any disciplinary action taken by the platform. […]
>
> This sophisticated capacity to game the system's technological affordances as well as to game the media culture at large (through "weaponizing irony" for example) are significant dimensions of online misogyny and TFCC [technology-facilitated coercive control] alike; and present major governance challenges to platforms.[190]

These characteristic elements of not just TFGBV, but *platformed TFGBV*, result in additionally devastating consequences for the targets of such attacks, compared to if they occurred exclusively in physical spaces. First, that the abuse takes place online means it is no longer constrained by physical boundaries. For example, victim/survivors of TFCC have spoken of no longer being able to escape just by moving to a different city or country, because the "characteristics of digitally mediated communication such as storage, synchronicity, replicability, and mobility enhance abusers' ability to persistently intrude on their targets regardless of their location. As a result, TFCC expands abusers' sphere of control beyond previous spatial boundaries."[191] Outside the context of TFCC, online abuse can take on an element of omnipresence and relentlessness that intrudes particularly jarringly when the victim goes online while physically in a private or intimate location, such as their home or bedroom.[192] Second, TFGBV is characterized by the "near-limitless reach of the Internet", which

> exponentially multiplies the harm to one's reputation, social standing, future prospects, personal relationships and, even, personal security when intimate information (or in some cases, misinformation) about her is distributed through those means. The sheer magnitude of the exposure can itself be understood as an aggravating factor—transforming what might ordinarily be understood as a private law harm (say, defamation) into a criminal law one.[193]

Third, individuals can both intentionally and unintentionally cause their abuse of someone, as Dragiewicz et al. point out, to go viral across multiple platforms: "misogynist peer networks on social media mobilise to harass women using information and images provided by DV [domestic violence] perpetrators."[194] For example, "A number of high-profile [NCDII] victims have described anonymous online groups of men, motivated by a shared misogyny, persistently re-circulating images released

---

[190] Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 614, 616 (inline citations omitted).

[191] *Ibid* at 611 (inline citations omitted).

[192] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 33; Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 611.

[193] Jane Bailey & Carissima Mathen, "Technology-Facilitated Violence against Women & Girls: Assessing the Canadian Criminal Law Response" (2019) 97 La Revue du Barreau Canadien 664 at 677.

[194] Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 614.

without their consent to maximum reputational damage."[195] Fourth, digital platforms, particularly social media websites, can give rise to what Markwick and boyd have termed 'context collapse', where multiple social spheres that may have otherwise remained separate converge (for instance, family members, close friends, work colleagues, and sports teammates).[196] This convergence "can exacerbate the effects of [TFGBV], as perpetrators can upload defamatory or humiliating content 'that effectively poisons the user's social world'"[197]—especially if the perpetrator recruits their own social network and/or members of the targeted person's social network to extend and amplify the abuse. Fifth, the permanence and persistence of content on the Internet, particularly in the case of abusive content that specifically targets an individual, such as intimate photos posted without consent, can lead to lasting damage and countless instances of revictimization over time. Dragiewicz et al. write:

> The storage, reach, and replicability [...] of digital media communication and content means that texts and media objects used in abuse may be persistently visible and connected to the victim's identity. For example, some victims of image based sexual abuse experience constant anxiety about who has seen an image and where it may next appear [...]. Young women interviewed about abusive text messages from violent partners say that "it gets into your head" more than it does in person as they have a permanent record on their phones. Because they have their phones with them all the time, they say, the abuse "stays with you" [...].[198]

Despite the widely documented proliferation of TFGBV on digital platforms and its impacts on women and girls—as well as online abuse perpetrated against platform users with other marginalized identities—platform companies have on the whole done poorly to address the issue.[199] Those who have been the targets of TFGBV have criticized the major platform companies for ignoring individual requests for help in specific cases of abuse, as well as ignoring the broader issue of TFGBV while continuing to support and build features that contribute to optimizing the platform for abuse (such as when Snapchat released a location-tracking feature that automatically broadcast the user's precise location to their entire friends list upon opening the app).[200] Digital platforms have also displayed a certain degree of selective attentiveness and double standards to the extent they have developed and applied their content moderation policies, such as determining that abusive content does not violate any policies while removing the content or suspending the account of users who were the victims of abuse, or for

---

[195] *Ibid*. They also note: "Salter's (2017b) analysis of the 'Gamergate' controversy provides a case study of viral TFCC, in which an embittered video game developer wrote and circulated a defamatory post about his ex-partner, ultimately recruiting tens of thousands of social media users into a sustained campaign of targeted abuse and harassment // that attracted global media attention. Such misogynist campaigns can be highly organised and coordinated across multiple online platforms, exploiting the specific affordances and loopholes of each platform. *Ibid* at 613-14 (inline citations omitted).

[196] *Ibid* at 613.

[197] *Ibid*.

[198] *Ibid* at 611 (in-text citations omitted).

[199] Digital platform companies' efforts to address TFGBV on their platforms are discussed in Section 3.3 ("Platform Content Moderation Policies and Practices") and Section 3.4 ("Critiques of Platform Approaches to Speech-Based TFGBV").

[200] Maggie Nicholson, "Snapchat's new Snap Map feature may pose a threat to victims and survivors" (19 July 2017), online: *National Latin@ Network*, <https://enblog.nationallatinonetwork.org/snapchats-new-snap-map-feature-may-pose-a-threat-to-victims-and-survivors/>.

pointing out the existence of such abuse.[201] As Dragiewicz et al. point out, "The slow and uneven response of online platforms to women's complaints of harassment and abuse suggests that the issue of gender-based violence and harassment has not been a priority within the tech industry. This exemplifies how the social context of gendered inequality enables abuse by creating the conditions in which women's efforts to seek assistance are blocked or ignored."[202] Many have connected these conditions to the lack of gender, race, and other forms of diversity throughout the technology sector, including among employees (with influence or decision-making power), management, and leadership at the major digital platform companies.[203]

At the same time, those targeted by TFGBV on digital platforms have limited options for redress or accessing justice through systems outside of the platform. Many forms of harassment and abuse may not amount to causes of action that would succeed in a civil lawsuit or chargeable crimes under criminal law. The facts of a particular situation may not meet the elements of any legal test, and the law has not yet caught up to the many new ways in which technology enables what should be actionable or chargeable acts of violence, abuse, or harassment. Moreover, pursuing a legal claim or criminal charge would require being able to identify the specific wrongdoer to sue or lay charges against. This may be difficult in the case of anonymous abusers,[204] or in the case of mob-style abuse where hundreds or thousands of individuals may only send one or two messages insufficient to ground legal action, but results in a level of impact and harm that warrants legal recognition and redress for the victim.

Even in cases that more clearly from the outset constitute recognized legal wrongs (e.g., invasion of privacy, intentional infliction of mental suffering) or criminal offences (e.g., criminal harassment, stalking, intimidation), there are significant barriers that the justice system presents to individuals who wish to pursue or report a claim. In the case of civil lawsuits, the recipient of abuse may not have the time or financial resources to be able to afford going through the litigation process and all that it would entail. In the case of a criminal offence, law enforcement and police services have long and widely documented track records of dismissing violence against women and girls, including sexual assault, intimate partner violence, and TFGBV[205]—which combines lack of understanding of gender-based violence, abuse, and harassment with inadequate appreciation of non-physical harms and ignorance regarding technology and how technosocial forces operate in the context of gendered violence and abuse. This lack of understanding adds to the dismissive and victim-blaming attitudes that women and girls already encounter when reporting gender-based violence and abuse to police, even where the

---

[201] See Section 3.4.1 ("Inconsistent and Unprincipled: 'Free Speech' Rhetoric").

[202] Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 618.

[203] See e.g., Donovan X Ramsey, "Twitter's White-People Problem", *The Nation* (6 January 2016), online: <https://www.thenation.com/article/archive/twitters-white-people-problem/>; and "Mark Zuckerberg Hates Black People" (17 May 2017), online: *DiDi Delgado*, <https://thedididelgado.medium.com/mark-zuckerberg-hates-black-people-ae65426e3d2a>.

[204] Increasingly, however, both the law and platform companies themselves are providing for processes where identity information may or must be disclosed under certain circumstances.

[205] Molly Dragiewicz et al, "Technology-facilitated coercive control" in Walter S. DeKeseredy, Callie Marie Rennison & Amanda K. Hall-Sanchez, eds, *The Routledge International Handbook of Violence Studies* (Abingdon: Routledge, 2018) 244 at 247-28; Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 86-87; Jessica West, "Cyber-Violence Against Women" (May 2014) at 19, online (pdf): *Battered Women's Support Services* <https://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf>.

harm is overtly physical and no technology is involved.[206] Nor is this lack of understanding limited to police officers; lawyers, judges, and other actors in the Canadian legal system also often lack the understanding necessary to fully and appropriately handle or adjudicate claims concerning TFGBV, as the following passage demonstrates:

> Regardless of which side they took on the verdict in my case, many observers commented on how the judge's decision revealed what was, on his part, a very limited grasp of how the Internet works, and Twitter in particular, which is where the harassment occurred. […] But how do you explain to someone who has never used Twitter what it's like to be someone who uses Twitter as your primary means of sharing your voice with the world? How do you explain to that person who never uses Twitter just how much it impacted your life to no longer be able to use it freely, and to feel fear every time you sign in that your harasser is going to be there to greet you? The answer is that you can't, but that person who doesn't use the Internet will have the power to determine the official public narrative of what happened to you on the Internet.[207]

Thus, those who are impacted by TFGBV either are pre-emptively dissuaded from engaging the formal legal system at all, or try but are dismissed, revictimized, or offered victim-blaming and potentially dangerous advice such as simply shutting down their social media accounts and 'staying offline'.[208]

As a result of all of the above forces leaving many targets of TFGBV with nowhere to turn, pressure has steadily risen on the platform companies themselves and on governments and lawmakers to impose legal obligations on these companies,[209] as the actors best positioned to address the types of harm that their services enable and facilitate, in some cases uniquely so. Digital platform companies also benefit materially from TFGBV where they are ad-driven or their business model relies on maximizing user activity and extracting as much user data as possible, thus creating perverse incentives to refrain from implementing measures that would reduce TFGBV on their platforms. Some jurisdictions have begun imposing or are considering imposing liability on digital platforms through legislation, making them

---

[206] "Even when a specialized legal framework is in place, legal and regulatory mechanisms, including law enforcement officials, are not always trained or equipped to implement it effectively owing to the lack of adequate gender-sensitive training and the general perception that online abuse is not a serious crime." *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNGAOR, 38th Sess, UN Doc A/HRC/38/47 (18 June 2018) at 85; "The lack of response from police with regards to violence against women is a regular pattern known to most women and helps explain why very few women choose to report sexual assault and abuse to the police. This is reflected in our survey, in which only 6.7% of women responded that they appealed to the police and of those, only half saw their attacker arrested or obtained a restraining order." *ibid* at 19.

[207] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 45, quoting Stephanie Guthrie, whose complaints led to the criminal harassment case *R v Elliott*, 2016 ONCJ 35.

[208] Jane Bailey, Valerie Steeves & Suzie Dunn, "Submission to the Special Rapporteur on Violence Against Women Re: Regulating Online Violence and Harassment Against Women" (27 September 2017) at 12 and 14, online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2017/12/Bailey-Steeves-Dunn-Submission-27-Sep-2017.pdf>.

[209] See e.g., Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 615, Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 51, 57; Raine Liliefeldt, "How cyberviolence is threatening and silencing women" (14 June 2018), online: *Policy Options* <https://policyoptions.irpp.org/magazines/june-2018/how-cyberviolence-is-threatening-and-silencing-women/>.

legally responsible for expeditiously removing harmful content created and posted by users. These platform liability frameworks, both enacted and proposed, are discussed below in Part 4 ("Platform Liability for TFGBV in Canadian Law") and Part 5 ("Platform Liability Models: Jurisdictional Scan").

There are many legal challenges and complexities associated with imposing liability on digital platforms for user wrongdoing. At the forefront are concerns regarding the right to freedom of expression, and how to meaningfully implement harm reduction measures without leading to inadvertent removal of beneficial content (including political expression by women, girls, and other marginalized identities, and information about issues such as sexual health, sex education, and reproductive rights). The Citizen Lab succinctly summarizes the central tensions involved:

> Liability-based mechanisms have almost uniformly led to poor outcomes. There is ample evidence of the extensive over-enforcement that occurs when intermediaries are compelled to identify allegedly abusive users or to remove allegedly illegal content under threat of liability. This is particularly so in the absence of narrowly defined court orders or other legal safeguards. Such over-enforcement inevitably leads to disproportionate interference with the rights to privacy and free expression, including the rights of women and girls. On the other hand, current mechanisms are predominantly voluntary, leading to inconsistent outcomes and under-enforcement related to harmful and abusive content. In either case, the various economic, social, and moral harms that flow from online and technology-facilitated violence, harassment, and abuse often remain unmitigated.[210]

Indicating the importance of applying a cross-disciplinary approach to online platform liability, the UN Special Rapporteur on violence against women, its causes and consequences, in her report about online violence against women and girls from a human rights perspective, recommended that her office coordinate with that of the Special Rapporteur on the right to privacy and the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, regarding how best to address technology-related human rights violations, including TFGBV.[211] In addition to the human rights implications, digital platform liability also raises issues concerning jurisdiction (as the platforms themselves may not be based in Canada), enforcement, and practical workability given the astronomical scale of content that must be moderated.

3.1.3. Platform Design and Business Models Optimize for TFGBV

The main ad-driven business model of several major digital platforms combines with their technological affordances to optimize their environments for the proliferation of speech (or expression)-based abuse and harassment, including hate speech and harassing or abusive campaigns

---

[210] Robert J. Deibert et al, "Submission of the Citizen Lab (Munk School of Global Affairs, University of Toronto) to the United Nations Special Rapporteur on violence against women, its causes and consequences, Ms. Dubravka Šimonović" (2 November 2017) at 11, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2017/11/Final-UNSRVAG-CitizenLab.pdf>.

[211] *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNGAOR, 38th Sess, UN Doc A/HRC/38/47 (18 June 2018) at 92

targeting individuals.[212] Platforms' features and content moderation measures are exploited or gamed by abusive users to achieve their objectives in harassing and silencing their victims, while commercial incentives and platforms' business logic tacitly and in some cases explicitly discourage interventions that may impede such abuse.[213]

In developing the concept of 'platformed racism', Ariadna Matamoros-Fernández specified that the term does two things:

> it (1) evokes platforms as tools for amplifying and manufacturing racist discourse both by means of users' appropriations of their affordances and through their design and algorithmic shaping of sociability and (2) suggests a mode of governance that might be harmful for some communities, embodied in platforms' vague policies, their moderation of content and their often arbitrary enforcement of rules.[214]

These attributes and impacts of digital platforms and platform governance operate similarly in the context of other and intersecting forms of systemic oppression, such as sexism, misogyny, homophobia, and transphobia, for example.

Numerous additional scholars, journalists, and human rights advocates have noted that digital platforms' decisions with respect to content moderation, community standards, and feature design are largely driven by and responsive to the platforms' own economic interests, particularly in their sustained failures to address online abuse: "Hate online triggers traffic to online content and interaction about it, which translates in economic revenue for platforms and could explain their lack of response to online abuse."[215] According to Philippa Hall, "hate speech can be conceptualized as a by-product of the current privatized form of internet technology developed to serve the requirements of globalized capital".[216]

---

[212] See e.g., Anat Ben-David & Ariadna Matamoros-Fernández, "Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain" (2016) 10 International Journal of Communication 1167.

[213] See e.g., *ibid* at 1168: "We join with critical studies of social media, which argue that the corporate logic of these platforms, alongside their technical intrinsic characteristics (algorithms, buttons, and features), condition the social interactions they host, as well as effect broader social and political phenomena."; and Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930 at 933.

[214] *Ibid*.

[215] Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 617 (in-text citations omitted); See also Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930 at 933; Philippa Hall, "Disability Hate Speech: Interrogating the Online/Offline Distinction" in Karen Lumsden & Emily Harmer, eds, *Online Othering: Exploring Digital Violence and Discrimination on the Web* (London: Palgrave Macmillan, 2019) 309 at 317; Rebecca Lewis, "Alternative Influence: Broadcasting the Reactionary Right on YouTube" (2018) at 43, online (pdf): *Data & Society Research Institute*, <https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf>; Taina Bucher & Anne Helmond, "The Affordances of Social Media Platforms" in Jean Burgess, Thomas Poell & Alice Marwick, eds, *The SAGE Handbook of Social Media* (London and New York: SAGE Publications Ltd., 2017).

[216] Philippa Hall, "Disability Hate Speech: Interrogating the Online/Offline Distinction" in Karen Lumsden & Emily Harmer, eds, *Online Othering: Exploring Digital Violence and Discrimination on the Web* (London: Palgrave Macmillan, 2019) 309 at 317.

Rebecca Lewis has demonstrated how YouTube in particular has contributed to the creation of the Alternative Influence Network (AIN), a network of "political influencers who adopt the techniques of brand influencers to build audiences and 'sell' them on far-right ideology… cross-promotion of ideas forms a broader 'reactionary' position: a general opposition to feminism, social justice, or left-wing politics".[217] Lewis concludes:

> *YouTube is built to incentivize the behavior of these political influencers.* YouTube monetizes influence for everyone, regardless of how harmful their belief systems are. The platform, and its parent company, have allowed racist, misogynist, and harassing content to remain online–and in many cases, to generate advertising revenue–as long as it does not explicitly include slurs. YouTube also profits directly from features like Super Chat which often incentivizes "shocking" content. In other words, the type of content and engagement created by the AIN fits neatly into YouTube's business model.[218]

Media coverage has revealed how executives at both YouTube and Facebook ignored or shelved internal research at each company that demonstrated each platform's propensity to systematically amplify and promote abusive speech and hate-based rhetoric. Such executives also deliberately quashed or undermined efforts to mitigate such effects, as proposed programs would run counter to their pursuit of maximized user engagement and company growth. Bloomberg journalist Mark Bergen reported the following about YouTube:

> Wojcicki and her deputies know this [that YouTube promotes abusive content by design]. In recent years, scores of people inside YouTube and Google, its owner, raised concerns about the mass of false, incendiary and toxic content that the world's largest video site surfaced and spread. One employee wanted to flag troubling videos, which fell just short of the hate speech rules, and stop recommending them to viewers. Another wanted to track these videos in a spreadsheet to chart their popularity. A third, fretful of the spread of "alt-right" video bloggers, created an internal vertical that showed just how popular they were. Each time they got the same basic response: Don't rock the boat.
>
> The company spent years chasing one business goal above others: "Engagement," a measure of the views, time spent and interactions with online videos. Conversations with over twenty people who work at, or recently left, YouTube reveal a corporate leadership unable or unwilling to act on these internal alarms for fear of throttling engagement.[219]

---

[217] Rebecca Lewis, "Alternative Influence: Broadcasting the Reactionary Right on YouTube" (2018) at 1, online (pdf): *Data & Society Research Institute*, <https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf>.

[218] *Ibid* at 43 (first-line emphasis in original). See also Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930, documenting similar mechanisms in networked racism against Indigenous peoples in Australia.

[219] Mark Bergen, "YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant" (2 April 2019), online: *Bloomberg* <https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant>.

Similar dynamics between concerned employees and bottom-line-minded executives played out at Facebook, which "knew that its recommendation algorithm exacerbated divisiveness… Building features to combat that would require the company to sacrifice engagement—and by extension, profit—according to a later document from 2018 which described the proposals as 'antigrowth' and requiring 'a moral stance.'"[220] Other internal documents at Facebook included the following findings:

> "Our algorithms exploit the human brain's attraction to divisiveness," read a slide from a 2018 presentation. […] Worse was Facebook's realization that its algorithms were responsible for their growth [of extremist content]. The 2016 presentation states that "64% of all extremist group joins are due to our recommendation tools" and that most of the activity came from the platform's "Groups You Should Join" and "Discover" algorithms: "Our recommendation systems grow the problem."[221]

Internal researchers at Facebook also found that much of the problematic content came from "hyperactive users—who were usually far more partisan than average users and engaged in behaviour similar to spammers", and proposed a design tweak that would have reduced such content's reach (titled 'Sparing Sharing').[222] Reportedly, Facebook CEO Mark Zuckerberg approved the program only after demanding its impact be cut by 80 percent,[223] and stated that "he was losing interest in the effort to change the platform for the good of its users and asked not to have that subject brought to him again."[224]

Under platforms' current models, algorithms respond to indicators of preference, such as sharing or reposting, the 'like' button on Facebook,[225] view time on YouTube, or retweeting or 'favouriting' on Twitter, to push more of the same and similar toxic content to encase each user in their own personalized 'public sphere' over time, while "award[ing] it [the abusive content] with certain legitimacy" and boosting such contents' algorithmic rankings generally.[226] Thus, those who engage with sexist and misogynistic content on digital platforms such as social media are more likely to see more such content in their feeds and in recommendation panels, increasing in intensity and extremity of views as well as in frequency of exposure over time with continued engagement. Simultaneously, algorithmically boosted content may push down and out of the user's information environment

---

[220] Adam Smith, "Facebook knew its algorithm made people turn against each other but stopped research", *Independent* (28 May 2020), online: <https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-algorithm-bias-right-wing-feed-a9536396.html>.

[221] Jeff Horwitz & Deepa Seetharaman, "Facebook Executives Shut Down Efforts to Make the Site Less Divisive", Wall Street Journal (26 May 2020), online: <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>.

[222] *Ibid*.

[223] "Facebook's employee revolt" (4 June 2020), online (podcast): *Reset* <https://podcast9.com/share/episode/fyWhwBHTI/facebook-s-employee-revolt>.

[224] Adam Smith, "Facebook knew its algorithm made people turn against each other but stopped research", *Independent* (28 May 2020), online: <https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-algorithm-bias-right-wing-feed-a9536396.html>.

[225] Anat Ben-David & Ariadna Matamoros-Fernández, "Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain" (2016) 10 International Journal of Communication 1167 at1171.

[226] Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930 at 938.

altogether content that challenges sexist or misogynistic views, credible journalism on the same topics, sources of fact-checking, or peer-reviewed academic research that has not been widely discredited.

As for other platform affordances, there are no shortage of ways in which they facilitate sexist, misogynistic, and sexually violent abuse and harassment, or are exploited and gamed by users to achieve the same objectives. “Easy feedback systems…lead to discursive loops, in which influencers build audiences that ask for, or reward, certain types of content. […] Such audiences can, in turn, drive political influencers to deliver ever more extreme content,”[227] resulting in self-reinforcing feedback loops of content increasingly bordering or crossing into hate speech. Coordinated campaigns that exploit simple platform engagement features include the following: mass upvoting or positively boosting extremist or hate-based content so that it appears on homepages, is promoted under trending topics, or is recommended to more users; mass downvoting feminist content or otherwise the content of marginalized individuals exposing or bringing attention to abuses so that their speech is effectively buried; falsely mass flagging or reporting feminist, anti-racist, or 2SLGBTQQIA users, posts, and pages for violating community standards in order to have their accounts or pages suspended, banned, or deleted;[228] or upvoting, liking, and engaging with harmful content posted without consent so the algorithms will pick up and assist in further disseminating the content, including intimate photos or personal information that resulted from doxing, such as the person’s home address.[229]

Users have also repurposed seemingly ‘neutral’ or ‘innocuous’ platform features such as Twitter’s hashtags and Facebook’s reaction emojis for abusive ends. For example, in *R v Elliott*, part of the harassment against one of the feminist activists was attributed to the defendant tweeting to a hashtag that had been specifically created for and tied to an event celebrating her, and which was a play on her Twitter handle (username). However, because a hashtag channel on Twitter is a public feature by

---

[227] Rebecca Lewis, “Alternative Influence: Broadcasting the Reactionary Right on YouTube” (2018) at 40, online (pdf): *Data & Society Research Institute* <https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf>.

[228] In one situation, a Facebook user, citing the Bible and believing he was “delivering God’s punishment to the ‘perverts and sodomites’”, reported hundreds of drag queens for violating the platform’s “real name” policy, resulting in mass suspensions. Gillespie points out the complexity in how to view content moderation features in such a case: “[T]he paradox here is that while @RealNamePolice’s motivations may have been political, and to some reprehensible, he did flag ‘appropriately’: he did understand the policy correctly, and he did identify names that violated it. Was this a misuse of the flagging system, then, or exactly what it was designed for?”: Tarleton Gillespie, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media* (New Haven: Yale University Press, 2018) at 94-95. Arguably, the responsibility ultimately lies with Facebook to ensure to the best of its ability that its policies cannot be interpreted and enacted towards abusive or oppressively discriminatory ends, regardless of whether the interpretation was technically correct or incorrect relative to what the platform intended. In this specific case, Facebook responded appropriately by changing the policy, which in its original formulation, constituted a form of harmful discrimination against certain historically marginalized groups of people (such as LGBTQ2S+ users, sex workers, survivors of intimate partner violence, and others who rely on maintaining pseudonymous identities for personal safety in accessing community and resources online).” See e.g., Lil Miss Hot Mess, "Facebook's 'real name' policy hurts real people and creates a new digital divide", *Guardian* (3 June 2015), online: <https://www.theguardian.com/commentisfree/2015/jun/03/facebook-real-name-policy-hurts-people-creates-new-digital-divide>.

[229] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Evidence*, 42nd Parl, 1st Sess, No 37 (5 December 2016) (Matthew Johnson); see also Kate Crawford & Tarleton Gillespie, “What is a flag for? Social media reporting tools and the vocabulary of complaint” (2016) 18:3 New Media & Society 410; and “In response, the Indigenous activist and writer Celeste Liddle posted the video in her Facebook page with a written message to denounce Facebook’s standards. What she did not imagine was that ‘malicious people’ would repeatedly flag her post until she was temporarily locked out of Facebook and the video removed”. Ariadna Matamoros-Fernández, “Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube” (2017) 20:6 Information, Communication & Society 930 at 931.

design, regardless of what the hashtag is or the context or purpose driving its creation, the judge determined that the defendant's usage of the hashtag attached to the activist's event "should not be treated as direct or indirect communication absent proof of intention to use them for that purpose".[230] In other contexts, trolls have deliberately flooded, with violent or abusive content, hashtags known to be channels of communications and resources for particular civil rights causes or social movements, such as when the hashtags #TakeBacktheTech and #ImagineAFeministInternet were inundated with "thousands of anti-feminist and misogynistic tweets and memes […] The volunteer who was organising the tweet chat also received an email in her personal inbox declaring the launch of the attack to 'destroy' the [anti-online violence] campaign."[231] Similarly, what may otherwise be considered "[i]nnocent and light-hearted" emojis are repurposed to perpetrate and pile onto online abuse, such as use of the eggplant emoji as a subtle form of sexual harassment in online spaces;[232] the vomit emoji being algorithmically linked to and entrenched as representative of 'liberals' and 'feminism';[233] and the use of the pig emoji and pig stickers as part of Islamophobic posts and rhetoric on social media, playing into a long history of "porcine hate acts" against Muslims throughout Western countries.[234]

Abusive users have also taken advantage of platforms' paid advertising features. For example, one Twitter account impersonating an Australian feminist paid for a promoted tweet (paid ads in the form of tweets) that *(Content Warning)* "encourag[ed] transgender people to kill themselves", which successfully ran across the site before Twitter responded to user reports and deleted the ads.[235] Twitter has also run paid advertisements promoting stalkerware, a form of consumer spyware closely tied to, and whose marketing explicitly encourages, intimate partner violence.[236] Researchers have shown that Facebook, Google, and Twitter all allow customers to buy targeted ads based on "a range of bigoted

---

[230] *R v Elliott*, 2016 ONCJ 35.

[231] "Take Action for #TakeBackTheTech and #ImagineAFeministInternet" (10 October 2015), online: *Association for Progressive Communications* <https://www.apc.org/en/pubs/take-action-takebackthetech-and-imagineafeministin>.

[232] "This symbolism of the eggplant emoji has become so commonplace that people not only use it for benevolent online sexual flirt but as an online harassment proxy (Dooling and Cuen, 2015). It is common for women to receive unsolicited eggplant emoji in different digital spaces, a practice that is at its core "a symbolic representation of old fashioned masculinity and dominance over women" (Dooling and Cuen, 2015)." Ariadna Matamoros-Fernández, "Inciting anger through Facebook reactions in Belgium: The use of emoji and related vernacular expressions in racist discourse" (2018) 23:9 First Monday, online: *First Monday* <https://firstmonday.org/ojs/index.php/fm/article/view/9405>.

[233] Ken Yeung, "Facebook fixing 'bug' showing vomit sticker when searching for 'liberals' or 'feminism'" (31 August 2016), online: *Venture Beat* <https://venturebeat.com/2016/08/31/facebook-fixing-bug-showing-vomit-sticker-when-searching-for-liberals-or-feminism/>.

[234] Ariadna Matamoros-Fernández, "Inciting anger through Facebook reactions in Belgium: The use of emoji and related vernacular expressions in racist discourse" (2018) 23:9 First Monday, online: *First Monday* <https://firstmonday.org/ojs/index.php/fm/article/view/9405>.

[235] Kia Kokalitcheva, "Troll Uses Twitter Ads to Spread Transphobic Message", *Time* (20 May 2015), online: <https://time.com/3891189/twitter-troll-transgende/>.

[236] "This week Twitter pushed sponsored tweets advertising a piece of spyware that is marketed to spy on a spouse. The advert heavily suggested the monitoring could be done without the subject's consent; it is illegal to use spyware in this way in the U.S. ... 'What is she hiding from you? Find our [sic] with mSpy!' the advert reads, according to a screenshot posted to Twitter. The ad is for 'mSpy Lite Phone Tracker App.' The advert then shows some notifications a customer might expect if they used the product. 'Helen entered the Night Club,' reads one. 'Helen left the office,' says another, as a man lays in bed reading the pop-ups." Joseph Cox, "Twitter Pushed Adverts for Spyware to Monitor Girlfriends" (3 July 2019), online: *Vice* <https://vice.com/en_uk/article/3k3wx5/twitter-pushed-adverts-for-spyware-to-track-girlfriends>.

and derogatory terms".[237] In at least one instance, the platform's own advertising capitalized on speech-based violence towards women, when Instagram used a screenshot of journalist Olivia Solon's post displaying an email threat she had received.[238]

To get around automated content detection systems (such as platform algorithms that automatically detect and remove photos containing nudity, or hate speech), those intent on disseminating content that they know may violate community standards "often modify their discourse online through deliberate misspellings or word choices".[239] Another tactic is to turn on the "sensitive media" filter on Twitter, generally intended for graphic or sexually explicit content but which is also used to conceal abusive or hate-based speech and reduce the chances of it being flagged or reported.[240] Users also rely on the so-called 'tweet and delete' strategy: tweeting abuse so that the target sees it, and deleting the tweet before it is reported or flagged.[241] While targeted or witnessing users have captured such abuses through screenshots, Twitter requires all reports of abuse to provide "links to exact Tweets or Twitter accounts" and is known for being "unable to accept attachments or screenshots"[242]—thus curtailing the ability to hold abusive users to account.

3.2. How Platform Dynamics Characterize TFGBV

Expression-based TFGBV on digital platforms has three characteristics in particular. These attributes, in part shaped by users' interactions with the platform's environment and affordances, amplify the power of TFGBV to intimidate, silence, and impose a coercive force on targeted individuals, with further repercussions for marginalized communities throughout society as a whole. First, speech-based abuse is weaponized by its perpetrators, going beyond mere expression and employed deliberately to attain

[237] Sam Levin, "Instagram uses 'I will rape you' post as Facebook ad in latest algorithm mishap", *Guardian* (21 September 2017), online: <https://www.theguardian.com/technology/2017/sep/21/instagram-death-threat-facebook-olivia-solon>; see e.g. Julia Angwin, Madeleine Varner & Ariana Tobin, "Facebook Enabled Advertisers to Reach 'Jew Haters'", *ProPublica* (14 September 2017), online: <https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters>; Brian Patrick Byrne, "Twitter Says It Fixed 'Bug' That Let Marketers Target People Who Use the N-Word" (16 September 2017), online: *Daily Beast* <https://www.thedailybeast.com/twitter-lets-you-target-millions-of-users-who-may-like-the-n-word>; Alex Kantrowitz, "Google Allowed Advertisers To Target People Searching Racist Phrases" (15 September 2017), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/alexkantrowitz/google-allowed-advertisers-to-target-jewish-parasite-black#.vsAxJvLbK>.

[238] (*Content Warning)* The email featured the subject line "Olivia, you fucking bitch!!!!!!!" and the statement, "I will rape you before I kill you, you filthy whore". Sam Levin, "Instagram uses 'I will rape you' post as Facebook ad in latest algorithm mishap", *Guardian* (21 September 2017), online: <https://www.theguardian.com/technology/2017/sep/21/instagram-death-threat-facebook-olivia-solon>.

[239] Anat Ben-David & Ariadna Matamoros-Fernández, "Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain" (2016) 10 International Journal of Communication 1167 at 1171. "Facebook's algorithm EdgeRank, which tracks what users like and the links they click on, recommends similar information based on the user's prior interests. Such algorithmic logic creates what Pariser (2011) describes as a 'filter bubble' to refer to the increasing personalization of the Web. One consequence of such algorithmic logic is that a user's racist behavior on Facebook triggers recommendations of similar content from the platform."

[240] Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930 at 938.

[241] J. Matias et al, "Reporting, Reviewing, and Responding to Harassment on Twitter" (13 May 2015) [unpublished], online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2602018>.

[242] "Staying safe on Twitter and sensitive content", online: *Twitter* <https://help.twitter.com/forms/abusiveuser>.

specific objectives with respect to their victims. Second, speech-based abuse on digital platforms is networked, socialized, and gamified, bringing together large groups of friends, acquaintances, collaborators, and strangers to enact online violence and harassment against targeted individuals or communities. Due to being networked, this form of abuse is also distributed, where perpetrators are able to widely distribute and dramatically lower the cost of engaging in abusive speech and behaviours, while victims of such abuse suffer the impact and costs of being the target of the entire network of abuse alone. Third, the ubiquity and ease of perpetrating speech-based abuse across online platforms, and its particular modes of expression (including reliance on humour and coded language to escape detection), both are perpetuated by and further result in normalization and mainstreaming, over time, of gender-based violence and sexist and misogynistic values and beliefs, seeping into the 'offline' world through conventional politics and physical attacks such as mass shootings. The remainder of this section will discuss each of these characteristics in turn.

3.2.1. Platformed TFGBV Weaponizes Expression to Harm Women

Although platformed TFGBV is often enacted through speech, it is not 'merely' speech, but constitutes substantive behaviour with immediate and long-term material impacts on women's lives, including their ability to exercise fundamental human rights and freedoms. Such expression, in the context of racism, has been described as "assaultive speech, […] words that are used as weapons to ambush, terrorize, wound, humiliate, and degrade".[243] The Supreme Court of Canada (SCC) characterized the purpose and impacts of hate speech as follows:

> Hate speech is, at its core, an effort to marginalize individuals based on their membership in a group. Using expression that exposes the group to hatred, hate speech seeks to delegitimize group members in the eyes of the majority, reducing their social standing and acceptance within society. When people are vilified as blameworthy or undeserving, it is easier to justify discriminatory treatment. The objective of s. 14(1)(b) [of the *Saskatchewan Human Rights Code*] may be understood as reducing the harmful effects and social costs of discrimination by tackling certain causes of discriminatory activity.[244]

J.L. Austin's speech act theory provides a salient lens through which to analyze sexist and misogynistic expression on digital platforms. The theory provides that words and speech do not simply impart their contents to recipients but, in their very utterance, "perform all kinds of actions"; put simply, "to say something is to *do* something".[245] According to Langton, speech acts subordinate a targeted group if they have three characteristics: "They rank [the targeted group] as having inferior worth. They

---

[243] Charles R Lawrence III et al, "Introduction" in Mari J Matsuda et al, eds, *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (New York & London: Routledge, 1993) 1 at 1, writing specifically about racist speech; however, this description applies to sexist or misogynistic speech as well, with respect to women, including racialized women. Mari Matsuda, a founding legal scholar of critical race theory, points out, "There is much speech that comes close to action. Conspiratorial speech, inciting speech, fraudulent speech, obscene speech, and defamatory speech are examples of words that seem to emerge from human mouths as more than ideas." Mari J Matsuda, "Public Response to Racist Speech: Considering the Victim's Story" in Mari J Matsuda et al, eds, *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (New York & London: Routledge, 1993) 17 at 32.

[244] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 71.

[245] Rae Langton, "Speech Acts and Unspeakable Acts" (1993) 22:4 Philosophy and Public Affairs 293 at 295 (emphasis in original).

legitimate discriminatory behavior on the part of [the dominant group]. And finally, they deprive [the targeted group] of some important powers: for example, the power to go to certain areas and the power to vote."[246] Speech act theory can enhance our understanding of abusive speech aimed towards women and intersecting marginalized identities on Internet platforms. In explaining how the Internet itself further destabilizes any lines between speech and action, Danielle Citron notes, "Indeed, the Internet's very essence is to aggregate expressions so as to convert them into actions."[247] In *Keegstra*, the SCC also recognized that expression can amount to behaviour: "In the context of sexual harassment, for example, this court has found that words can in themselves constitute harassment […]. In a similar manner, words and writings that wilfully promote hatred can constitute a serious attack on persons belonging to a racial or religious group".[248]

Researchers, legal scholars, governments, journalists, and the collective experiences of women and girls online globally have widely established that sexist and misogynistic statements online '*do* things' that, in their very utterances, impose a range of negative impacts on those who dare to engage online while being female. These impacts track the three characteristics of subordinating speech acts as set out by Langton.[249] First, such speech explicitly and implicitly demeans, dehumanizes, and treats women and their speech as inferior.[250] Second, it encourages or provides overt or tacit justification for discriminating against women and targeting them for further violence, abuse, and harassment.[251] As Matsuda writes, "The deadly violence that accompanies the persistent verbal degradation of those subordinated because of gender or sexuality explodes the notion that there are clear lines between words and deeds."[252] Third, abusive speech silences women, intimidates them into self-censorship, and drives them off of central forums of public discussion if not off the Internet altogether, depriving women

---

[246] *Ibid* at 303.

[247] Danielle Keats Citron, "Cyber Civil Rights" (2009) 89 Boston University Law Review 61 at 99.

[248] *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 64 (WL).

[249] Langton herself explicitly draws the connection between speech act subordination and silencing of an oppressed group: "To subordinate is to rank, to legitimate discrimination, to unfairly deprive of a power; to silence is to deprive of a power. So there is a link between the subordination claim and the silencing claim: one way of subordinating is to silence, to deprive someone of certain liberties that are available to others—the opportunity, for example, freely to speak." Rae Langton, "Speech Acts and Unspeakable Acts" (1993) 22:4 Philosophy and Public Affairs 293 at 329.

[250] See e.g., "Online hate 'undermines the well-being and sense of security of victims' as well as their 'sense of belonging.' More generally, it increases discord in society and contributes to the marginalization of certain groups 'by convincing listeners of the inferiority of the targeted group.'" Canada, Parliament, House of Commons, *Taking Action to End Online Hate: Report of the Standing Committee on Justice and Human Rights*, 42nd Parl, 1st Sess (June 2019) (Chair: Anthony Housefather) at 8 (footnotes omitted); Jocelyn Maclure, "The Regulation of Hateful and Hurtful Speech: Liberalism's Uncomfortable Predicament" (2017) 63 McGill Law Journal 133 at 141-42 (footnotes omitted); and Richard Delgado, "Words that Wound: A Tort Action for Racial Insults, Epithets, and Name Calling" in Mari J Matsuda et al, eds, *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (New York & London: Routledge, 1993) 89.

[251] See e.g., Richard Moon, "A turning point for misogynist and Islamophobic speech?" (19 February 2019), online: *Policy Options* <https://policyoptions.irpp.org/magazines/february-2019/turning-point-misogynist-islamophobic-speech/>. See also Section 3.2.3 ("Platformed TFGBV Normalizes and Escalates Violence against Women").

[252] Mari J Matsuda, "Public Response to Racist Speech: Considering the Victim's Story" in Mari J Matsuda et al, eds, *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (New York & London: Routledge, 1993) 17 at 23. Matsuda further notes, "In considering the emerging theory that patriarchy and heterosexism are cornerstones of violence in our society, I note that in researching hundreds of incidents of racist violence in preparation for this chapter, I found in virtually every case the perpetrators were men. Thus although the focus of this chapter is racist speech, other forms of subordination are always, uneasily close at hand."

of important freedoms such as the ability to exercise freedom of expression online,[253] and associated rights and liberties such as enjoying the benefits and necessities of the Internet without trepidation, and participating fully in the world as political actors.[254]

3.2.2. Platformed TFGBV Is Networked, Socially Gamified, and Distributed

Online abuse can involve large-scale coordination across multiple platforms, such that assaults on targeted individuals or groups are not necessarily the result of single attacks that occur independently of each other, but instead, deliberately orchestrated onslaughts that take advantage of the networked properties of digital platforms.[255] In *The Internet of Garbage*, Sarah Jeong explains how "sustained harassment campaigns […] are often coordinated out of another online space"; for example, the planning and coordination may take place on a forum such as 4chan or within a Facebook group, while the attack itself is carried out on Twitter or another platform that is home to the targeted individual or group.[256] Jeong further notes, "When one platform links to another platform in these cases, it creates a pipeline of hate with very little friction. Even if the targeted platform maintains certain norms, the oncoming invaders ignore them, operating only under the norms of their originating platform."[257]

This networked and coordinated nature of online abuse also creates and is sustained by what may be considered a troubling 'gamification' of abuse, where it amounts to a social activity and source of bonding, entertainment, and competition between those involved. According to Anita Sarkeesian, "We

---

[253] See e.g., Jenny Sundén & Susanna Paasonen, "Shameless hags and tolerance whores: feminist resistance and the affective circuits of online hate" (2018) 18:4 Feminist Media Studies 643 at 646 (in-line citation omitted); and "Silencing is what many harassers are after. As a [TFGBV] victim confided to me, she felt she was left with no choice but to withdraw from online life because whenever she engaged online, harassers went after her, and whenever she stopped, so did they." Danielle Keats Citron, "Restricting Speech to Protect It" in Susan J Brison & Katharine Gelber, eds, *Free Speech in the Digital Age* (New York: Oxford University Press, 2019) 122 at 131 (inline citations omitted).

[254] See e.g., "Cyber harassment destroys victims' ability to interact in ways that are essential to self-governance. Online abuse prevents targeted individuals from realizing their full potential as digital citizens. Victims cannot participate in online networks if they are under assault. Rape threats, defamatory lies, the non-consensual disclosure of nude photos, and technological attacks destroy victims' ability to interact with others. They sever a victim's connections with people engaged in similar pursuits." Danielle Keats Citron, "Restricting Speech to Protect It" in Susan J. Brison & Katharine Gelber, eds, *Free Speech in the Digital Age* (New York: Oxford University Press, 2019) 122 at 130; Danielle Keats Citron, *Hate Crimes in Cyberspace* (Cambridge & London: Harvard University Press, 2014) at 126-27 (footnotes omitted)"; and LW Sumner, "Incitement and the Regulation of Hate Speech in Canada: A Philosophical Analysis" in Ivan Hare & James Weinstein, eds, Extreme Speech and Democracy (Oxford: Oxford University Press, 2009) 205 at 208.

[255] "Coordinated harassment campaigns are increasingly organized by more-and-less organized groups, who synchronously flood a target's social media feeds (Heron, Belford & Goker, 2014; Phillips, 2015)." R Stuart Geiger, "Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space" (2016) 19:6 Information, Communication & Society 787 at 787; Alice E. Marwick & Robyn Caplan, "Drinking male tears: language, the manosphere, and networked harassment" (2018) 18:4 Feminist Media Studies 543 at 544.

[256] See e.g., "Such networked misogyny is often organized in subcultural online spaces such as Reddit, 4Chan, and chat rooms, where participants collectively frame feminists like Sarkeesian as 'villains.' This provides justification for the harassing behavior and gives those engaging in it a moral high ground (Shagun Jhaver, Larry Chan, and Amy Bruckman 2018)." Alice E. Marwick & Robyn Caplan, "Drinking male tears: language, the manosphere, and networked harassment" (2018) 18:4 Feminist Media Studies 543 at 545; and "In the case of the harassment of Zoë Quinn, Quinn documented extensive coordination from IRC chat rooms, replete with participation from her ex-boyfriend. Theoretically, sustained harassment can take place entirely on a single platform without having to receive reinforcement from an outside platform, but I have come across no such instances." Sarah Jeong, *The Internet of Garbage* (Vox Media, 2018) at 1346-61.

[257] *Ibid* at 1346-61.

don't usually think of online harassment as a social activity, but we do know from the strategies and tactics that they used that they were not working alone, that they were actually loosely coordinating with one another. The social component is a powerful motivating factor that works to provide incentives for perpetrators to participate and to actually escalate the attacks by earning the praise and approval of their peers."[258] Taking the gamification element further, "[u]sers on 8chan frequently lionize mass gunmen using jokey internet vernacular, referring to their body counts as 'high scores' and creating memes praising the killers."[259] In *R v BLA*, the court noted this precise dynamic as part of the defendant's motivations, as summarized by the eQuality Project:

> [The accused] posted a false ad on Craigslist pretending to be another girl stating that she was looking for sex, along with her name and address. He claimed to have nude photos of another girl and threatened to post them online. He used bots to send over 200 texts to one girl. B's pre-sentencing report makes note of his misogynistic attitudes and finds that his actions were primarily motivated by pleasure from his victim's distress and prestige gained within an online peer group.[260]

Online platforms do not only support such networks, but may also create them where they would not have existed otherwise. This primarily occurs through recommendation features which suggest, promote, and push content at users that they may not have seen otherwise, related to content they have engaged with or otherwise indicated interest in. Kaiser and Rauchfleisch have demonstrated how this occurs in the particular case of YouTube inexorably, once its algorithms were established, forging a network of right-wing channels that would be delivered to its target audience on a platter:

> YouTube's algorithms are not creating something that is not already there. These channels exist, they interact, their users overlap to a certain degree. YouTube's algorithm, however, connects them visibly via recommendations. […] [Algorithms] potentially shape future [users'] behaviour. As our data shows, the channel recommendation connects diverse channels that might be more isolated without the influence of the algorithm, and thus helps to unite the right.[261]

The networked nature of online platforms and platformed TFGBV also allows perpetrators to easily piece together, infiltrate, or otherwise poison victims' own online networks across the same and other platforms. As Jeong writes, "A simple Google search can connect together all the disparate aspects of a person's digital life, allowing bad actors to attack each and every part even without knowing them particularly well to begin with."[262] Many cases of online abuse in the context of intimate partner violence

---

[258] Alice E. Marwick & Robyn Caplan, "Drinking male tears: language, the manosphere, and networked harassment" (2018) 18:4 Feminist Media Studies 543 at 545.

[259] Kevin Roose, "'Shut the Site Down,' Says the Creator of 8chan, a Megaphone for Gunmen", *The New York Times* (4 August 2019), online: <https://www.nytimes.com/2019/08/04/technology/8chan-shooting-manifesto.html>.

[260] "Technology-Facilitated Violence: Criminal Harassment Case Law" (3 July 2020), online (pdf): *eQuality Project* <http://www.equalityproject.ca/wp-content/uploads/2020/07/TFVAW-Criminal-Harassment-3-July-2020.pdf> at 31-32 summarizing *R v BLA*, 2015 BCPC 203.

[261] Jonas Kaiser & Adrian Rauchfleisch, "Unite the Right? How YouTube's Recommendation Algorithm Connects the U.S. Far Right" (11 April 2018), online: *D&S Media Manipulation: Dispatches from the Field* <https://medium.com/@MediaManipulation/unite-the-right-how-youtubes-recommendation-algorithm-connects-the-u-s-far-right-9f1387ccfabd>.

[262] Sarah Jeong, *The Internet of Garbage* (Vox Media, 2018).

involve abusers "recruit[ing] other people to participate in TFCC [technology-facilitated coercive control]"; "us[ing] friends' and family members' devices and accounts to contact survivors"; and enlisting their own friends and family as well as those of the targeted individual to pass on harassing communications or pressure the victim/survivor to re-establish contact or reconcile with the abuser, as a method to get around the abuser themselves being blocked by the victim across platforms.[263]

Because online abuse can involve so many actors spread throughout large networks, such abuse and the costs of perpetrating it are *distributed*: each participant contributes only one or a few messages or clicks, and spends a nominal amount of time and energy to harass the targeted individual. However, the total impact of such a network's collective volume of abuse landing on a single individual can be profound and devastating. Geiger describes the "disparities of scale" inherent to much online abuse, in this case within the context of Twitter:

> This capacity for collective action in counter-harassment work is important given the disparities of scale that are associated with online harassment. As many scholars note, a particularly problematic form of harassment takes the form of 'piling on,' where a large number of people each send a small number of messages, overwhelming the target [...]. The work of harassment can be efficiently distributed and decentralized, with anonymous imageboards serving as one of many key sites for the selection of targets. Some prominent anti-feminist individuals also use Twitter itself to direct their tens of thousands of followers to particular accounts. In such a situation, it only takes a short amount of time and energy to send a single harassing reply. In contrast, the work of responding to harassment is much more difficult to scale, as each of those messages must be dealt with by the recipient.[264]

The networked, socially gamified, and distributed nature of speech-based (and other) abuse on digital platforms must inform proposed solutions to mitigate or prevent such abuse. Significant harm to targeted individuals or groups results in part from the coalescence of mass collective action, online social dynamics, and enabling technical infrastructure, posing a challenge to the ability to identify or hold legally accountable any one actor within the network of abuse.

3.2.3. Platformed TFGBV Normalizes and Escalates Violence against Women

The proliferation of misogynistic expression across online platforms fulfills an important function for those perpetuating it, which is to lend a normalizing and eventually legitimizing validity to such values

---

[263] Molly Dragiewicz et al, *Domestic violence and communication technology: Survivor experiences of intrusion, surveillance, and identity crime* (Sydney: Australian Communications Consumer Action Network, 2019) at 24.

[264] R Stuart Geiger, "Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space" (2016) 19:6 Information, Communication & Society 787 at 795 (in-text citations omitted). See also: "Moreover, the Internet's powerful aggregative capacity converts seemingly individual expressions (e.g., visiting a website or sending an e-mail) into criminal acts through their repetition (e.g., denial-of-service attacks and image reaping). The Internet also routinely allows individuals to aggregate their efforts with strangers." Danielle Keats Citron, "Cyber Civil Rights" (2009) 89 Boston University Law Review 61 at 99-100.

and beliefs in everyday discussion[265] and in public and political discourse, both online and offline.[266] As the SCC stated in *Whatcott*, hate speech "can have a societal impact. If a group of people are considered inferior, subhuman, or lawless, it is easier to justify denying the group and its members equal rights or status. […] 'hate speech always denies fundamental rights'."[267]

Research has repeatedly tied virulent and sustained strains of misogynistic belief systems and discourse to online communities across various social networking platforms,[268] through the loosely yet consistently connected web of sociocultural and political online commentators, content creators, conservative activists, self-styled intellectuals, and average users with key overlapping beliefs, known as the 'manosphere'.[269] These various communities promote and popularize—both as a matter of course in their day-to-day online activities and deliberately and systematically through more coordinated, short-term and long-term strategies[270]—misogynistic ideologies in their own right,[271] or as part of a broader constellation of overlapping far-right positions, including white supremacy, homophobia and transphobia, racism, nationalism, fascism, Islamophobia, and anti-Semitism.[272]

---

[265] "These spaces, in which extremist ideas meet banal talk of video games and other facets of internet culture, also serve to dehumanise communities and negate the impact of actions against them." Mike Stuchbery, "The New Zealand terror attack shows how far-right violence is cultivated by the internet and populist politicians", *Independent* (15 March 2019), online: <https://www.independent.co.uk/voices/new-zealand-shooting-terror-christchurch-mosque-far-right-suspect-a8824186.html>.

[266] See e.g., Canada, Parliament, House of Commons, *Taking Action to End Online Hate: Report of the Standing Committee on Justice and Human Rights*, 42nd Parl, 1st Sess (June 2019) (Chair: Anthony Housefather) at 8; Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 39-40.

[267] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 74, citing *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 147 (WL).

[268] See e.g., Aaron Winter, "'Online Hate: From the Far-Right to the 'Alt-Right', and from the Margins to the Mainstream'" in Karen Lumsden & Emily Harmer, eds, *Online Othering: Exploring Digital Violence and Discrimination on the Web* (London: Palgrave Macmillan, 2019) 39 at 50 (inline citations omitted).

[269] See Debbie Ging, "Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere" (2019) 22:4 Men and Masculinities 638; and "The internet has been key to the popularization of men's rights activism and discourse [...]. While the manosphere includes a variety of groups, including MRAs, pickup artists, MGOW (men going their own way), incels (involuntary celibates), father's rights activists, and so forth, they share a central belief that feminine values dominate society, that this fact is suppressed by feminists and "political correctness," and that men must fight back against an overreaching, misandrist culture to protect their very existence": Alice E Marwick & Robyn Caplan, "Drinking male tears: language, the manosphere, and networked harassment" (2018) 18:4 Feminist Media Studies 543 at 546 (inline citations omitted).

[270] See explanation of 'information laundering' in Anat Ben-David & Ariadna Matamoros-Fernández, "Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain" (2016) 10 International Journal of Communication 1167 at 1170-71, 1187; and Jacob Davey & Julia Ebner, "The Fringe Insurgency: Connectivity, Convergence and Mainstreaming of the Extreme Right" (2017), online (pdf): *ISD* <www.isdglobal.org/wp-content/uploads/2017/10/The-Fringe-Insurgency-221017.pdf> at 14-16, 25.

[271] *Ibid* at 27-28.

[272] Rebecca Lewis, "Alternative Influence: Broadcasting the Reactionary Right on YouTube" (2018) at 35, online (pdf): *Data & Society Research Institute*, <https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf>; and "GamerGate laid the groundwork for what was to come. A significant part of what made it such a militant hate movement was the fact that it was leapt upon by right-wing media—most notoriously, former Trump advisor Steve Bannon's website Breitbart—as an opportunity to convert these angry people to their cause and expose them to more explicitly right-wing ideas." Andrew Todd, "NZ Authorities Have Been Ignoring Online Right-Wing Radicalisation For Years" (21 March 2019),

Sexist and misogynistic ideas, allusions, and dog whistles that saturate online environments do not stay there, but rather go on to enter mainstream political discourse, achieving even greater validity, institutional legitimacy, and political power.[273] This has happened in Canada[274] and around the world.[275] For example, in March 2017, during a campaign running for leader of the federal Conservative Party, former cabinet minister Maxime Bernier posted a meme on Twitter making reference to "taking the red pill"[276]—a phrase that originated with the film *The Matrix* but by that point had become (and continues to be) widely synonymous with a worldview that "purports to awaken men to [what is misogynistically considered] feminism's misandry and brainwashing" which has purportedly "elevated women to the position of dominance over men. The red pill has an anti-feminist political philosophy built in that rejects the equality of women and the post-1960s politics of women's empowerment."[277] One particular dynamic that makes it difficult to hold politicians to account for employing and tacitly legitimizing such language is the constant availability of plausible deniability, due to such communities' ongoing cultivation of "a culture layered in sarcasm and satire; this veil is challenging for a dilettante to

online: *Vice* <https://www.vice.com/en_nz/article/pan9yg/nz-authorities-have-been-ignoring-online-right-wing-radicalisation-for-years>.

[273] See e.g., "[Those] who share a similar aversion can feel validated and encouraged to express their sentiments publicly. When hateful or contemptuous speech erupts in the public sphere, it lowers the social cost of expressing negative attitudes toward the targeted group, especially when it is expressed by people in positions of authority or influence. The diffusion and circulation of hate speech favour the coordination of actions among those who share a common aversion and can, as a consequence, increase the vulnerability of the targeted groups." Jocelyn Maclure, "The Regulation of Hateful and Hurtful Speech: Liberalism's Uncomfortable Predicament" (2017) 63 McGill Law Journal 133 at 142; and Zachary Kamel, Martin Patriquin & Alheli Picazo, "Maxime Bernier's alt-right problem", *Toronto Star* (15 February 2019), online: <https://thestar.com/politics/federal/2019/02/08/maxime-berniers-alt-right-problem.html>.

[274] "Increasingly, we are seeing evidence that the far right has already had success in reshaping the boundaries of acceptable political discourse in Canada. A number of different groups have latched onto the ideas of the far right, blending them into their political agendas and movements. For example, the United We Roll protest movement, while ostensibly focused on criticizing the federal government's alleged disregard for Alberta's oil economy, has also featured critical rhetoric of illegal immigration and globalism. Canada's newest federal political party is also rooting its appeal to Canadians in the language championed by the far right. The People's Party has constructed the core of its policy agenda around a commitment to reducing immigration, protecting borders and preserving Euro-Canadian heritage." Brian Budd, "Starving online trolls won't stop far-right ideas from going mainstream" (14 April 2019), online: *Conversation* <https://theconversation.com/starving-online-trolls-wont-stop-far-right-ideas-from-going-mainstream-115220>.

[275] See generally Aaron Winter, "'Online Hate: From the Far-Right to the 'Alt-Right', and from the Margins to the Mainstream'" in Karen Lumsden & Emily Harmer, eds, *Online Othering: Exploring Digital Violence and Discrimination on the Web* (London: Palgrave Macmillan, 2019) 39; Ariadna Matamoros-Fernández, "Inciting anger through Facebook reactions in Belgium: The use of emoji and related vernacular expressions in racist discourse" (2018) 23:9 First Monday, online: *First Monday* <https://firstmonday.org/ojs/index.php/fm/article/view/9405>; and Aaron Winter & Aurelien Mondon, "Understanding the mainstreaming of the far right" (26 August 2018), online: *openDemocracy* <https://www.opendemocracy.net/en/can-europe-make-it/understanding-mainstreaming-of-far-right/>.

[276] Ryan Maloney, "Maxime Bernier Criticized For Using 'Red Pill' Meme Popular Among Anti-Feminists" (7 March 2017), online: *Huffington Post* <https://www.huffingtonpost.ca/2017/03/07/maxime-bernier-matrix-red-pill-_n_15205762.html>; David Bell, "Red pill rhetoric infiltrating political parties, Calgary prof cautions", *CBC News* (8 March 2017), online: <https://www.cbc.ca/news/canada/calgary/mra-political-parties-1.4016680>.

[277] Bharath Ganesh, "What the Red Pill Means for Radicals" (7 June 2018), online: *Fair Observer* <https://www.fairobserver.com/world-news/incels-alt-right-manosphere-extremism-radicalism-news-51421/>, citing Debbie Ging, "Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere" (2019) 22:4 Men and Masculinities 638. The article continues, situating the concept of the "red pill" in the context of the broader extremist right movement, "Taking the red pill, for both the manosphere and the alt-right, is the beginning of a process of radicalization in which an individual becomes enculturated in an extreme, reactionary worldview."

penetrate",[278] in a context where "defences of satire and irony to disguise racist and sexist commentary are a common practice online [...] that fosters discrimination and harm".[279]

Consistently repeated and amplified harmful or hateful speech online does not only normalize harm to women and intersecting marginalized communities, but in fact serves to justify such harm in the minds of the converted. Moreover, such discourse can make violence appear to be a logical and necessary response for adherents of misogynistic or related oppressive beliefs, such as anti-immigration ideology.[280] Free expression scholar Richard Moon asserts, in the context of racism and hate speech laws, that the "concern is that individuals, or small groups, who are already inclined to bigoted or racist thinking, might be encouraged or emboldened to take extreme action against the target group's members."[281] In the context of platformed misogyny, the following passage illustrates how violence against women is advocated as both justified and necessary, among the so-called "incel"[282] community:

---

[278] Jacob Ware, Bruce Hoffman & Ezra Shapiro, "Remembering Toronto: Two Years Later, Incel Terrorism Threat Lingers" (6 May 2020), online: *Global Network on Extremism & Technology* <https://gnet-research.org/2020/05/06/remembering-toronto-two-years-later-incel-terrorism-threat-lingers/>.

[279] Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930 at 936 (inline citations omitted); see also Roose's description of the kind of language characteristic to such communities, in this case, illustrated in a mass shooting manifesto comprising "a wordy mixture of white nationalist boilerplate, fascist declarations and references to obscure internet jokes — [what] seems to have been written from the bottom of an algorithmic rabbit hole." Kevin Roose, "A Mass Murder of, and for, the Internet", *The New York Times* (15 March 2019), online: <https://www.nytimes.com/2019/03/15/technology/facebook-youtube-christchurch-shooting.html>.

[280] As noted by Bradley Galloway from the Organization for the Prevention of Violence, "[t]he perpetuation of associated rhetoric can create an environment where discrimination, harassment and violence are viewed by individuals as not only a reasonable response or reaction but also as a necessary one." Canada, Parliament, House of Commons, *Taking Action to End Online Hate: Report of the Standing Committee on Justice and Human Rights*, 42nd Parl, 1st Sess (June 2019) (Chair: Anthony Housefather) at 8 (footnotes omitted); see also Canada, Parliament, House of Commons, *Taking Action to End Online Hate: Report of the Standing Committee on Justice and Human Rights*, 42nd Parl, 1st Sess (June 2019) (Chair: Anthony Housefather) at 9.

[281] Richard Moon, "A turning point for misogynist and Islamophobic speech?" (19 February 2019), online: *Policy Options* <https://policyoptions.irpp.org/magazines/february-2019/turning-point-misogynist-islamophobic-speech/>; see also Mike Stuchbery, "The New Zealand terror attack shows how far-right violence is cultivated by the internet and populist politicians", *Independent* (15 March 2019), online: <https://www.independent.co.uk/voices/new-zealand-shooting-terror-christchurch-mosque-far-right-suspect-a8824186.html> ("A recent Hope Not Hate study showed that five of the top ten most influential far right social media personalities in the world today hail from these shores. It was inevitable that inflammatory pronouncements about Islam's threat, made over and over again by myriad people around the world would eventually be answered with violence.")

[282] ""Incel" signifies an "involuntary celibate" — [someone who considers himself to be] a male oppressed by the injustice of women who refuse to have sex with him. The term came into prominence in the corners of the "manosphere," a loose coalition of men's rights activists, bloggers, participants in pick-up artist forums in addition to audiences across social media platforms, primarily Reddit and 4chan. The incel identity is violently misogynistic." Bharath Ganesh, "What the Red Pill Means for Radicals" (7 June 2018), online: *Fair Observer* <https://www.fairobserver.com/world-news/incels-alt-right-manosphere-extremism-radicalism-news-51421/>; see also Jia Tolentino, "The Rage of the Incels", *The New Yorker* (15 May 2018), online: <https://newyorker.com/culture/cultural-comment/the-rage-of-the-incels> ("In the past few years, a subset of straight men calling themselves 'incels' have constructed a violent political ideology around the injustice of young, beautiful women refusing to have sex with them. These men often subscribe to notions of white supremacy […] They're also diabolically misogynistic. *[…]* The idea that this misogyny is the real root of their failures with women does not appear to have occurred to them. […] Incels aren't really looking for sex; they're looking for absolute male supremacy. Sex, defined to them as dominion over female bodies, is just their preferred sort of proof.").

> On one incel website, the Toronto van attack suspect is called "our hero," while the gunman who killed 14 women at a Montreal engineering school is a "prophet for the incel cause." "There's only one path to acknowledgement as an incel and it happens to be violence," a user wrote on a different online forum. "In order for your ideology to get across, violence is inevitably required," another wrote. Internet sites for incels, or involuntary celibates, are a swamp of self-pity, conspiracy theory and outright justification of violence. But despite growing recognition that attacks by incels are a form of domestic terrorism, online discussion forums that cater to the misogynist subculture continue to operate openly.[283]

Sustained hateful speech online can escalate and has escalated into physical and lethal violence offline, exactly as called for by those represented in the passage above.[284] This is perhaps manifested most prominently in the form of misogyny- or racism-fuelled mass shootings that have been characterized as "of, and for, the Internet […] an internet-native mass shooting, conceived and produced entirely within the irony-soaked discourse of modern extremism."[285]

Canadian examples include the January 2017 Islamophobic shooting at a mosque in Quebec City,[286] and the April 2018 misogynistic murder of ten people in Toronto, through the killer deliberately plowing a van into a busy sidewalk shortly after posting a foreshadowing Facebook message. Despite a court finding that advancing the 'incel movement' was not the perpetrator's primary motive, so much as a means by which to achieve notoriety, it bears interrogating whether it may not have been precisely the "of, and for, the Internet" nature of what is essentially a form of politicized and mobilized TFGBV[287] that made the perpetrator select this movement, out of all possible ones, to champion, however

---

[283] Stewart Bell, "Despite crackdown on incels, their discussion forums are still online" *Global News* (9 June 2020), online: <https://globalnews.ca/news/7022100/incel-discussion-forums-still-online-crackdown/>.

[284] See e.g., Mitchell Gracie, "The Rise of the Alt-Right in Canada", *Ontarion* (15 November 2018), online: <https://theontarion.com/2018/11/15/the-rise-of-the-alt-right-in-canada/>; Aaron Winter, "'Online Hate: From the Far-Right to the 'Alt-Right', and from the Margins to the Mainstream'" in Karen Lumsden & Emily Harmer, eds, *Online Othering: Exploring Digital Violence and Discrimination on the Web* (London: Palgrave Macmillan, 2019) 39 at 54-56; and Maura Conway, "Violent Extremism and Terrorism Online in 2018: The Year in Review" (2018) at 20-21, online (pdf): *Vox Pol* <https://www.voxpol.eu/download/vox-pol_publication/Year-in-Review-2018.pdf> .

[285] Kevin Roose, "A Mass Murder of, and for, the Internet", *The New York Times* (15 March 2019), online: <https://www.nytimes.com/2019/03/15/technology/facebook-youtube-christchurch-shooting.html>. See also Georgia Wells & Ian Lovett, "'So What's His Kill Count?': The Toxic Online World Where Mass Shooters Thrive", *The Wall Street Journal* (4 September 2019), online: <https://www.wsj.com/articles/inside-the-toxic-online-world-where-mass-shooters-thrive-11567608631>.

[286] [C]ourt evidence and uncovered online activity have revealed that the Internet can be powerful in motivating far-right extremists to commit acts of violence. For example, after Bissonnette's Twitter searches were introduced to the court by Crown attorneys, the Washington Post reported: "Bissonnette also appears to have obsessively visited the Twitter accounts of Tucker Carlson and Laura Ingraham, Fox News personalities; David Duke, the former leader of the Ku Klux Klan; Alex Jones of Infowars; conspiracy theorist Mike Cernovich; Richard Spencer, the white nationalist; and senior White House adviser Kellyanne Conway. Bissonnette checked in on the Twitter account of Ben Shapiro, editor-in-chief of the conservative news site the Daily Wire, 93 times in the month leading up to the shooting." Mitchell Gracie, "The Rise of the Alt-Right in Canada", *The Ontarion* (15 November 2018), online: <https://theontarion.com/2018/11/15/the-rise-of-the-alt-right-in-canada/>.

[287] [Incels] use online forums to spread their messages of hate, convincing other would-be incels they can blame their social and sexual difficulties on others. Some fantasise about committing acts of violence." Sian Tomkinson, Katie Attwell & Tauel Harper, "'Incel' violence is a form of extremism. It's time we treated it as a security threat" (26 May 2020), online: *Conversation* <https://theconversation.com/incel-violence-is-a-form-of-extremism-its-time-we-treated-it-as-a-security-threat-138536>.

'incidentally'.[288] This would be in addition to his pre-existing interest in incels and misogynistic resentment towards women, as further stoked by the incel forums themselves even if not to the point of becoming a *primary* motive.

In addition, the judge in *R v Sears* referenced the Toronto attack in Canada's first decision that applied the criminal hate speech provision to women as an identified targeted group:

> The preeminent concern noted a half century ago, that hate propaganda could contribute to violence, is starkly relevant today. The Toronto van attack in April 2018, the Quebec mosque attack in January 2017, and the Pittsburgh synagogue attack a few months ago, are all present day displays of extreme hatred of identifiable groups. The extent to which hate propaganda, such as that which YWN publishes, bears responsibility in these cases, is still undergoing investigation.[289]

Similar attacks, with clear Internet-oriented elements planned and built in, have occurred in other jurisdictions, including once in New Zealand and more frequently in the United States:

> Moments before the El Paso shooting on Saturday, a four-page message whose author identified himself as the gunman appeared on 8chan. The person who posted the message encouraged his "brothers" on the site to spread the contents far and wide. In recent months, 8chan has become a go-to resource for violent extremists. At least three mass shootings this year—including the mosque killings in Christchurch, New Zealand, and the synagogue shooting in Poway, Calif.—have been announced in advance on the site, often accompanied by racist writings that seem engineered to go viral on the internet.[290]

---

[288] *R v Minassian*, 2021 ONSC 1258. The perpetrator of the Toronto attack was subsequently found guilty of ten counts of first-degree murder and 16 counts of attempted murder. The court accepted the evidence of expert witnesses that the "incel movement" was "not a primary driving force behind the attack", though additionally stated that "resentment towards women [purportedly as a result of women not having demonstrated attraction to him] was a factor in this attack" (at paras 193 and 196). It should also be noted that while the defendant admitted having exaggerated his investment in and connections to the 'incel movement' in his police statement, in the same interviews used as part of the expert evidence to make the court's findings, he "continued to say that he was influenced by Elliot Rodger's manifesto and by the incel movement" (at para 157). Several expert witnesses told the court the defendant was "'hyper-focused' on", "indoctrinated" by, and "obsessed with" Elliot Rodger, his manifesto, and mass shootings (at paras 158, 165, and 188). Moreover, the defendant told expert witnesses in these same interviews that "he would have preferred to hit more women", "his preference was to kill women", and "if he had the opportunity to do this all over again, he would be more specific in targeting women between the ages of 18 and 30" (at paras 177, 179, 191). The defendant also "reported that he enjoyed reading negative, hate-filled comments about women and felt relieved that there was an explanation for why women never seemed interested in him" (at para 170).

[289] *R v Sears*, 2019 ONCJ 104 at para 13.

[290] Kevin Roose, "'Shut the Site Down,' Says the Creator of 8chan, a Megaphone for Gunmen", *The New York Times* (4 August 2019), online: <https://www.nytimes.com/2019/08/04/technology/8chan-shooting-manifesto.html>. See also "We also know that many recent acts of offline violence bear the internet's imprint. Robert Bowers, the man charged with killing 11 people and wounding six others at the Tree of Life synagogue in Pittsburgh, was a frequent user of Gab, a social media platform beloved by extremists. Cesar Sayoc, the man charged with sending explosives to prominent critics of President Trump last year, was immersed in a cesspool of right-wing Facebook and Twitter memes." Kevin Roose, "A Mass Murder of, and for, the Internet", *The New York Times* (15 March 2019), online: <https://www.nytimes.com/2019/03/15/technology/facebook-youtube-christchurch-shooting.html>; and "These social media tools were used for the express purpose of turning the killings into a spectacle, one to be consumed over and over again." Mike Stuchbery, "The New Zealand terror attack shows how far-right violence is cultivated by the internet and populist politicians", *Independent* (15 March 2019), online:

Again, these actions emerged from repeated hateful and misogynistic or racist and Islamophobic discourse—'just' speech, but a particular type of speech taken to its logical violent endpoint.

In some cases, online movements hostile to women have grown into movements now considered high-level threats on par with other national security concerns, such as in the case of incel-driven killings.[291] Specifically, in May 2020, the RCMP laid terrorism charges against a 17-year-old male defendant for fatally stabbing a female employee, Ashley Arzaga, and injuring two others at a massage parlour in Toronto, based on evidence that the attacker was motivated by incel ideology and had carried out his attack in its name.[292] This is the first time in Canadian history that criminal violence not linked to racialized and religious minority groups such as al-Qaeda or ISIS has been formally charged as terrorist activity,[293] let alone criminal violence driven by misogyny specifically.[294] In addition, both the RCMP and the Canadian Security Intelligence Service (CSIS) now recognize incel ideology, and violent misogyny, as a form of ideological extremism that can fuel terrorist acts: the RCMP by adding incels to its Terrorism and Violent Extremism Awareness Guide,[295] and CSIS by including incels as an example of gender-driven violence under the category Ideologically Motivated Violent Extremism (IMVE) in its 2019 annual public report.[296] The escalation and flourishing of misogynistic rhetoric and ideology and its subsequent violence, alongside both overt and subtle consequences for women in everyday life, cannot be separated from this platformed TFGBV's ability to thrive, proliferate, and recruit adherents online.

3.3. Platform Content Moderation Policies and Practices

Digital platforms have implemented a range of content moderation features and processes to address content that may constitute TFGBV, as well as content that may not constitute TFGBV but may violate terms of use or community standards for other reasons (such as intellectual property infringement).

<https://www.independent.co.uk/voices/new-zealand-shooting-terror-christchurch-mosque-far-right-suspect-a8824186.html>.

[291] "The incel ideology has already inspired the murders of at least sixteen people. Elliot Rodger, in 2014, in Isla Vista, California, killed six and injured fourteen in an attempt to instigate a 'War on Women' for 'depriving me of sex.' (He then killed himself.) Alek Minassian killed ten people and injured sixteen, in Toronto, last month [in April 2018]; prior to doing so, he wrote, on Facebook, 'The Incel Rebellion has already begun!' You might also include Christopher Harper-Mercer, who killed nine people, in 2015, and left behind a manifesto that praised Rodger and lamented his own virginity." Jia Tolentino, "The Rage of the Incels", *The New Yorker* (15 May 2018), online: <https://newyorker.com/culture/cultural-comment/the-rage-of-the-incels>.

[292] Chris Herhalt, "Massage parlour stabbing was act of 'incel' terrorism: RCMP" (19 May 2020), online: *CTV News* <https://toronto.ctvnews.ca/massage-parlour-stabbing-was-act-of-incel-terrorism-rcmp-1.4945411>.

[293] "There are no far-right groups on Canada's terror watchlist. This expert says we need to talk about that", *CBC News* (20 March 2019), online: <https://www.cbc.ca/radio/thecurrent/the-current-for-march-20-2019-1.5063841/there-are-no-far-right-groups-on-canada-s-terror-watchlist-this-expert-says-we-need-to-talk-about-that-1.5063845>.

[294] Mack Lamoureux, "Police Charge Canadian Teenager With Terrorism in Alleged Incel Murder" (19 May 2020), online: *Vice* <https://www.vice.com/en_ca/article/wxq8n4/police-charge-canadian-teenager-with-terrorism-in-alleged-incel-murder>.

[295] Stewart Bell, "RCMP adding incels to terrorism awareness guide" (9 June 2020), online: *Global News* <https://globalnews.ca/news/7021882/rcmp-incel-terrorism-guide/>.

[296] Canada, Canadian Security and Intelligence Service, *CSIS Public Report 2019* (Ottawa: Public Works and Government Services Canada, 2020) at 13, online (pdf): *Government of Canada* <https://www.canada.ca/content/dam/csis-scrs/documents/publications/PubRep-2019-E.pdf>; and Stewart Bell, "Incels labelled violent extremists in latest CSIS annual report" (20 May 2020), online: *Global News* <https://globalnews.ca/news/6965806/incels-violent-extremism-csis-report/>.

Many such initiatives emerged on an improvisational and reactive basis, responding to user outcry, threat or fear of regulation, public or political pressure, or negative media attention. Although platforms have become increasingly more deliberative and proactive with respect to content moderation over time, many of their decisions, enforcement actions, and policies remain problematic and deficient, particularly with respect to gender-based violence, abuse, and harassment. Such measures differ from platform to platform in their finer details, though they can be grouped by certain commonalities.

For example, Robyn Caplan has classified content moderation systems into three models that each strike a different "balance between context-sensitivity and consistency", depending on platform companies' "missions, business models, and size of [the content moderation] team": artisanal, community-reliant, and industrial.[297] Under the "artisanal" model, smaller companies such as Medium, Vimeo, Patreon, and Discord rely entirely on in-house staff to form content moderation teams—commonly deemed the Trust and Safety team—that could fit entirely within a modestly sized office.[298] Automated content detection or algorithmic moderation is limited or not used at all, and scale and stakes (along with reach, reputation, and financial consequences) are considered less of an issue as compared to their gargantuan counterparts such as Facebook and Twitter—relatively less content is flagged across the platform, and staff spend more time reviewing content that is reported ("10 to 20 minutes" at Discord,[299] in contrast to "less than 30 seconds" at Facebook[300]), while policies evolve incrementally and improvisationally, according to staff capacity.[301]

Under the "community-reliant" model, platforms such as Reddit and Wikipedia "have created structures for large groups of volunteer users to implement and add to the overarching policy decisions of a small team employed by the company."[302] The platform "sets minimum standards they should be able to enforce with [volunteer users within different] subcommunities responsible for adopting specific rules in relation to their own communities."[303] Consequently, Reddit's volunteer subreddit moderators and Wikipedia's volunteer editors and contributors play a far more central and infrastructural role compared to the average user on other platforms, leading to unique complications in the relationships between the platform's employees and their respective user bases, who are essentially relied on as a fundamental pillar of unpaid content moderation labour.[304] In the context of

---

[297] Robyn Caplan, "Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches" (2018) at 1, online (pdf): *Data & Society* <https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf>.

[298] *Ibid* at 17-18.

[299] *Ibid* at 18.

[300] Casey Newton, "The Trauma Floor: The secret lives of Facebook moderators in America" (25 February 2019), online: *Verge* <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>.

[301] Robyn Caplan, "Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches" (2018) at 17, 19, online (pdf): *Data & Society* <https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf>.

[302] *Ibid* at 20.

[303] *Ibid* at 23.

[304] *Ibid* at 22-23.

TFGBV, it bears mentioning that 84-91% of Wikipedia's active editors are men and just under 18% of Wikipedia's more than 1.5 million biographies are about women (as of a 2019 article).[305]

Under the "industrial" model, platforms such as Facebook and Google's YouTube use a variety of methods to moderate content at extraordinary scale, relying on a higher degree of formalization of content moderation rules, outsourced content moderation "decision factories" with up to tens of thousands of low-level paid moderators, and significant leveraging of automated detection and takedown tools.[306] The variety across different types of digital platforms, including their respective sizes and purposes, suggests that proposed legal reforms may need to incorporate a certain degree of flexibility, or a sliding scale approach, to account for smaller or non-profit platforms as well as massive corporate platforms.

The remainder of this subsection will assess key content moderation mechanisms common to several major platforms, with respect to expression-based TFGBV. The mechanisms are as follow: community standards; user flagging and reporting; human review (in-house and third-party content moderators); automated moderation (algorithmic detection and takedowns, artificial intelligence, and content filters); ranking and recommendation algorithms; fact-checking, external partnerships and industry initiatives; and external or quasi-external content moderation bodies.

3.3.1. Community Standards

Nearly all digital platforms bind their users to terms of service (also known as terms of use or terms and conditions), and a set of community standards with respect to user content and behaviour, as a condition of using the platform. The community standards may be a separate document or may be embedded within the terms of service (either in full, or by reference, through a clause that states users must adhere to the platform's community standards as a term of service). Community standards cover a range of guidelines and potential infractions, such as (in the case of Facebook) prohibiting "content that: is hate speech, threatening, or pornographic; incites violence; or contains nudity or graphic or gratuitous violence",[307] where hate speech is defined as

> a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for

[305] Jessica Kantor, "Wikipedia still hasn't fixed its colossal gender gap" (13 November 2019), online: *Fast Company* <https://www.fastcompany.com/90429161/wikipedia-still-hasnt-fixed-its-colossal-gender-gap>; See also: Nicole Torres, "Why Do So Few Women Edit Wikipedia?" (2 June 2016), online: *Harvard Business Review* <https://hbr.org/2016/06/why-do-so-few-women-edit-wikipedia>; and Emma Paling, "Wikipedia's Hostility to Women", *The Atlantic* (21 October 2015), online: <https://www.theatlantic.com/technology/archive/2015/10/how-wikipedia-is-hostile-to-women/411619/>.

[306] Robyn Caplan, "Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches" (2018) at 23-24, online (pdf): *Data & Society* <https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf>.

[307] Anat Ben-David & Ariadna Matamoros-Fernández, "Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain" (2016) 10 International Journal of Communication 1167 at 1169.

> immigration status. We define "attack" as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation.[308]

Similarly, Google's hate speech policy "explicitly prohibit[s] YouTube videos promoting violence or hatred to justify discrimination, segregation or exclusion based on qualities like age, gender, race, caste, religion, sexual orientation or veteran status" and the site has "remove[d] content on YouTube denying that well-documented violent events, like the Holocaust, took place."[309] In June 2020, Reddit introduced a new content policy that explicitly stated, "Communities and users that incite violence or that promote hate based on identity or vulnerability will be banned."[310]

The contents of community standards, as well as platforms' enforcement of such standards and associated terms of service, have continually been the focus of substantial criticism over the years,[311] particularly in the context of online abuse targeting women and girls, including sexist or misogynistic memes, posts, videos, groups, and pages. For example, certain exceptions to rules prohibiting hateful or harmful speech have constituted major loopholes allowing demonstrably hateful or harmful content to remain and spread on the platform. According to the 2020 Facebook Civil Rights Audit, "humor [as an exception to hate speech] was not well-defined and was largely left to the eye of the beholder — increasing the risk that the exception was applied both inconsistently and far too frequently."[312] Platforms have also been criticized for having too narrow definitions within their community standards to begin with, such as the fact that Facebook has banned the terms "white nationalism" and "white separatism" yet continues to permit "content that explicitly espouses the very same ideology without using those exact phrases".[313]

Other exceptions such as "newsworthiness",[314] alongside exceptions regarding politicians and public figures, have also allowed hate speech or otherwise harmful speech that clearly violated the relevant platform's community standards to remain available and disseminated across the platform, moreover

---

[308] Evelyn Douek, "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2019) 21:1 North Carolina Journal of Law & Technology 1 at 27.

[309] Canada, Parliament, House of Commons, *Taking Action to End Online Hate: Report of the Standing Committee on Justice and Human Rights*, 42nd Parl, 1st Sess (June 2019) (Chair: Anthony Housefather) at 26.

[310] See Rule 1 in "Reddit Content Policy", online: *Reddit* <https://redditinc.com/policies/content-policy>.

[311] "While Facebook's Community Standards prohibit hate speech, harassment, and attempts to incite violence through the platform, civil rights advocates contend that not only do Facebook's policies not go far enough in capturing hateful and harmful content, they also assert that Facebook unevenly enforces or fails to enforce its own policies against prohibited content. Thus harmful content is left on the platform for too long." "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 42, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>.

[312] On the auditors' recommendation, Facebook has "eliminated humor as an exception to its prohibition on hate speech, instead allowing only a narrower exception for content meeting the detailed definition of satire. Facebook defines satire as content that 'includes the use of irony, exaggeration, mockery and/or absurdity with the intent to expose or critique people, behaviors, or opinions, particularly in the context of political, religious, or social issues. Its purpose is to draw attention to and voice criticism about wider societal issues or trends.'": *Ibid* at 44.

[313] *Ibid* at 50-51.

[314] See e.g., "Another example is Facebook's exception to its Community Standards for content it deems to be 'newsworthy.' These cases require a balancing of the harm caused by allowing speech that breaches Facebook's rules to remain on the platform against the public interest in being informed about the particular matter." Evelyn Douek, "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2019) 21:1 North Carolina Journal of Law & Technology 1 at 14-15 (footnotes omitted).

with the added strength of such views being espoused by a powerful public figure or politician.[315] Such exceptions, allowing certain individuals to remain active and maintain massive public reach on platforms where they would have been long banned if they were an everyday user, contribute to platformed TFGBV against women and girls. For instance, after the 45th US president attacked an 18-year-old girl on Twitter for challenging him at a forum in October 2015, his followers posted photos of her alongside her phone number and other personal information, including her address, and she subsequently received "threatening, sexually explicit calls, voicemails, emails, and Facebook messages" for a *year*.[316]

Kadri and Klonick have shown how the "public figure" exception in privacy and defamation law in the United States have influenced platforms' own approaches to such an exception in enforcing community standards,[317] and this exception perpetuates online abuse against women in two ways. The first way is as described above, where abusive speech by individuals deemed public figures, such as male politicians, is not removed or otherwise moderated, but allowed to remain, be amplified, and used to mobilize followers into further harassing the victim. The second way is where the targeted individual is herself considered a public figure, and thus often thought to be less 'deserving' of protection against abusive speech. As an example,

> actress and comedienne Leslie Jones […] was inundated with racist and sexist comments on Twitter after she starred in the all-female Ghostbusters remake. There is no doubt that Jones's fame makes her a public figure under defamation and privacy law. But whether—as a normative matter—she deserves the harsher […] rules [which expose her to increased TFGBV] that accompany public-figure status on Facebook is a far harder question.[318]

Platforms' community standards and enforcement policies have also been shown to be ignorant of historical and contemporary context related to systemic discrimination and substantive inequality. For example, in June 2017, *Pro Publica* reported that internal documents at Facebook trained content

---

[315] See, e.g., the many instances in which Facebook and Twitter refused to enforce their own content moderation policies against posts by the U.S. president, Donald Trump: Naomi Nix & Kurt Wagner, "Facebook Watchdog Rips Inaction Over Misleading Trump 2020 Posts" (8 July 2020), online: *Bloomberg* <https://www.bloomberg.com/news/articles/2020-07-08/facebook-trump-stance-paves-way-to-voter-suppression-audit-says?sref=dZ65CIng>; Mike Isaac & Cecilia Kang, "Facebook Says It Won't Back Down From Allowing Lies in Political Ads", *The New York Times* (9 January 2020), online: <https://www.nytimes.com/2020/01/09/technology/facebook-political-ads-lies.html>; Nick Corasaniti, "Will Twitter Draw a Line for Trump?", *The New York Times* (26 May 2020), online: <https://www.nytimes.com/2020/05/26/us/politics/trump-twitter-kara-swisher.html>; Rebecca Shabad, "Rep. Ilhan Omar says Trump tweet 'spread lies that put my life at risk'" (18 September 2019), online: *NBC News* <https://www.nbcnews.com/politics/congress/rep-ilhan-omar-says-trump-tweet-spread-lies-put-my-n1055951>.

[316] (*Content Warning*) Libby Nelson, "Donald Trump has weaponized Twitter — with dangerous consequences" (10 December 2016), online: *Vox* <https://www.vox.com/2016/12/10/13901238/trump-twitter-harassment-criticism-jones> The threats consisted of statements in the vein of the following: "Wishing I could f—ing punch you in the face. id then proceed to stomp your head on the curb and urinate in your bloodied mouth and i know where you live, so watch your f—ing back punk." Claire Landsbaum, "Donald Trump's Harassment of a Teenage Girl on Twitter Led to Death and Rape Threats" (9 December 2016), online: *The Cut* <https://thecut.com/2016/12/trumps-harassment-of-an-18-year-old-girl-on-twitter-led-to-death-threats.html>.

[317] Thomas E Kadri and Kate Klonick, "Facebook v Sullivan: Public Figures and Newsworthiness in Online Speech" (2019) 93 Southern California Law Review 37.

[318] *Ibid* at 84-85.

moderators to remove speech attacking "white men"—due to both race and gender being protected categories—but allow speech attacking "female drivers", "black children", and "radicalized Muslims", because for each of those described groups, "one of their characteristics is not protected".[319] Such a blunt approach all but defeats the purpose of attempts to moderate hate speech and otherwise abusive expression to begin with, by neglecting the central fact that it is precisely women, Black people, and Muslims who are members of historically marginalized groups and disproportionately subjected to online abuse and the ones in need of protection through content moderation policies, regardless of their mode of transport, age, or political views, respectively.

Moreover, Facebook has a long history of refusing to take down posts and entire pages containing violent rape jokes[320] or dedicated to advocating for rape or endorsing intimate partner violence,[321] yet is markedly active in enforcing policies for the most minor slights against men. In 2017, for example, "Facebook kicked [comic Marcia] Belsky off the platform for 30 days" after she commented, "Men are scum" in response to *Full Frontal with Samantha Bee* writer "Nicole Silverberg [sharing] on her Facebook page a trove of bilious comments directed at her after she'd written a list of ways men 'need to do better'" in context of the #MeToo movement against sexual assault of women by men in positions of power.[322] That is, Facebook's community standards and content moderation policies have permitted explicit endorsements of sexual violence against women, yet *kicked off the platform* female users for mild insults about men hurling online abuse at a woman for speaking out about how men can help reduce sexual assault. Similarly, Twitter's suspension of Rose McGowan at an early height of #MeToo, for tweeting a phone number among commentary indicting male actors' complicity, brought to the forefront that "while victims of abuse and marginalized users who deal with harassment are frequently censured over strict readings of Twitter's abuse and safety rules, like McGowan, users who are widely seen as perpetuating real ideological violations of those rules are rarely censored."[323]

There is no shortage of additional instances of seeming double standards and hypocrisy among platforms' content moderation policies and decisions, cutting across sexism, misogyny, and racism,[324]

---

[319] Julia Angwin, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children", *ProPublica* (28 June 2017), online: <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>.

[320] Simon van Zuylen-Wood, "'Men Are Scum': Inside Facebook's War on Hate Speech", *Vanity Fair* (March 2019), online: <https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech>.

[321] Rory Carroll, "Facebook gives way to campaign against hate speech on its pages", *Guardian* (29 May 2013), online: <https://www.theguardian.com/technology/2013/may/29/facebook-campaign-violence-against-women?CMP=gu_com>.

[322] Simon van Zuylen-Wood, "'Men Are Scum': Inside Facebook's War on Hate Speech", *Vanity Fair* (March 2019), online: <https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech>.

[323] Aja Romano, "Twitter's suspension of Rose McGowan epitomizes the site's most infuriating problem" (12 October 2017), online: *Vox* <https://www.vox.com/culture/2017/10/12/16464752/twitter-suspended-rose-mcgowan>. Another Twitter user revealed, in response to the McGowan incident, "Some alt-right dickbag tweeted my phone number last winter, and when I reported it Twitter denied it was a violation of terms of service." *Ibid*.

[324] See e.g., Sam Levin, "Civil rights groups urge Facebook to fix 'racially biased' moderation system", *Guardian* (18 January 2017), online: <https://www.theguardian.com/technology/2017/jan/18/facebook-moderation-racial-bias-black-lives-matter>: "Activists in the Movement for Black Lives have routinely reported the takedown of images discussing racism and during protests, with the justification that it violates Facebook's Community Standards. At the same time, harassment and threats directed at activists based on their race, religion, and sexual orientation is thriving on Facebook. Many of these activists have reported such harassment and threats by users and pages on Facebook only to be told that they don't violate Facebook's Community Standards." See also Olivia Solon, "Facebook ignored racial bias research, employees say", *NBC News* (23 July 2020), online: <https://www.nbcnews.com/tech/tech-news/facebook-management-ignored-internal-

while reflecting harmful cultural biases[325] and stereotypes around female anatomy, female sexuality, and sexual objectification of women. Dragiewicz et al. point out:

> Facebook enforces its nudity and obscenity policy in a narrow fashion that often ignores the context and cultural specificities of nude bodies. This was the case when the platform removed photos of breastfeeding [...] or pictures of female Indigenous elders with uncovered breasts participating in cultural celebrations [...]. However, when women's organisations reported a page that glorified DV [domestic violence], Facebook responded by only deleting the most controversial posts [...]. It was not until the media started covering the case that Facebook closed the page [...].[326]

When questioned about "treating different groups differently"—i.e., applying a substantive equality approach to content moderation, rather than a flawed "neutral" approach that exacerbates inequality—a spokeswoman for Facebook stated "that it is 'very difficult to parse out who is privileged and who is marginalized globally' and so the company has not changed its policies."[327] This reason for inaction seems tenuous at best, given both the many legal, political, public policy, and sociological contexts in which governing actors must and already do determine who is privileged or marginalized for the purposes of particular decisions; the availability of copious research, academic literature, and statistics to support such determinations; and the fact that digital platforms such as Facebook already modify their content policy regimes from jurisdiction to jurisdiction, due to differing legal requirements, and thus already do not adhere to one universal policy across the globe.

---

research-showing-racial-bias-current-former-n1234746>: "This inequity is reflected in the levels of hate speech that is reported versus taken down automatically. According to a chart posted internally in July 2019 and leaked to NBC News, Facebook proactively took down a higher proportion of hate speech against white people than was reported by users, indicating that users didn't find it offensive enough to report but Facebook deleted it anyway. In contrast, the same tools took down a lower proportion of hate speech targeting marginalized groups including Black, Jewish and transgender users than was reported by users, indicating that these attacks were considered to be offensive but Facebook's automated tools weren't detecting them."

[325] See e.g., "Although Liddle explained the cultural relevance of posting a picture of two topless Aboriginal women performing a traditional ceremony, Facebook decided to continue banning the photograph and temporary block Liddle for repeatedly posting it." At the same time, "Facebook has also been notorious for refusing to ban racist pages towards Aboriginal people. In 2012 and 2014, the Online Hate Prevention Institute (OHPI) went through extensive negotiations with Facebook to get the platform to remove several pages containing racist attacks on Indigenous Australians (Oboler, 2013). Facebook initially ruled that the pages did not breach its terms of service and instead compelled their creators to rename them to note that they were 'controversial content' (Oboler, 2013). Not until the Australian Communications and Media Authority was involved did Facebook decide to block these pages, but even then only in Australia (Oboler, 2013)." Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930 at 941, 931.

[326] Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 617 (in text citations omitted); Laura Meachim & Kym Agius, "Blokes Advice: Facebook refuses to remove group despite members' threats against women", *ABC News* (29 July 2016), online: <https://www.abc.net.au/news/2016-07-29/facebook-will-not-remove-blokes-advice-page-over-threats/7668174>.

[327] Olivia Solon, "Facebook ignored racial bias research, employees say", *NBC News* (23 July 2020), online: <https://www.nbcnews.com/tech/tech-news/facebook-management-ignored-internal-research-showing-racial-bias-current-former-n1234746>.

3.3.2. User Flagging and Reporting

User flagging and reporting is one constant across nearly all digital platforms, and one of the main front lines of content moderation where speech-based abuse is concerned, even where a platform also relies on automated moderation or content filters. For instance, "Twitter has a mechanism that allows the user to download a written report in cases of [TFGBV]; it contains the specific tweet, the URL of the tweet, the time stamp, the URL and name of the user who shared it, and a link to the law enforcement guideline in your jurisdiction."[328] However, this does not help if the initial author has deleted the tweet before it was reported or downloaded.

On Facebook, "users flag millions of pieces of content worldwide every week. […] Facebook's reporting process attempts not only to guide users toward categorizing which aspect of the Community Standards is being violated to expedite review but also to urge users toward self-resolution of disputes that likely fall outside the Standards' purview."[329]

Two main observations explain the deficiency of user flagging and reporting as a reliable mechanism to stem the tide of sexist and misogynistic content and similarly abusive speech on digital platforms. First, the scale of platforms such as Facebook and corresponding volume of reported content contains a high noise-to-signal ratio, including users misunderstanding or purposely gaming the flagging feature to silence members of marginalized communities.[330] "Users have many reasons for flagging content, and much of what they flag does not violate the Community Standards. Rather, a vast majority of flagged content reflects personal opinion, conflict between groups or users, or even abuse of the flagging system to harass other users."[331] This diverts from and dilutes efforts to address abusive content that is flagged, and given that this mechanism is both lightweight and entirely dependent on user discretion, user flagging "may be structurally insufficient to serve the platforms' obligations to public discourse".[332]

Second, abusive speech addressed only through user flagging and reporting constitutes a complaint-driven system that addresses TFGBV entirely on a one-off, case-by-case, reactive basis, assuming a user has the wherewithal to report content in the first place, rather than a systemic, platform-wide, and proactive response that would mitigate or prevent the proliferation of online abuse to begin with. Such a system 'responsibilizes' users[333]—and disproportionately marginalized users at that—who are targets of speech-based abuse, and forces them to bear the burden of seeking accountability, redress, and

---

[328] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 52.

[329] Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129 Yale Law Journal 2418 at 2432-33.

[330] "Anti-racism activists and other users have reported being subjected to coordinated reporting attacks designed to exploit this potential for content reviewing errors. Those users have reported difficulty managing the large number of appeals, resulting in improper use restrictions and other penalties." "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 48, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>.

[331] Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129 Yale Law Journal 2418 at 2432.

[332] Kate Crawford & Tarleton Gillespie, "What is a flag for? Social media reporting tools and the vocabulary of complaint" (2016) 18:3 New Media & Society 410 at 424.

[333] Lisa Sugiura & April Smith, "Victim Blaming, Responsibilization and Resilience in Online Sexual Abuse and Harassment" in Jacki Tapley & Pamela Davies, eds, *Victimology: Research, Policy and Activism* (London: Palgrave Macmillan, 2020) 45 at 55, 62.

functioning content moderation in addition to living through the impacts of the abuse itself. This reflects a form of technological solutionism (a term coined by Evgeny Morozov), where suggested remedies "are often based in an autonomous individualist mindset. Technology is often used to shift the burden of solving these problems to the individual, frequently assuming that having such a responsibility is empowering."[334] In the context of addressing platformed TFGBV, the platform has offloaded its content moderation work onto the small subset of users willing to, or who have no choice but to, shoulder the additional labour as part of their personal price of using the platform.[335]

3.3.3. Human Review (Content Moderators)

Human review simply refers to a platform's reliance on human staff, contractors, or designated users to review content that has been flagged for removal or otherwise reported, as opposed to algorithmic review done by automated content moderation tools.[336] Human reviewers are meant to assess flagged content for context and nuance (such as applying broader cultural context, or evaluating whether a user is engaging in hate speech or quoting hate speech to condemn it),[337] and human review (or 'manual review') is also used to check content moderation decisions made by algorithms[338] or by lower-level human moderators.

Results across platforms, particularly the largest ones, frequently fall far short of what human review is meant to accomplish. This is due to reasons such as insufficient training, lack of guidance around how to interpret and apply community standards, the community standards themselves being flawed as described above, or incorrect interpretation and application of community standards even where they are not inherently faulty. Other issues include human moderators lacking the necessary context, knowledge, or even language to assess certain pieces of content, especially where moderators in one country must review content based in another country, such as those in the Philippines trying to assess whether or not a particular phrase constitutes hate speech in Canada, or a moderator in the United

---

[334] R Stuart Geiger, "Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space" (2016) 19:6 Information, Communication & Society 787 at 792.

[335] This is, moreover, on top of the gendered and racialized nature of major platforms' paid content moderators who work under abysmal working conditions. See e.g., Lindsay Bartkowski, "Caring for the Internet: Content Moderators and the Maintenance of Empire" (2019) 4:1 Journal of Working Class-Studies 66; Bryan Dosono & Bryan Semaanm "Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit" (Paper delivered at the CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), Glasgow, Scotland, 4-9 May 2019), online (pdf): <http://library.usc.edu.ph/ACM/CHI2019/1proc/paper142.pdf>.

[336] "Human decision-making is an essential part of speech moderation, especially as Facebook increasingly uses automated methods to both initially identify and adjudicate content. Historically, Facebook's proactive moderation was largely confined to certain kinds of extreme content and limited by the nascent state of video- and photo-recognition technology. The platform predominantly depended on users to flag violating speech." Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129 Yale Law Journal 2418 at 2417-18 (footnotes omitted). See also the various role of human reviewers in the artisanal, community-reliant, and industrial content moderation models described by Caplan: Robyn Caplan, "Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches" (2018) at 1, online (pdf): *Data & Society* <https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf>.

[337] "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 42, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>.

[338] Daphne Keller, "Internet Platforms: Observations on Speech, Danger, and Money" (2018) at 7, online: *Hoover Institution* <https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf>.

States trying to evaluate the local context of a post in Hong Kong. As Matamoros-Fernández notes, "Subjectivity is unavoidable in content moderation and some decisions can be attributed to the cultural background of platforms' moderators".[339]

Most of all, content moderators working for some of the largest platforms must function under notoriously dismal working conditions and review content at rates that make it virtually impossible to truly apply context or nuance to borderline cases or otherwise where the correct decision is not obvious, even if the above deficiencies were remedied.[340] Sarah T. Roberts has written extensively on the labour conditions of third-party content moderators, including low pay, little to no autonomy (including to go to the bathroom or take breaks), high production pressures to process as much content as quickly as possible, and most prominently and concerningly, short-term and long-term psychological damage from examining thousands of highly graphic, violent and traumatizing images, video, and speech all day, every day.[341] Human content moderators under the industrial model are not treated with the same respect as, for instance, platform companies' (disproportionately white and male) product engineers or other in-house staff based in company headquarters. Bartkowski points out that this divide (between a disproportionate number of white men among well-paid engineers, and a disproportionate number of racialized men and women among underpaid contractor content moderators) reproduces "the gendered, racialized division of labor", where content moderation is the invisible and underpaid "care work" of the Internet.[342] "This 'unseen work' tends to favour platforms' profit seeking and legal demands rather than responding to social justice or advocacy-related goals".[343]

Human content moderation at scale is thus already a hugely labour-intensive and problematically gendered and racialized undertaking even when flawed and deficient, as demonstrated when Facebook stated in 2017, in response to public and political pressure, that it would hire up to 20,000 more

---

[339] Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930 at 937.

[340] See e.g., "Casey Newton, "The Trauma Floor: The secret lives of Facebook moderators in America" (25 February 2019), online: *Verge* <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>; Casey Newton "Bodies in Seats" (19 June 2019), online: *Verge* <https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa>; Sarah Emerson, "'A Permanent Nightmare': Pinterest Moderators Fight to Keep Horrifying Content Off the Platform" (28 July 2020), online: *OneZero* <https://onezero.medium.com/a-permanent-nightmare-pinterest-moderators-fight-to-keep-horrifying-content-off-the-platform-4d8e7ec822fe>.

[341] Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (New Haven and London: Yale University Press, 2019); see also Paul M Barrett, "Who Moderates the Social Media Giants? A Call to End Outsourcing" (June 2020), online (pdf): *NYU Stern Center for Business and Human Rights* <https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf>.

[342] Lindsay Bartkowski, "Caring for the Internet: Content Moderators and the Maintenance of Empire" (2019) 4:1 Journal of Working-Class Studies 66 at 69. See also "By distributing the care work of content moderation among digital laborers of the Global South, U.S. corporations, acting in cooperation with state powers, repurpose historical domestic labor practices on a transnational scale." *Ibid* at 70.

[343] Ariadna Matamoros-Fernández, "Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube" (2017) 20:6 Information, Communication & Society 930 at 931. See also "[O]ne issue with this model is that it requires a substantial amount of emotional labor performed by communities or nonprofit organizations, which ultimately benefits the for-profit corporation Twitter, Inc." R Stuart Geiger, "Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space" (2016) 19:6 Information, Communication & Society 787 at 792.

moderators.[344] Legal or policy reforms that require more of human reviewers must thus also include the necessary labour reforms to ensure such workers are appropriately compensated, have access to mental health care such as therapy and counselling, and otherwise are provided with professional working conditions that treats them with dignity.

3.3.4. Automated Moderation: Algorithms and Artificial Intelligence

Digital platforms have increasingly relied on a variety of automated content moderation tools, some involving artificial intelligence or machine learning algorithms, as problematic content continues to cause high-profile issues across a variety of contexts in addition to that of TFGBV—including disinformation, electoral integrity, and public health—and continued public and political pressure has forced companies to take content moderation issues more seriously. Such tools include automated sorting of content that has been flagged by users, automated filtering or blocking of content, and automated detection and takedowns of certain specific kinds of content,[345] such as content that is deemed to have infringed copyright, images containing nudity, intimate photos that have been uploaded without consent, and child sexual exploitation materials.[346] In some cases, human review and automated review are combined and leveraged to improve the accuracy of each other.[347] Hate speech is one of the categories on Facebook that even if machine detected, will be sent to human moderators for review.[348] Facebook and Google have both stated they are developing algorithmic tools to address

---

[344] Elizabeth Dwoskin, Jeanne Whalen & Regine Cabato, "Content moderators at YouTube, Facebook and Twitter see the worst of the web — and suffer silently", *The Washington Post* (25 July 2019), online: <https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/>.

[345] See e.g., "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 42, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>: "Facebook reports that it removes some posts automatically, but only when the content is either identical or near-identical to text or images previously removed by its content review team as violating Community Standards, or where content very closely matches common attacks that violated policies. Facebook states that automated removal has only recently become possible because its automated systems have been trained on hundreds of thousands of different examples of violating content and common attacks." It is notable that these descriptions are available only through Facebook itself, without indication that they have been independently verified, which speaks to broader transparency issues when it comes to platform accountability and liability.

[346] See e.g., Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129 Yale Law Journal 2418 at 2430-31 (Facebook); landoflobsters, "Changes to Our Policy Against Bullying and Harassment" (30 September 2019), online: *Reddit*, <https://www.reddit.com/r/announcements/comments/dbf9nj/changes_to_our_policy_against_bullying_and/> ("You should also know that we'll also be harnessing some improved machine-learning tools to help us better sort and prioritize human user reports. But don't worry, machines will only help us organize and prioritize user reports. They won't be banning content or users on their own. A human user still has to report the content in order to surface it to us. Likewise, all actual decisions will still be made by a human admin.")

[347] See e.g., "Facebook can often algorithmically identify new spam postings based on the behavior of the poster, the proliferation on the site, and the targeting of the content. But Facebook also informs its algorithms and automatic takedowns with information gathered from users reactively and manually reporting spam. Thus, the manual and automatic systems work together to iteratively develop a database of content to be automatically removed from the site." Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129 Yale Law Journal 2418 at 2431.

[348] "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 42, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>.

harassment and abuse specifically;[349] however, it remains to be seen how these tools will work or if they will achieve their intended purpose.[350] Researchers have also attempted to develop content moderation algorithms that would automatically detect hate speech and related abuse, including misogynistic content, though not without challenges.[351] Outside of platform companies, some have developed automated tools to respond to abusive speech in certain contexts, such as the Edmonton-based ParityBOT, which used "artificial intelligence to send positive tweets in response to abusive ones directed at women running in the [2019 Canadian] federal election",[352] and anti-racism bots, which responded to people tweeting racial slurs with a reminder of the targeted individual's humanity.[353]

Automated content moderation, including algorithmic detection and automatic content removal, faces particular difficulties in the context of abusive, sexist, or misogynistic speech on digital platforms. This is because such speech "is contextually and culturally specific, and is often disseminated through coded language, images, gifs, and memes".[354] Researchers working on such technology have encountered the issue firsthand, such as in trying to develop automated detection of misogynistic tweets:

> Misogynistic tweet detection is challenging for text classification methods because social media users very commonly use offensive words or expletives in their online dialogue. For example, the bag-of-words approach is straightforward and usually has a high recall, but it results in a higher number of false positives because the presence of misogynistic words causes these tweets to be misclassified as abusive tweets […]
>
> Misogynistic abusive tweets may contain misogynistic keywords, but tweets can also be misogynistic abuse without explicitly containing these slurs. Further, not all tweets that contain misogynistic keywords are abusive. Classifying misogynistic abuse in tweets

---

[349] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu) at 54.

[350] See e.g., "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 81, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>: "When it comes to Facebook's own algorithms and machine learning models, the Auditors cannot speak to the effectiveness of any of the pilots Facebook has launched to better identify and address potential sources of bias or discriminatory outcomes."

[351] See e.g., Md Abul Bashar et al, "Misogynistic Tweet Detection: Modelling CNN with Small Datasets" (Paper delivered at Data Mining: 16th Australasian Conference, AusDM2018, 28-30 November 2018), in Rafiqul Islam et al, eds, *Data Mining: 16th Australasian Conference, AusDM 2018, Bahrust, NSW, Australia, November 28-30, 2018, Revised Selected Papers* (Singapore: Springer, 2019); Paula Fortuna & Sérgio Nunes, "A Survey on Automatic Detection of Hate Speech in Text" 51:4 ACM Computing Surveys 1; Sean MacAvaney et al, "Hate speech detection: Challenges and solutions" (2019) 14:8 PLoS ONE.

[352] Emily Mertz, "'ParityBOT' uses AI to combat negative tweets towards female candidates" (27 September 2019), online: *Global News* <https://globalnews.ca/news/5951313/paritybot-twitter-women-candidates-politics-ai/>.

[353] Ananya Bhattacharya, "Racist tweeters can be convinced to stop spreading hate—if a white man asks them to" (18 November 2016), online: *Quartz* <https://qz.com/840060/racist-tweeters-can-be-convinced-to-stop-spreading-hate-if-a-white-man-asks-them-to/>.

[354] Katherine Feenan & Kathleen Donovan, "Online Culture Shift: Safer Platforms for Women in Politics" (August 2019) at 22, online (pdf): *Public Policy Forum* <https://ppforum.ca/wp-content/uploads/2019/08/OnlineCultureShift-PPF-Aug2019-EN.pdf>; see also "For every category except bullying and harassment, and hate speech, Facebook found over 95% of the content it took down as violating its Community Standards before it was reported by a user, in large part because of its AI. But the exception of bullying and harassment, and hate speech is telling: these two categories of content are harder for Facebook to proactively identify because they are so highly context-dependent. … Hate speech is notoriously difficult to detect through automated processes, because it depends so much on linguistic nuance, intention, and local norms." Evelyn Douek, "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2019) 21:1 North Carolina Journal of Law & Technology 1 at 12-13.

> requires close reading, and even humans can struggle to classify these tweets accurately.[355]

Algorithms also introduce the problem of algorithmic bias, where algorithms trained on biased data sets end up perpetuating and entrenching that bias through the algorithms' subsequent decisions, whose outputs may then be fed into future algorithms, in an increasingly biased feedback loop. For example, Sap et al. have demonstrated racial bias in certain hate speech detection algorithms, where "tweets inferred to be in AAE [African American English] and tweets from self-identifying African American users are more likely to be classified as offensive" compared to other users.[356] If deployed across social media platforms, this would result in those who use (what the study defines as) African American English being disproportionately silenced through wrongful removals, "further suppressing already-marginalized voices".[357]

Moreover, automated takedowns of user content on digital platforms have long been associated with over-removal and wrongful removal of legitimate or beneficial speech,[358] including that of women, girls, and users with other or intersecting marginalized identities, and posts that bring attention to or censure abusive speech. Automated wrongful removals, combined with lack of due process or appeal mechanisms, risk further compounding the systemic marginalization of women and girls online, on top of the silencing effects of being the recipients of the high volumes of abusive speech that are not captured by either automated or human moderation processes.

3.3.5. Ranking and Recommendations

There is increasing recognition that content moderation does not have to involve binary decisions of leaving a flagged post up and allowing it to reach as many people as it will, and taking the post down altogether.[359] For some types of content, it may be appropriate to instead algorithmically or manually adjust its ranking, reach, or whether or not it appears in recommendations and other channels of promoted content; place a screen before such content ('quarantining' it); or tweak the platform's overall curation and newsfeed algorithms such that abusive content surfaces less than it might

---

[355] Md Abul Bashar et al, "Misogynistic Tweet Detection: Modelling CNN with Small Datasets" (Paper delivered at Data Mining: 16th Australasian Conference, AusDM2018, 28-30 November 2018), in Rafiqul Islam et al, eds, *Data Mining: 16th Australasian Conference, AusDM 2018, Bahrust, NSW, Australia, November 28-30, 2018, Revised Selected Papers* (Singapore: Springer, 2019) at 3-4.

[356] Maarten Sap et al, "The Risk of Racial Bias in Hate Speech Detection" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics, 2019) 1668 at 1668. See also Charlotte Joe, "Google's algorithm for detecting hate speech is racially biased" (13 August 2019), online: *MIT Technology Review* <https://www.technologyreview.com/2019/08/13/133757/googles-algorithm-for-detecting-hate-speech-looks-racially-biased/>.

[357] *Ibid*.

Daphne Keller, "Empirical Evidence of Over-Removal by Internet Companies Under Intermediary Liability Laws: An Updated List" (8 February 2021), online: *Center for Internet and Society at Stanford Law School* <https://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws>.

[359] See e.g., "Often, content moderation conversations revolve around a 'take down / leave up' dichotomy. But platforms have far greater capacity to control the content on their sites than this paradigm suggests." Evelyn Douek, "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2019) 21:1 North Carolina Journal of Law & Technology 1 at 42.

otherwise on public pages. For instance, before deciding to ban a virulently misogynistic and racist subreddit r/The_Donald in 2020, Reddit took an intermediate step of "placing it behind a warning screen after it was found to host content that incited violence. The company had previously prevented posts on the forum from reaching Reddit's front page. Former users of the forum began moving to a new site off Reddit last year."[360] Similarly, Twitter decided to "stop recommending accounts and content related to QAnon [a set of far-right conspiracy theories], including material in email and follow recommendations, and it will take steps to limit circulation of content in features like trends and search."[361] The authors of the Facebook Civil Rights Audit recommended that "Facebook should do everything in its power to prevent its tools and algorithms from driving people toward self-reinforcing echo chambers of extremism, and that the company must recognize that failure to do so can have dangerous (and life-threatening) real-world consequences."[362]

The main issue with relying on ranking, recommendation, and curation algorithms and policies is their current lack of transparency. While transparency is a broader issue for all content moderation practices and platform regulation generally, its absence is particularly pertinent in this case given the potentially 'black box' nature of a platform's algorithms, on top of lacking institutional transparency by the company. For instance, Douek writes, "Facebook's decision to 'downrank a piece of content (or distribute it less) in users' News Feeds is currently much less transparent than a decision to take down a piece of content. Users are typically notified when a post is removed entirely, but, because users are not told how their posts are treated by the News Feed algorithm, may be entirely unaware when their post is left up but just not shown to other users."[363] This is a deliberate feature to prevent, for instance, abusive users from becoming aware their comments are not reaching their recipients and thus "providing fewer incentives to the commenting user to spam the page or attempt to circumvent the social networking system filters".[364] However, such opacity raises concerns for users whose content is unjustly or mistakenly subjected to downranking or otherwise suppressed by a content moderation algorithm or policy. Similarly, there is lack of transparency regarding Google's search algorithms and why certain results are provided over others, such as misogynistic results regarding Black girls,[365] and Instagram is rife with users who believe they have been 'shadowbanned' by the platform without notice or explanation, including users who post 2SLGBTQQIA or sex education content.[366]

---

[360] Casey Newton, "Reddit bans r/The_Donald and r/ChapoTrapHouse as part of major expansion of its rules" (29 June 2020), online: *Verge* <https://theverge.com/2020/6/29/21304947/reddit-ban-subreddits-the-donald-chapo-trap-house-new-content-policy-rules>.

[361] Ben Collins & Brandy Zadrozny, "Twitter bans 7,000 QAnon accounts, limits 150,000 others as part of broad crackdown", *NBC News* (21 July 2020), online: <https://nbcnews.com/news/amp/ncna1234541>.

[362] "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 56, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>.

[363] Evelyn Douek, "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2019) 21:1 North Carolina Journal of Law & Technology 1 at 42-43.

[364] *Ibid* at 43.

[365] Safiya Noble, *Algorithms of Oppression* (New York: NYU Press, 2018).

[366] 'Shadowbanning' means that "a user can continue posting as normal, but their posts will be hidden from the rest of the community". Danielle Blunt et al, "Posting Into the Void" (2020) at 15, online (pdf): *Hacking//Hustling* <https://hackinghustling.org/wp-content/uploads/2020/09/Posting-Into-the-Void.pdf>.

3.3.6. Fact-Checking, Labelling, and External Linking

Some social media platforms have implemented fact-checking programs, which operate in different ways and to different extents. Fact-checking labels as a form of content moderation for harmful speech garnered media and public attention when Twitter added such a label to tweets by the 45th US president that contained false or misleading information about mail-in ballots, inviting users to obtain information about mail-in ballots.[367] However, this occurred only after Twitter reversed its initial laissez-faire position in response to racist and threatening tweets by the now-former president about the George Floyd protests in May 2020.[368] In 2018, YouTube began linking directly to Wikipedia articles from conspiracy videos as a fact-checking measure; however, it did so without first coordinating with or even notifying the non-profit platform, which is entirely based on the community-reliant model of content moderation.[369] This led one user to comment, "Does linking result in increased traffic [from conspiracy theorists]? […] Increased vandalism? It's not polite to treat Wikipedia like an endlessly renewable resource with infinite free labor; what's the impact?"[370]

Fact-checking has more often been raised as a content moderation solution to disinformation in the context of electoral campaigns, some instances of speech by politicians, climate change, and public health—specifically, the COVID-19 pandemic and to a certain extent the anti-vaccine movement—than a potential response to TFGBV. However, much sexist and misogynistic content also includes demonstrably false or misleading assertions, disinformation or misinformation, and conspiracy theories. Issues such as sexual assault and sexual trauma, abortion and reproductive rights, and intimate partner and dating violence are also matters of public health, involve facts based on established science, and can be and often are spoken misleadingly about by politicians and other public figures. There is no reason these issues should not also benefit from any effective fact-checking or labelling measures that digital platforms have implemented for other issues, next to which gender-based violence, abuse, and harassment is no less important.

For example, on a webpage describing its fact-checking program and how it identifies misinformation, Facebook states, "We also use machine learning models to continuously improve our ability to predict misinformation. We feed ratings from our fact-checking partners back into this model, so that we get better and better over time at predicting content that could be false."[371] Potentially, Facebook could consult non-governmental organizations in Canada that work on TFGBV and possess expertise

---

[367] Alex Kantrowitz & Ryan Broderick, "Twitter Fact-Checked A Trump Tweet For The First Time" (26 May 2020), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/alexkantrowitz/twiter-fact-checked-trump>.

[368] "Twitter hides Trump tweet for 'glorifying violence'", *BBC* (29 May 2020), online: <https://www.bbc.com/news/technology-52846679>.

[369] Megan Farokhmanesh, "YouTube didn't tell Wikipedia about its plans for Wikipedia" (14 March 2018), online: *Verge* <https://www.theverge.com/2018/3/14/17120918/youtube-wikipedia-conspiracy-theory-partnerships-sxsw>. "Wikimedia added that its content is possible because of the millions of people who make donations, as well as those who edit and contribute to the site. In a series of follow-up tweets, Wikimedia notes that it has thousands of editors monitoring content and that those tracking conspiracy theories specifically have sometimes spent years doing so."

[370] Phoebe Ayers, "Also! @YouTube should probably run some A/B tests with the crew at @WikiResearch first. Does linking result in increased traffic? Increased vandalism? It's not polite to treat Wikipedia like an endlessly renewable resource with infinite free labor; what's the impact?" (13 March 2018 at 22:06), online: Twitter <https://twitter.com/phoebe_ayers/status/973742197857284096>.

[371] "How Our Fact-Checking Program Works" (11 August 2020), online: *Facebook* <https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works>.

regarding gender-based hate and discrimination and intimate partner violence—such as LEAF, the YWCA, and the BC Society of Transition Houses—as well as women's health organizations with proven qualifications and credibility, as part of this program. Such groups, and their counterparts in other jurisdictions, could become 'fact-checking partners' with respect to misinformation concerning TFGBV, violence against women and girls, sexual violence, and reproductive rights, for example, as well as provide digital platform companies with a more in-depth and accurate understanding to assist in more effectively identifying and responding to other kinds of TFGBV on their platforms.

3.3.7. External Content Moderation Bodies

Further merging between regulation *of* platforms and regulation *by* platforms is reflected in growing interest in delegating content moderation powers to administrative entities that are intended to be independent of the platform companies themselves and have binding power over them, but are still private entities rather than government agencies such as a regulator or administrative tribunal.

The sole example of this model to date is the Facebook Oversight Board (FBOB). The FBOB, which has been colloquially referred to as the 'Facebook Supreme Court',[372] consists of 40 part-time board members from around the world, who are expected to "exercise independent judgment" over Facebook's content moderation decisions that have been contested.[373] These individuals ostensibly "are not Facebook employees and cannot be removed by Facebook",[374] although the company funds their salaries[375] and selected the inaugural four co-chairs.[376] The widely publicized global selection process sought to establish "a council of sage advisers—the group eventually included humanitarian activists, a former Prime Minister, and a Nobel laureate",[377] and Facebook has suggested the FBOB may eventually expand to be an oversight board for other platform companies, or inspire them to create their own.[378] The FBOB began accepting cases in October 2020, "from a random [5%] of users, like a new Instagram feature".[379]

---

[372] See e.g., Mathew Ingram, "The Facebook Supreme Court will see you now", *Columbia Journalism Review* (19 September 2019), online: <https://www.cjr.org/the_media_today/facebook-supreme-court.php>; and Leo Kelion, "Facebook 'Supreme Court' to begin work before US Presidential vote", *BBC* (24 September 2020), online: <https://www.bbc.com/news/technology-54278788>.

[373] Nick Clegg, "Welcoming the Oversight Board" (6 May 2020), online: *Facebook* <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/>.

[374] *Ibid*.

[375] Allana Akhtar, "Facebook's Oversight Board members reportedly earn 6-figure salaries and only work 'about 15 hours a week'", *Business Insider* (13 February 2021), online: <https://www.businessinsider.com/facebook-oversight-board-members-get-six-figure-salaries-report-2021-2>.

[376] Nick Clegg, "Welcoming the Oversight Board" (6 May 2020), online: *Facebook* <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/>.

[377] Kate Klonick, "Inside the Making of Facebook's Supreme Court", *New Yorker* (12 February 2021), online: <https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court>.

[378] Nick Clegg, "Welcoming the Oversight Board" (6 May 2020), online: *Facebook* <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/>.

[379] Kate Klonick, "Inside the Making of Facebook's Supreme Court", *New Yorker* (12 February 2021), online: <https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court>.

If a user is dissatisfied with how Facebook has handled a certain content moderation decision regarding a specific post, and they have exhausted all of Facebook's internal resolution and appeal processes, then they may appeal Facebook's final decision to the FBOB.[380] At time of writing (April 2021), however, only decisions to *remove* content may be appealed, not decisions to *leave up* content, resulting in a built-in bias where potentially harmful content can only ever be restored to the site, but not taken down.[381] This is a problem for TFGBV in particular. "As it stands, the board could become a forum for trolls and extremists who are angry about being censored. But if a user believes that the company should crack down on certain kinds of speech, she has no recourse."[382]

The FBOB has been studied extensively since it was first announced, in particular by Klonick and Evelyn Douek.[383] Additional issues that they and others have pointed out include, for example, the fact that the FBOB's binding power only applies to the specific pieces of content in the cases it decides, out of the billions of pieces of content across Facebook, or the thousands of decisions for which users may request an appeal.[384] The FBOB may issue decisions containing more meaningful systemic changes, such as platform-wide policy changes, but only as non-binding recommendations.[385] On the other hand, both because of and despite Facebook's global reach, the FBOB cannot become, nor should it become, "the ultimate arbiter of free speech norms around the world."[386] Douek suggests that the "true value" of the FBOB "lies between these two extremes of individual error correction and the settlement of globally applicable speech rules."[387]

Critics of the FBOB—including a group of experts who call themselves the 'Real Facebook Oversight Board'[388]—have suggested the FBOB inappropriately prioritizes freedom of expression at the expense

---

[380] Nick Clegg, "Welcoming the Oversight Board" (6 May 2020), online: *Facebook* <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/>.

[381] Kate Klonick, "Inside the Making of Facebook's Supreme Court", *New Yorker* (12 February 2021), online: <https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court>.

[382] *Ibid*.

[383] See e.g., Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129 Yale Law Journal 2418; Evelyn Douek, "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2019) 21:1 North Carolina Journal of Law & Technology 1. In addition, the *Lawfare* blog has put together a centralized resource that monitors, analyzes, and conducts research on the FBOB and its cases: "Welcome to the FOB Blog: Overseeing the Facebook Oversight Board" (2021), online: Lawfare <https://www.lawfareblog.com/welcome-fob-blog-overseeing-facebook-oversight-board>.

[384] Evelyn Douek, "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2019) 21:1 North Carolina Journal of Law & Technology 1 at 11.

[385] Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129 Yale Law Journal 2418 at 2464.

[386] Evelyn Douek, "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2019) 21:1 North Carolina Journal of Law & Technology 1 at 7.

[387] *Ibid* at 7.

[388] Sue Halpern, "The Ad-hoc Group of Activists and Academics Convening a 'Real Facebook Oversight Board'", New Yorker (15 October 2020), online: <https://www.newyorker.com/tech/annals-of-technology/the-ad-hoc-group-of-activists-and-academics-convening-a-real-facebook-oversight-board>; and Billy Perrigo, "Facebook's Oversight Board Is Reviewing Its First Cases. Critics Say It Won't Solve the Platform's Biggest Problems", Time (7 December 2020), online: <https://time.com/5918499/facebook-oversight-board-cases/>.

of harm reduction and the rights of historically marginalized and vulnerable groups.[389] Sejal Parmar points out the potential implications of focusing on the right to freedom of expression to the exclusion of other human rights:

> It is significant that, under the [FBOB] charter, the board's decision-making is to be guided by "human rights norms" in relation to one particular right, namely freedom of expression, only. This emphasis on freedom of expression consolidates Facebook's own focus on "voice and free expression" through its values and in the public statements of its senior leaders, notably Zuckerberg. Yet, although freedom of expression may be most obviously adversely impacted by content-moderation decisions, other internationally recognized human rights – including the right to life, the right to equality before the law, the right to privacy, freedom of assembly, and the right to an adequate standard of living – can be impacted too, as the 2019 review of the board indicates. The pre-eminence of freedom of expression in the charter creates a hierarchy of human rights for the board's evaluation of content decisions, which is antithetical to an international human rights approach in principle. But it also means that, in practice, the board may be reluctant to prioritize cases which harm human rights other than freedom of expression, even if those harms are severe.[390]

Parmar's concerns appear to have been realized in the first set of decisions released by the FBOB.[391] To counter the dangers that she notes, both digital platforms and any lawmakers purporting to regulate them should ensure that the right to substantive equality and freedom from discrimination is given due weight and prioritized alongside the right to freedom of expression.

Beyond the FBOB, another example of external or quasi-external content moderation is industry collaboration between multiple platforms. Such institutions are not independent from any of the digital platforms themselves, but involve major platform companies coordinating with each other and agreeing to moderate certain types of content in a standardized way among themselves, including through sharing a centralized database of content to block or remove. The primary example of this is the Global Internet Forum to Counter Terrorism (GIFCT), which is discussed further in Section 6.2.3 ("Privatized Regulation of Speech and Public Discourse").

3.4. Critiques of Platform Approaches to Speech-Based TFGBV

While specific content moderation measures to address abusive speech each have their own respective deficiencies as discussed above, some issues cut across platforms' approaches to moderating abusive expression as a whole. This section will address each of those issues in turn: the inconsistent and

---

[389] Elena Debré, "The Independent Facebook Oversight Board Has Made Its First Rulings", Slate (28 January 2021), online: <https://slate.com/technology/2021/01/facebook-oversight-boards-content-moderation-rulings.html>.

[390] Sejal Parmar, "Facebook's Oversight Board: A Meaningful Turn Towards International Human Rights Standards?" (20 May 2020), online: *Just Security* <https://www.justsecurity.org/70234/facebooks-oversight-board-a-meaningful-turn-towards-international-human-rights-standards/>.

[391] Adi Robertson, "Facebook Oversight Board overturns hate speech and pandemic misinformation takedowns" (28 January 2021), online: *Verge* <https://www.theverge.com/2021/1/28/22254155/facebook-oversight-board-first-rulings-coronavirus-misinformation-hate-speech>.

hypocritical use of 'free speech' or 'freedom of expression' as a rhetorical shield for inaction; the overly reactive and selective approach to banning and suspending the posts and accounts of certain users or groups for hateful or harmful speech; and platforms' inability and lack of will to truly quell the problem of abusive speech so long as they continue to prioritize business growth and appeasing political power.

3.4.1. Inconsistent and Unprincipled: "Free Speech" Rhetoric

Many have written about the influence of the United States' legal and social norms and forces around the First Amendment, and freedom of expression as a cultural concept more broadly, on digital platforms' approaches to content moderation, particularly where abusive speech is concerned.[392] Dragiewicz et al. write, "[T]he cultures of US social media companies like Facebook and Twitter are deeply entangled with American ideals of freedom of expression, openness, and the free market [...]—values that in practice often work to favour the prerogatives of privileged groups over the rights of others."[393] In the context of TFGBV specifically,

> Policies that govern user behaviour on social media platforms betray irresolvable conflicts between Silicon Valley's libertarian ideals and the challenges of inclusion and safety. For example, controversial humour is generally protected in most social media sites' policies, which facilitates the disguise of online abuse by means of sexist or racist jokes [...]. Often, misogynistic humour is mediated through visual content, such as memes that, due their catchy aesthetics and their potential to go viral, can be a "ceaseless flickering hum of low-level emotional violence" [...]. Women's grievances with regards to online abuse have sometimes been contested by evoking libertarian principles of freedom of expression that frame any form of intervention as "censorship" [...], and blaming victims by emphasising their personal "responsibility" for the harms that befall them online [...].[394]

According to Franks, "free speech rhetoric has for years been employed to justify these companies' laissez-faire approach to controversial content, from terrorist training videos to [NCDII]. These companies commonly invoke familiar First Amendment tropes to present their passivity as principled neutrality".[395] This purported neutrality is already a fallacy to begin with, in the context of substantive equality[396] and a more complete conceptualization of freedom of expression as established in Canadian law.[397] However, digital platforms have consistently not maintained any such neutrality in any case, even on their own terms.

---

[392] See generally Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harvard Law Review 1598.

[393] Molly Dragiewicz et al, "Technology facilitated coercive control: domestic violence and the competing roles of digital media platforms" (2018) 18:4 Feminist Media Studies 609 at 617 (in-text citations omitted).

[394] *Ibid*.

[395] Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (21 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment>.

[396] See Section 6.1.2 ("Right to Equality Must Inform Proportionality Analysis").

[397] See Section 6.1.3 ("TFGBV Is Low-Value Expression Far from the Core of Section 2(b)") and Section 6.1.4.1 ("Platform Dynamics and a Dysfunctional 'Marketplace of Ideas'").

Exceptions to their alleged principles abound, not least of which has been online platforms' "affirmative and proactive steps to identify and remove harmful content that is surfacing in response to the current COVID-19 pandemic".[398] For example, Facebook has looked to "immunologists, doctors, and the medical establishment" to inform their content moderation policies around pandemic-related posts; incorporated awareness of pandemic-related racism into its content moderator hate speech guidelines; and shown fact-checking messages to users "who have interacted with […] harmful misinformation about COVID-19 that was later removed as false […] us[ing] these messages to connect people to the WHO's COVID-19 mythbuster website that has authoritative information."[399] Facebook has been similarly active in addressing anti-vaccination content on its own platform and on Instagram, limiting the reach of such posts and ensuring search results "prioritize information and links from reputable sources like the World Health Organization".[400] Twitter has acted similarly to quell harmful speech related to COVID-19.[401] As a former Facebook employee stated, "Facebook would be looking for—what is the medical consensus, not what is the political consensus".[402]

The above proactive harm reduction measures beg the question of why platforms cannot or should not do the same when it comes to dangerous content that clearly poses a danger to women's health and well-being. For example, platforms could proactively apply removal, reach reduction, fact-checking, and user redirection measures to misinformation around sexual violence, assault, harassment, and trauma; intimate partner and dating violence; abortion; and sexual health and education, which are also matters of science and medicine with equally authoritative sources to link to, including the WHO.[403]

---

[398] "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 52, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>.

[399] Alex Kantrowitz, "Facebook Is Taking Down Posts That Cause Imminent Harm - But Not Posts That Cause Inevitable Harm" (23 May 2020), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/alexkantrowitz/facebook-coronavirus-misinformation-takedowns?bfsource=relatedmanual>; "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 53, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>.

[400] Dax D'Orazio, "Freedom of Expression, Misinformation, and Anti-Vaxxers: The Right Thing to Do Is Not Obvious" (25 March 2020), online (blog): *Centre for Free Expression* <https://cfe.ryerson.ca/blog/2020/03/freedom-expression-misinformation-and-anti-vaxxers-right-thing-do-not-obvious>; Yana Tatevosian, "Facebook Will Crack Down on Anti-Vaccine Content" (7 March 2019), online: *Wired* <https://www.wired.com/story/facebook-anti-vaccine-crack-down/>.

[401] "Amid the coronavirus outbreak, Twitter has become more aggressive about combating misinformation. At the end of March, Twitter deleted two tweets by Brazilian President Jair Bolsonaro because they contained false or misleading information about COVID-19, the disease caused by the novel coronavirus. The platform also cited its COVID-19 content policy to delete a tweet from Rudy Giuliani, Trump's lawyer, which quoted Talking Points USA's Charlie Kirk and claimed the use of hydroxychloroquine was '100% effective' in treating COVID-19. It also temporarily locked the account of the right-wing news site Federalist and deleted one of its tweets that promoted 'controlled voluntary infection' of COVID-19." Alex Kantrowitz & Ryan Broderick, "Twitter Fact-Checked A Trump Tweet For The First Time" (26 May 2020), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/alexkantrowitz/twiter-fact-checked-trump>.

[402] Alex Kantrowitz, "Facebook Is Taking Down Posts That Cause Imminent Harm - But Not Posts That Cause Inevitable Harm" (23 May 2020), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/alexkantrowitz/facebook-coronavirus-misinformation-takedowns?bfsource=relatedmanual>.

[403] See e.g., "Violence against women" (9 March 2021), online: *World Health Organization* <https://www.who.int/news-room/fact-sheets/detail/violence-against-women>; "Abortion", online: *World Health Organization* <https://www.who.int/health-topics/abortion#tab=tab_1>; "Sexual Violence", online: *World Health Organization* <https://www.who.int/reproductivehealth/topics/violence/sexual_violence/en/>; "International technical guidance on sexuality education: An evidence-informed approach" (14 March 2018), online: *World Health Organization* <https://www.who.int/reproductivehealth/publications/technical-guidance-sexuality-education/en/>.

3.4.2. Reactive, Arbitrary, and Selective: Damage-Control Approach

While platforms' community standards and enforcement policies may have evolved over time in attempts to establish a higher or more equitable standard of inclusivity and equality, this did not occur without years of tremendous effort on the part of feminists,[404] anti-racists, and other activists and advocates working tirelessly on issue after issue,[405] combined with multiple waves of media scrutiny, public outrage, and political pressure. Several high-profile bans, suspensions, or policy changes also only occurred on some platforms after many years of permitting such content to stand and proliferate unabated, even in the face of user complaints and reports, or only after all of their peers and competitors had already acted and left the remaining hold-out platform standing alone.

Examples of content moderation decisions that match one or more of the above patterns of years of committed inaction followed by eventual forced action or pressure-induced reaction include:

- Facebook's modification of its policy against breastfeeding photos[406] and, in 2019, ban of "far-right political commentator Faith Goldy [and] various extremist groups" such as "Canadian Nationalist Front, Aryan Strikeforce, Wolves of Odin and Soldiers of Odin (also known as Canadian Infidels)", including affiliated content, pages, and groups;[407]

---

[404] *(Content Warning)* See e.g., "The climbdown followed a week-long campaign by Women, Action and the Media, the Everyday Sexism Project and the activist Soraya Chemaly to remove supposedly humorous content endorsing rape and domestic violence. Examples included a photograph of the singer Rihanna's bloodied and beaten face, captioned with 'Chris Brown's Greatest Hits', a reference to the assault by her ex-boyfriend. A photograph of a woman in a pool of blood had the caption 'I like her for her brains'. Another photograph, of a man holding a rag over a woman's mouth, was captioned 'Does this smell like chloroform to you?'. More than 100 advocacy groups joined the protest and demanded Facebook recognise such content as hate speech and train moderators to remove it. Facebook, which is based in Menlo Park, California, initially rebuffed the complaints, citing freedom of speech. ... The campaign gathered momentum, however, when tens of thousands of tweets and emails using the hashtag #Fbrape were sent to the social network's advertisers." Rory Carroll, "Facebook gives way to campaign against hate speech on its pages", *Guardian* (29 May 2013), online: <https://www.theguardian.com/technology/2013/may/29/facebook-campaign-violence-against-women>.

[405] See e.g., Amanda Marcotte, "Can These Feminists Fix Twitter's Harassment Problem?" (7 November 2014), online: *Slate* <https://slate.com/human-interest/2014/11/women-action-media-and-twitter-team-up-to-fight-sexist-harassment-online.html>; Rory Carroll, "Facebook gives way to campaign against hate speech on its pages", *Guardian* (29 May 2013), online: <https://www.theguardian.com/technology/2013/may/29/facebook-campaign-violence-against-women>; Sheila Dang, "Exclusive: Facebook ad boycott campaign to go global, organizers say" (28 June 2020), online: *Reuters* <https://www.reuters.com/article/us-facebook-ads-boycott-exclusive/exclusive-facebook-ad-boycott-campaign-to-go-global-organizers-say-idUSKBN23Z0O>.4>.

[406] Rachel Moss, "Facebook Clarifies Nudity Policy: Breastfeeding Photos Are Allowed (As Long As You Can't See Any Nipples)" (16 March 2015), online: *Huffington Post* <https://www.huffingtonpost.co.uk/2015/03/16/breastfeeding-facebook-nudity-policy_n_6877208.html>.

[407] Kathleen Harris, "Facebook bans Faith Goldy and 'dangerous' alt-right groups", *CBC News* (8 April 2019), online: <https://www.cbc.ca/news/politics/facebook-faith-goldy-ban-alt-right-1.5088827>.

- Reddit's quarantine[408] of and then, later, bans of various misogynistic and racist subreddits such as r/Incels, including their July 2020 shut-down of "2,000 subreddits under new rules that ban certain violent and hateful content";[409]
- the gaming livestreaming platform Twitch directly, albeit temporarily, banning the 45th US president for "hateful conduct";[410]
- YouTube's banning of a number of prominent far-right and white supremacist speakers and public figures for hate speech, including Stefan Molyneux, David Duke, and Richard Spencer[411] (while Twitter hesitated[412]);
- Twitter eventually banning some of the same individuals,[413] and taking down thousands of accounts associated with the conspiracy theory network QAnon;[414] and
- major social media platforms banning or suspending Alex Jones to varying degrees.[415]

---

[408] "Around the same time that it banned nonconsensual pornography, Reddit began to 'quarantine' some of the site's most controversial subreddits and to ban others outright. When a subreddit is banned, it is deleted altogether from the site; when a subreddit is quarantined it is still accessible but is flagged with a warning prompt and cannot host ads." Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (21 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment>.

[409] Jillian C York, "Reddit banned a pro-Trump subreddit. Here's what that means for hate speech." (2 July 2020), online: *NBC News* <https://www.nbcnews.com/think/opinion/reddit-banned-pro-trump-subreddit-here-s-what-means-hate-ncna1232797>; see also Olivia Solon, "'Incel': Reddit bans misogynist men's group blaming women for their celibacy", *Guardian* (8 November 2017), online: <https://www.theguardian.com/technology/2017/nov/08/reddit-incel-involuntary-celibate-men-ban>; and Adi Robertson, "Reddit has broadened its anti-harassment rules and banned a major incel forum" (30 September 2019), online: *Verge* <https://www.theverge.com/2019/9/30/20891920/reddit-harassment-bullying-threats-new-policy-change-rules-subreddits>.

[410] Kellen Browning, "Twitch suspends Trump's channel for 'hateful conduct'", *Seattle Times* (29 June 2020), online: <https://www.seattletimes.com/business/twitch-suspends-trumps-channel-for-hateful-conduct/>; Jacob Kastrenakes, "Twitch temporarily bans President Trump" (29 June 2020), online: *Verge* <https://www.theverge.com/2020/6/29/21307145/twitch-donald-trump-ban-campaign-account>.

[411] Julia Alexander, "YouTube bans Stefan Molyneux, David Duke, Richard Spencer, and more for hate speech" (29 June 2020), online: *Verge* <https://www.theverge.com/2020/6/29/21307303/youtube-bans-molyneux-duke-richard-spencer-conduct-hate-speech>.

[412] Matt Novak, "Twitter Defends Giving David Duke a Platform: He's 'Not Currently a Member of the KKK'" (11 July 2020), online: *Gizmodo* <https://www.gizmodo.com.au/2020/07/twitter-defends-giving-david-duke-a-platform-hes-not-currently-a-member-of-the-kkk/>; Lois Beckett, "Twitter bans white supremacist David Duke after 11 years", *Guardian* (31 July 2020), online: <https://www.theguardian.com/technology/2020/jul/31/david-duke-twitter-ban-white-supremacist>.

[413] Oliver Darcy, "A Twitter spokesperson tells me Molyneux's account 'was suspended for spam and platform manipulation, specifically operating fake accounts.'" (7 July 2020), online: *Twitter* <https://twitter.com/oliverdarcy/status/1280679072444678150>.

[414] Ben Collins & Brandy Zadrozny, "Twitter bans 7,000 QAnon accounts, limits 150,000 others as part of broad crackdown" (21 July 2020), online: *NBC News* <https://nbcnews.com/news/amp/ncna1234541>. ("QAnon is a right-wing conspiracy theory that centers on the baseless belief that an anonymous tipster is revealing how President Donald Trump is leading a secret war against a so-called deep state — a collection of political, business and Hollywood elites who, according to the theory, worship Satan and abuse and murder children. The conspiracy theory's roots grew from Pizzagate, which claimed that Hillary Clinton ran a pedophilia ring from a Washington, D.C., pizza shop.")

[415] "Beginning in July 2018, several major online platforms began removing content produced by Alex Jones, a high-profile, far-right radio show host and creator of the conspiracy theorist website Infowars. Jones is notorious for claiming, among

Most if not all of these bans, suspensions, and policy changes have been met with justified skepticism and cynicism with respect to their timing and motivation, given that "many of the policy and enforcement decisions made by architects inside Facebook [and other companies] came about in reaction to external pressures from civil society, governments, the media, or users. Of these, negative media coverage had arguably the most powerful impact."[416] Journalists and others question "why the company waited until it became the subject of media reports and criticism from lawmakers to finally act"[417] and "what, exactly, prompted the decisions? After all, people have been complaining about this kind of speech for years. When it comes to Facebook, the answer is obvious: An advertiser boycott caused the company a $7 billion loss […]."[418] After major platform companies "decided, under pressure, to enforce existing 'acceptable use' policies and take action against groups that participated in the [deadly Unite the Right] rally [in Charlottesville, Virginia]", the Southern Poverty Law Centre stated that "it took 'blood in the streets for tech companies to take action'".[419]

Others have questioned the concerning degree of unilateral discretion that platforms display in making such decisions, finding the process problematic even if a specific ban is otherwise supported. As NDP MP Nathan Cullen has stated, high-profile bans and suspensions each represent "a 'one-off' that targeted particularly hateful and high profile groups and individuals. 'We think that's progress, but it's inconsistent. Without any set of guidelines or rules, then we're allowing self-regulation.'"[420]

To the extent that digital platforms have taken action against abusive speech and its most prominent purveyors, such efforts have been as discomfiting as much as they have been encouraging. They reflect not so much platforms' commitment to addressing abusive speech on a systemic level, but rather more so reflect their sensitivity to their respective reputations among users and—because it impacts their bottom lines—advertisers, and to the threat of regulation by legislators or other government intervention in their affairs.

---

other things, that the Sandy Hook shooting did not take place and for promoting the 'PizzaGate' conspiracy theory. […] For several weeks, the only major online platform not to take action against Jones was Twitter. On September 6, 2018, Twitter announced that it was permanently suspending Jones and the Infowars account." Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (21 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment> (footnotes omitted).

[416] Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129 Yale Law Journal 2418 at 2439 (footnotes omitted).

[417] Louise Matsakis, "Facebook Will Crack Down on Anti-Vaccine Content" (7 March 2019), online: *Wired* <https://www.wired.com/story/facebook-anti-vaccine-crack-down/>.

[418] Jillian C York, "Reddit banned a pro-Trump subreddit. Here's what that means for hate speech", *NBC News* (2 July 2020), online: <https://www.nbcnews.com/think/opinion/reddit-banned-pro-trump-subreddit-here-s-what-means-hate-ncna1232797>.

[419] Aaron Winter, "'Online Hate: From the Far-Right to the 'Alt-Right', and from the Margins to the Mainstream'" in Karen Lumsden & Emily Harmer, eds, *Online Othering: Exploring Digital Violence and Discrimination on the Web* (London: Palgrave Macmillan, 2019) 39 at 56. At the same time, "they argue that still 'some of the biggest tech companies keep hate group sites up and running', noting that Paypal, Bitcoin, Stripe, Network Solutions and others continue to provide services to designated hate groups (Southern Poverty Law Center 2018)."

[420] . Kathleen Harris, "Facebook bans Faith Goldy and 'dangerous' alt-right groups", *CBC News* (8 April 2019), online: <https://www.cbc.ca/news/politics/facebook-faith-goldy-ban-alt-right-1.5088827>.

3.4.3. Conflicting Incentives: Business Priorities and Political Influence

Platform companies have demonstrated undue business and political sensitivity through multiple actions, statements, and programs revealed through media reports, prioritizing financial growth and political power or appeasement at the expense of truly addressing online abuse and harassment.[421] Specifically, YouTube and Facebook have quashed, on multiple occasions, internal research implicating their algorithms in exacerbating online abuse, rejected or undermined employees' efforts to remedy the problem, and even rebuked employees for embarking on such initiatives and ordered them not to continue such lines of research or projects to address online bias and abuse.[422] In an in-depth profile of Facebook's 'Responsible AI' team, Karen Hao concluded that if the team ever attempted to "make headway against misinformation and hate speech" in earnest, it would be "set up for failure":

> Everything the company does and chooses not to do flows from a single motivation: Zuckerberg's relentless desire for growth. [The Responsible AI] team got pigeonholed into targeting AI bias [...] because preventing such bias helps the company avoid proposed regulation that might, if passed, hamper that growth. Facebook leadership has also repeatedly weakened or halted many initiatives meant to clean up misinformation on the platform because doing so would undermine that growth.
>
> In other words, the Responsible AI team's work—whatever its merits on the specific problem of tackling AI bias—is essentially irrelevant to fixing the bigger problems of misinformation, extremism, and political polarization. And it's all of us who pay the price.[423]

Current and former employees at digital platform companies, in addition to experts, civil society, and outside observers, have noted that political power unduly influences management's and executives' steadfast reluctance or outright opposition to addressing the problem of abusive speech at its roots, particularly hateful and harmful speech attacking women, gender and sexual orientation minorities, and members of racialized communities.[424] According to Yael Eisenstat, former election ads integrity

---

[421] See e.g., "There is a wide assumption, not unfounded, that Facebook has a financial stake in leaving total garbage up on its site. Reams of evidence, anecdotal and scholarly, suggest that its News Feed algorithm rewards inflammatory and addictive content." Simon van Zuylen-Wood, "'Men Are Scum': Inside Facebook's War on Hate Speech", *Vanity Fair* (March 2019), online: <https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech>.

[422] See e.g., Mark Bergen, "YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant" (2 April 2019), online: *Bloomberg* <https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant>; Jeff Horwitz & Deepa Seetharaman, "Facebook Executives Shut Down Efforts to Make the Site Less Divisive", *The Wall Street Journal* (26 May 2020), online: <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>; and "At around the same time as the Instagram episode, several pieces of research exploring race and racial bias on Facebook and Instagram were summarized and presented in a document to Zuckerberg and his inner circle, known as the M-Team. The team responded by instructing employees to stop all research on race and ethnicity and not to share any of their findings with others in the company, according to two current and one ex-employee." Olivia Solon, "Facebook ignored racial bias research, employees say" (23 July 2020), online: *NBC News* <https://www.nbcnews.com/tech/tech-news/facebook-management-ignored-internal-research-showing-racial-bias-current-former-n1234746>.

[423] Karen Hao, "How Facebook got addicted to spreading misinformation", *MIT Technology Review* (11 March 2021), online: <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/>.

[424] See e.g., Ryan Mac & Craig Silverman, "'Mark Changed The Rules': How Facebook Went Easy On Alex Jones And Other Right-Wing Figures", BuzzFeed News (22 February 2021), online: <https://www.buzzfeednews.com/article/ryanmac/mark-zuckerberg-joel-kaplan-facebook-alex-jones>; Ryan Mac & Craig Silverman, "'Hurting People at Scale': Facebook's

lead at Facebook, "They put political considerations over enforcing their policies to the letter of the law […] I can say for my time there that more than once the [Washington] policy team weighed in on appeals and decisions that made it clear there was a political consideration factoring into how we were enforcing our policy."[425] Moreover, journalists have reported on Facebook silencing or retaliating against employees attempting to discuss or address the issue internally.[426] For example, a senior engineer was purportedly dismissed or pushed to leave the company after he "collected multiple instances of conservative figures receiving unique help from Facebook employees, including those on the policy team, to remove fact-checks on their content."[427]

Facebook has gone even further and secretly launched a disinformation campaign—with the aid of a Republican political consultancy "specialized in applying political campaign tactics to corporate public relations"—against racial justice advocates to discredit them.[428] All the while, Facebook representatives were meeting with those same groups and individuals ostensibly to address the problem of violent and abusive speech and harassment on its platform.[429]

The documented extent of the US Republican Party's influence over content policies at Facebook is particularly concerning for addressing TFGBV on platforms, given right-wing positions on women's equality and other matters of social justice, generally speaking. It is relevant that Facebook's Washington policy team is led by Joel Kaplan, the company's vice president of global public policy and "a close, personal friend of [Supreme Court of the United States Justice Brett] Kavanaugh's, [who] sat behind the judge during his recent hearings before the Senate judiciary committee" about Christine

---

Employees Reckon with the Social Network They've Built", *BuzzFeed News* (23 July 2020), online: <https://www.buzzfeednews.com/article/ryanmac/facebook-employee-leaks-show-they-feel-betrayed>; Ryan Mac & Craig Silverman, "After The US Election, Key People Are Leaving Facebook And Torching The Company In Departure Notes", *BuzzFeed News* (11 December 2020), online: <https://www.buzzfeednews.com/article/ryanmac/facebook-rules-hate-speech-employees-leaving>.

[425] Ryan Mac and Craig Silverman, "'Hurting People at Scale': Facebook's Employees Reckon With the Social Network They've Built" (23 July 2020), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/ryanmac/facebook-employee-leaks-show-they-feel-betrayed>.

[426] Ryan Mac, "Facebook Forced Its Employees To Stop Discussing Trump's Coup Attempt", BuzzFeed News (6 January 2021), online: <https://www.buzzfeednews.com/article/ryanmac/facebook-trump-coup>.

[427] Craig Silverman & Ryan Mac, "Facebook Fired An Employee Who Collected Evidence Of Right-Wing Pages Getting Preferential Treatment", *BuzzFeed News* (6 August 2020), online: <https://www.buzzfeednews.com/article/craigsilverman/facebook-zuckerberg-what-if-trump-disputes-election-results>.

[428] "While Mr. Zuckerberg has conducted a public apology tour in the last year, Ms. Sandberg has overseen an aggressive lobbying campaign to combat Facebook's critics, shift public anger toward rival companies and ward off damaging regulation. Facebook employed a Republican opposition-research firm to discredit activist protesters, in part by linking them to the liberal financier George Soros. It also tapped its business relationships, lobbying a Jewish civil rights group to cast some criticism of the company as anti-Semitic." Sheera Frankel et al, "Delay, Deny and Deflect: How Facebook's Leaders Fought Through Crisis", *The New York Times* (14 November 2018), online: <https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html>.

[429] Alvaro Bedoya, "The activists at @ColorOfChange criticized Facebook, but they also engaged with them directly and in good faith. Meanwhile, Facebook was busy unleashing the alt-right fever dream." (16 November 2018), online: *Twitter* <https://twitter.com/alvarombedoya/status/1063430080905572353>; Color of Change, ".@facebook 's response to us challenging them 2 create safe conditions 4 Black ppl & marginalized groups on their platform? Fanning the flames of anti-Semitism resulting in a pipe bomb on George Soros' doorstep & campaigning against us using alt right media" (14 November 2018), online: *Twitter* <https://twitter.com/ColorOfChange/status/1062882090310664198>.

Blasey Ford's sexual assault allegations against the judge,[430] and who "played a key role in pushing for" the judge's nomination.[431]

Unfounded and strategic accusations of "anti-conservative bias" by prominent politicians in the United States have also caused digital platforms to drag their feet in meaningfully addressing hate speech and speech-based abuse,[432] a tactic known as "working the refs":[433]

> [C]onservatives are working the refs. If conservatives put media executives on their heels, constantly defending themselves or excusing themselves or apologizing for misunderstandings, then these companies are likely to bend toward conservatives out of fear or just exhaustion.
>
> Working the refs is still effective. Mark Zuckerberg of Facebook and Jack Dorsey of Twitter are not wise enough to understand what's happening. So both Facebook and Twitter have allowed themselves to be worked. Platforms do make some intentional decisions to moderate the content that appears on their websites. But Facebook, Twitter, and Google staff try to do so based on principles and standards that they agonize over. Calls to violence or gender-based harassment should not be considered expressions of political ideology. More often than not, these companies under-filter hate speech because they have such strong concern for free speech. Far from rushing to suspend even conspiracy slingers and hate-mongers such as Alex Jones and Milo Yiannopoulos, executives at Facebook and Twitter hemmed and hawed for years about whether to enforce their own terms of service.[434]

Facebook "has largely bent over backwards to appease Republican complaints. In 2018, it hired a former Republican senator to do an audit of bias on the site. The report accused non-partisan, neutral fact checkers of 'liberal bias,' and resulted in policy changes that allowed for more graphic anti-abortion ads."[435] In June 2020, Facebook also told employees it would not take any action on a "Trump campaign

---

[430] Arwa Mahdawi, "A Facebook executive's Kavanaugh support is a slap in the face to women", *Guardian* (6 October 2018), online: <https://www.theguardian.com/world/2018/oct/06/kavanaugh-joel-kaplan-facebook-women-survivors>.

[431] Tyler Sonnemaker, "A Facebook executive rallied support for Kavanaugh's Supreme Court nomination, a new book says" (22 November 2019), online: *Insider* <https://www.businessinsider.com/facebooks-joel-kaplan-rallied-support-for-kavanaugh-scotus-nomination-book-2019-11>.

[432] "Conservatives have a huge incentive to keep social media companies from moderating untrue or bigoted posts, since the narratives created by Trump allies such as Ben Shapiro and Tucker Carlson spread so effectively online — and helped Trump and Republicans rise to power." Rachel Kraus, "Once again, there is no 'anti-conservative' bias on social media" (28 July 2020), online: *Mashable* <https://mashable.com/article/anti-conservative-bias-facebook/>.

[433] "Experts say another reason conservatives engage in these arguments is to 'work the refs.' That is, if they accuse the people in charge of moderating content of bias loudly enough, moderators might be disinclined to do so again in the future to avoid looking biased. Conservatives have a huge incentive to keep social media companies from moderating untrue or bigoted posts, since the narratives created by Trump allies such as Ben Shapiro and Tucker Carlson spread so effectively online — and helped Trump and Republicans rise to power." *Ibid*.

[434] Siva Vaidhyanathan, "Why Conservatives Allege Big Tech is Muzzling Them", *Atlantic* (28 July 2019), online: <https://www.theatlantic.com/ideas/archive/2019/07/conservatives-pretend-big-tech-biased-against-them/594916/>.

[435] The article continues, "[Facebook] has also appointed an organization affiliated with Tucker Carlson's ultra-right wing website the Daily Caller as a "fact checking" partner, despite the Daily Caller's status as a routine peddler of misinformation. That appointment, and the audit, result in more than just lip service to conservatives: it undermines fact-checkers and the

ad that featured a triangle symbol used by Nazis to identify political prisoners" and that the ad did not violate content policies, despite being flagged by nine people, until "only after receiving questions from the *Washington Post*—more than 12 hours after it had been flagged by employees."[436]

The charges of 'anti-conservative bias' have been widely discredited as lacking any evidence, and if anything, evidence shows the contrary: "In our current media ecosystem, right-wing sources of news and propaganda spread much further and faster than liberal or neutral sources do".[437] If, however, it were the case that right-wing content undergoes more moderation than average on digital platforms, that would be because more "right-leaning ideologies and content overlap with behavior that's not allowed on social networks".[438] That is not "bias" but an accurate reflection of the fact that right-wing ideologies, which often advocate the inferiority of historically marginalized groups, will by nature result in more content that objectively violates community standards against hate speech, for example. To reframe this as 'anti-conservative bias' would be misleading and deeply problematic, as reacting the way digital platforms have would, in effect, shield abusive speech and hate-based ideologies from moderation purely by virtue of someone having adopted them as a political position. This would be tautological, and defeats the very purpose of content moderation in the first place, if the idea is not to allow hate speech to proliferate and spread to the point that it acquires that very prominence and legitimacy of becoming a political platform that can then be enacted through laws and policies with repercussions far beyond platforms themselves.

---

nature of truth and accountability itself." Rachel Kraus, "Once again, there is no 'anti-conservative' bias on social media" (28 July 2020), online: *Mashable* <https://mashable.com/article/anti-conservative-bias-facebook/>.

[436] Ryan Mac and Craig Silverman, "'Hurting People at Scale': Facebook's Employees Reckon With the Social Network They've Built" (23 July 2020), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/ryanmac/facebook-employee-leaks-show-they-feel-betrayed>.

[437] See Siva Vaidhyanathan, "Why Conservatives Allege Big Tech is Muzzling Them", *Atlantic* (28 July 2019), online: <https://www.theatlantic.com/ideas/archive/2019/07/conservatives-pretend-big-tech-biased-against-them/594916/> and Chris Stokel-Walker, "Fake news travels six times faster than the truth on Twitter", *New Scientist* (8 March 2018), online: <https://www.newscientist.com/article/2163226-fake-news-travels-six-times-faster-than-the-truth-on-twitter/>.

[438] Rachel Kraus, "Once again, there is no 'anti-conservative' bias on social media" (28 July 2020), online: *Mashable* <https://mashable.com/article/anti-conservative-bias-facebook/>.

4. Platform Liability for TFGBV in Canadian Law

Platform liability for users' abusive speech and behaviours is a dynamic and developing legal area in Canada, with little law yet established to address platform liability for TFGBV specifically. Civil society, academics, lawyers, human rights advocates, members of impacted communities, and members of Parliament have increasingly called for government action in recent years to regulate digital platforms to address TFGBV and similar forms of online abuse that target other and intersecting historically marginalized groups. [439] However, prior to 2019, the federal government appeared to shy away from regulating digital platform companies with respect to user content that would be harmful to vulnerable groups, instead opting for voluntary commitments, agreements, or partnerships. [440]

The government's sustained passivity and conciliatory approach to platform liability for harmful user content has stood in stark contrast to assertive if not aggressive government stances on platform regulation in two other contexts that rose to the forefront of ministers' concerns instead: copyright law, and industry funding proposals advanced by legacy media companies.[441] The priorities, stakeholder groups, and frameworks of Canadian copyright and media and broadcasting policy have thus shaped many Parliamentary debates around platform liability in Canada historically (and continue to this

---

[439] See e.g., Canada, Parliament, House of Commons, *Taking Action to End Online Hate: Report of the Standing Committee on Justice and Human Rights*, 42nd Parl, 1st Sess (June 2019) (Chair: Anthony Housefather) and Canada, Parliament, House of Commons, Standing Committee on Access to Information, Privacy and Ethics, *Democracy under Threat: Risks and Solutions in the Era of Disinformation and Data Monopoly*, 42nd Parl, 1st Sess (December 2018) (Chair: Bob Zimmer).

[440] See e.g., Jesse Brown, "1.All over the world, governments are cracking down on Facebook. Many countries are getting serious about privacy intrusion, 'fake news', media monopolization, tax avoidance…" (8 January 2018), online: *Twitter* <https://threadreaderapp.com/thread/950395900593082369.html>; "How Facebook Bought-Off Canada For Peanuts" (8 January 2018), online (podcast): *Canadaland* <https://www.canadalandshow.com/podcast/facebook-bought-off-canada-peanuts/>; Carl Meyer, "Carleton's new election-integrity scholar comes from Facebook. The NDP says that's like Dracula overseeing the blood supply" (27 January 2020), online: *National Observer* <https://www.nationalobserver.com/2020/01/27/news/carletons-new-election-integrity-scholar-comes-facebook-ndp-says-thats-dracula>; Murad Hemmadi, "What the NDP's ethics critic is not liking about Facebook's role in Ottawa", *Maclean's* (19 April 2018), online: <https://www.macleans.ca/politics/ottawa/what-the-ndps-ethics-critic-is-not-liking-about-facebooks-role-in-ottawa/>; Reuters Staff, "Facebook to launch election integrity effort in Canada" (14 September 2017), online: *Reuters* <https://www.reuters.com/article/us-facebook-canada-election/facebook-to-launch-election-integrity-effort-in-canada-idUSKCN1BP2ZT>.

[441] See e.g., Anja Karadeglija, "Canada is considering a proposal that would allow internet providers to block piracy websites", National Post (14 April 2021), online: <https://nationalpost.com/news/canada/federal-authorities-considering-proposal-to-allow-internet-providers-to-block-piracy-websites>; Amanda Connolly, "Netflix, other streaming services should be forced to create CanCon, pay digital tax: panel", Global News (29 January 2020), online: <https://globalnews.ca/news/6477105/broadcasting-review-panel-net-neutrality-netflix-tax/>; David Ljunggren, "Canada vows to be next country to go after Facebook to pay for news", Reuters (18 February 2021), online: <https://www.reuters.com/business/media-telecom/canada-vows-be-next-country-go-after-facebook-pay-news-2021-02-18/>; Alex Boutilier, "Internet giants should support local news, culture, Melanie Joly says", Toronto Star (14 March 2018), online: <https://www.thestar.com/news/canada/2018/03/14/internet-giants-should-support-local-news-culture-melanie-joly-says.html>; and Daniel Leblanc, "Ottawa warns internet platforms such as Facebook and Netflix the 'free ride' is over", Globe and Mail (5 June 2018), online: <https://www.theglobeandmail.com/politics/article-ottawa-warns-internet-platforms-such-as-facebook-and-netflix-the-free/>.

day),[442] tracing back to the original and overlapping, though not precisely interchangeable, concept of intermediary liability.[443] Legislators and policy-makers working on platform liability for TFGBV must thus take care in applying or extending these pre-existing discussions and their embedded policy considerations, and instead ensure that platform liability in the TFGBV context centers substantive equality and freedom from discrimination as applied through an intersectional feminist lens.

There are a number of Canadian laws that could theoretically be used to establish platform liability for TFGBV by a platform's user, given the right circumstances, but many of these have yet to be tested in court in this context. Canada has laws in force that do the following:

- laws that establish a general intermediary liability regime (e.g., section 22 of *Act to establish a legal framework for information technology* in Quebec[444]);
- laws that establish platform liability or legal obligations for non-TFGBV user content (e.g., direct liability for 'enabling' copyright infringement and the notice-and-notice regime for copyright infringement under the *Copyright Act*,[445] or what is effectively a notice-and-takedown regime that the courts have developed in defamation law);
- laws that address TFGBV in some form but are silent on the role of platforms (e.g., *Criminal Code* offences for NCDII and hate propaganda[446]); or
- laws of general application that address neither TFGBV nor intermediary liability specifically, but could apply to platform companies as organizations, provided the factual circumstances met the relevant legal test (e.g., criminal corporate negligence, product liability, or statutory human rights law).

There thus appears to be a gap in Canadian law to the extent of establishing *platform liability for TFGBV, specifically.* Under current laws, a platform cannot be held liable for TFGBV by a user, and has no legal obligation to act, unless, extrapolating from the categories laid out above:

- the user's post meets the legal definition of defamation or copyright infringement (in which latter case, the person targeted by TFGBV has only the extremely limited recourse of the forwarded notice, and the law would in some sense be addressing the wrong "mischief");
- the TFGBV committed by a user constitutes "illicit activity" in Quebec;

---

[442] See e.g., Canada, Parliament, House of Commons, Standing Committee on Industry, Science and Technology, *Statutory Review of the Copyright Act*, 42nd Parl, 1st Sess (June 2019) (Chair: Dan Ruimy) at 18-20, 74-83, 91-98; Gregory R Hagen, "'Modernizing' ISP Copyright Liability" in Michael Geist, ed, *"From Radical Extremism to Balanced Copyright": Canadian Copyright and the Digital Agenda* (Irwin Law, 2010) 361; Canada, Parliament, House of Commons, Standing Committee on Canadian Heritage, *Shifting Paradigms*, 42nd Parl, 1st Sess (May 2019) (Chair: Julie Dabrusin); and Dr Carys J Craig, "Meanwhile, in Canada… A Surprisingly Sensible Copyright Review", European Intellectual Property Review [forthcoming in 2021], at 1-5 (including comments on afore-cited Parliamentary reports).

[443] See Issue Spotlight No. 1 ("Copyright, Intermediary Liability, and Safeguarding Human Rights in Context").

[444] *Act to establish a legal framework for information technology*, CQLR c C-1.1, s 22.

[445] *Copyright Act*, RSC, 1985, c C-42, s 41.26.

[446] *Criminal Code*, RSC, 1985, c C-46, ss 162.1 and 320.1.

- the platform company can be shown to have been so involved that it meets the bar for "enabling", in copyright alone, or for direct liability, such as being party to a criminal offence that constitutes TFGBV (in which case, the platform is effectively no longer an intermediary); or
- a case can be made out against a platform company using a law of general application that did not specifically contemplate either TFGBV or platform liability.

The remainder of this Part 4 will provide an overview of the current Canadian legal landscape of platform liability for TFGBV—encompassing all of the laws mentioned above—organized as follows. Section 4.1 begins with an introduction to intermediary liability principles in Canadian law, which provide context and set the groundwork for assessing issues of platform liability for user wrongdoing. Section 4.2 will examine relevant federal legislation currently in force, namely parts of the *Copyright Act*, certain *Criminal Code* provisions, and, though it is not precisely federal legislation, the *Canada-United States-Mexico Agreement* (CUSMA). Section 4.3 will present potential forthcoming federal legislation (at time of writing, in early April 2021) and related Parliamentary studies on TFGBV. Section 4.4 will canvass relevant provincial legislation, specifically the intermediary liability regime in Quebec and several provincial NCDII statutes. Section 4.5 will discuss uniform and model legislation that has been adopted or proposed on a provincial level, and which represents some of the most substantive legislative work done to date in Canada that examines and sets out to resolve the issue of platform liability for abusive speech and actions by a platform's users. Section 4.6 will provide a brief overview of laws that fall into the fourth category above: those which do not explicitly contemplate either intermediary liability or TFGBV, but could be used to hold platform companies liable at an institutional and systemic level, based on pre-existing causes of action such as product liability, corporate or criminal negligence, or violation of statutory human rights law requiring non-discrimination in providing goods, services, and facilities.

4.1. Intermediary Liability Principles in Canadian Law

Digital platforms fall within the broader category of Internet intermediaries, which are entities that "bring together or facilitate transactions between third parties on the Internet".[447] Before digital platforms and their associated legal issues rose to the prominence they have today, the term Internet intermediary was commonly understood to include—if not primarily refer to—entities constituting some of the basic infrastructural layers of the Internet, such as Internet backbone providers, Internet service providers (ISPs), telecommunications companies, web hosting providers, and domain name registrars,[448] in addition to application-layer platforms such as search engines and social media websites. A formidable body of law, policy, and academic literature concerning the legal liability of such Internet intermediaries across various areas of law has thus been long established, under the umbrella term of "intermediary liability", and continues to evolve today. This background context of intermediary liability law thus informs and to some extent legally governs approaches to platform liability. After all, digital platforms are a type of Internet intermediary, even though some types—e.g.,

[447] Karine Perset, "The Economic and Social Role of Internet Intermediaries" (April 2010) at 9, online (pdf): *OECD* <https://oecd.org/internet/ieconomy/44949023.pdf>.

[448] *Ibid*.

social media platforms—have clearly departed from the quintessential notion of intermediaries as passive "mere conduits".[449]

The term "intermediary liability" has in the past, somewhat counterintuitively, served as a stand-in term to refer to the principle that online intermediaries *did not have* liability for their users' wrongdoing, reflecting what was often the case under relevant laws. As digital platforms and their role and impact in society have evolved, however, alongside public opinion and the law concerning such platforms, the full meaning of the term "intermediary liability" has become ever central: determining whether to attribute liability to a given intermediary for its users' activities, and if so, to what extent, on what basis, and under what circumstances.[450]

The measured, incremental approach that generally characterizes Canadian law is reflected in some laws currently in place that may be categorized as platform liability law, by virtue of imposing legal obligations on digital platforms, but in fact developed under the banner of other legal fields, such as copyright or defamation. Each case thus reflects the particular legislative, jurisprudential, or interpretive context of the governing statute or area of law as much as it does the law of intermediary liability that has emerged from them. Certain considerations may not be appropriate to transpose into the context of TFGBV, or may warrant a different balancing of certain scales. However, taken together, these cases present a collection of principles that the courts have relied on when it comes to determining intermediary liability for various kinds of user wrongdoing, which may serve as initial guidelines that could be applied with necessary modifications to assess platform liability for TFGBV.

As a starting point, the Supreme Court of Canada established, in *Society of Composers, Authors & Music Publishers of Canada v Canadian Assn. of Internet Providers* ("*CAIP*"), that Internet service providers are not liable for acts of copyright infringement by their subscribers.[451] This finding, as well as subsection 2.4(1) of the *Copyright Act* on which the finding is based, is rooted in Internet service providers' role as common carriers, a foundational concept in telecommunications law and policy.[452]

---

[449] "For more than a decade, social media platforms have portrayed themselves as mere conduits, obscuring and disavowing their active role in content moderation. When they acknowledge moderation at all, platforms generally frame themselves as open, impartial, and noninterventionist—in part because their founders fundamentally believe them to be so and in part to avoid obligation or liability. Their instincts have been to dodge, dissemble, or deny every time it becomes clear that, in fact, they powerfully shape and censor public discourse." Tarleton Gillespie, "Platforms are not Intermediaries" (2018) 2 Georgetown Law Technology Review 198 at 199.

[450] The term "platform liability" applies that same meaning to digital platforms specifically, as opposed to the full range of all possible Internet intermediaries such as those listed above.

[451] "Section 2.4(1)(b) shields from liability the activities associated with providing the means for another to communicate by telecommunication. "The 'means'", as the Board found, '. . . are not limited to routers and other hardware. They include all software connection equipment, connectivity services, hosting and other facilities and services without which such communications would not occur' (p. 452). I agree. So long as an Internet intermediary does not itself engage in acts that relate to the content of the communication, i.e., whose participation is content neutral, but confines itself to providing "a conduit" for information communicated by others, then it will fall within s. 2.4(1)(b)." *Society of Composers, Authors and Music Publishers of Canada v Canadian Assn of Internet Providers*, 2004 SCC 45 at para 92.

[452] "The common law notion of common carriage is central to the understanding of network discrimination issues. [T]raditional common carriers included coachmen, ferrymen and similar professions engaged in the transportation of people or merchandise. The concept was soon extended to railways and later came to include modern telecommunication systems like telegraph and telephone. In essence, common carriers are private companies which, due to their central role in transportation or telecommunications, are vested with some public duties. The traditional obligations of these companies are to offer reasonable rates to all customers, to ensure interconnection between their network and those of competitors and, crucially, to ensure a non-discriminatory treatment of passengers or merchandise transported over their network." Alex

In setting out the basis for this finding, the Supreme Court of Canada (SCC) suggests potential considerations that may favour a finding of intermediary liability and emphasizes that the analytic focus should be on *function* rather than category of entity:

> I conclude that the *Copyright Act*, as a matter of legislative policy established by Parliament, does not impose liability for infringement on intermediaries who supply software and hardware to facilitate use of the Internet. The attributes of such a "conduit", as found by the [Copyright] Board, include a lack of actual knowledge of the infringing contents, and the impracticality (both technical and economic) of monitoring the vast amount of material moving through the Internet, which is prodigious. …
>
> Of course an Internet Service Provider in Canada can play a number of roles. In addition to its function as an intermediary, it may as well act as a content provider, or create embedded links which automatically precipitate a telecommunication of copyrighted music from another source. In such cases, copyright liability may attach to the added functions. The protection provided by s. 2.4(1)(b) relates to a protected function, not to all of the activities of a particular Internet Service Provider. [453]

In the defamation context,[454] the SCC in *Crookes v Wikimedia Foundation Inc* (also known as *Crookes v Newton*) demonstrated its sensitivity to the potential repercussions of attaching legal liability to central building blocks of the Internet—in this case, the hyperlink. Although a hyperlink is not itself an online intermediary or platform, it is a key constituting element of them and serves a similar function of quickly and accessibly connecting users to information and to each other.[455] It is thus worth noting the following statement from the SCC's decision, which declined to apply the traditional legal approach in defamation concerning the test for what constitutes "publication", to the context of online communications:

> The Internet cannot, in short, provide access to information without hyperlinks. Limiting their usefulness by subjecting them to the traditional publication rule would have the effect of seriously restricting the flow of information and, as a result, freedom of expression. The potential "chill" in how the Internet functions could be devastating, since primary article authors would unlikely want to risk liability for linking to another article over whose changeable content they have no control. Given the core significance of the role of hyperlinking to the Internet, we risk impairing its whole functioning. Strict

---

Guindon & Danielle Dennie, "Net Neutrality in Canada and what it means for libraries" (2010) 5:1 Partnership: the Canadian Journal of Library and Information Practice and Research 1 at 7.

[453] *Society of Composers, Authors and Music Publishers of Canada v Canadian Assn of Internet Providers*, 2004 SCC 45 at paras 101-102.

[454] For a more detailed discussion of intermediary liability in the context of defamation law, or how Canadian law has approached platform liability for defamatory content by users, see Emily B Laidlaw & Hilary Young, "Internet Intermediary Liability in Defamation" (2018) 56:1 Osgoode Hall Law Journal.

[455] "While a hyperlinker is not an intermediary, she shares essential characteristics with most intermediaries, in that both play primarily facilitative roles. The intermediary provides access to content created by others, while the hyperlinker merely draws [the] reader's attention to that content." Tamir Israel, "Crookes v. Newton: Speculations on Intermediary Liability…." (2 November 2011), online: *Slaw* <http://www.slaw.ca/2011/11/02/crookes-v-newton-speculations-on-intermediary-liability/>.

> application of the publication rule in these circumstances would be like trying to fit a square archaic peg into the hexagonal hole of modernity.[456]

Writing for the majority, Abella J. adds that liability would apply where a person who links to defamatory content does not *only* link to the content, but in effect repeats the defamatory meaning itself through their surrounding words and context.[457] Thus, liability in this case still requires involvement in the wrongdoing to an extent that amounts to having exited the role of intermediary—or in this case, hyperlinker—into becoming a contributing participant. The SCC considered that "[s]uch an approach promotes expression and respects the realities of the Internet, while creating little or no limitations to a plaintiff's ability to vindicate his or her reputation".[458]

Another line of defamation law that has particular significance for intermediary liability as applied to online platforms and TFGBV is publication by omission.[459] Under this doctrine, liability can apply to a person—or an entity such as a platform company—if they control the venue where defamatory content was posted, even if they did not post the content themselves. Liability arises if the controller of the venue has been notified of the problematic content, has refused to remove the content, and "the refusal can be interpreted as endorsing it".[460] Emily Laidlaw and Hilary Young observe that generally speaking, "platforms have been treated as non-publishers (passive instruments) until notice, and then publishers by omission if they fail to remove content after notice."[461] However, the authors suggest that "[t]he vast amount of expression hosted, the nature of the intermediary, and the existence of terms of service may all militate against the conclusion that intermediaries endorse expression they fail to remove. As such, courts are, in our view, wrong to draw an inference of endorsement from a mere failure to remove."[462]

In *Google Inc v Equustek Solutions Inc.* ("*Equustek*"),[463] a trademark and trade secrets case, the SCC issued the country's first global de-indexing order to an online intermediary and non-party in the underlying case, Google Search, through an interim injunction (though in practice, the injunction was permanent given the facts at hand).[464] The majority decision, also written by Abella J., again placed particular emphasis on "the realities of the Internet" but this time for the plaintiff, in addition to access to justice concerns, the lack of practical remedies without such an injunction, and Google's role and corresponding responsibility as an intermediary, even in the absence of liability:

> Datalink [the defendant] and its representatives have ignored all previous court orders made against them, have left British Columbia, and continue to operate their business

---

[456] *Crookes v Newton*, 2011 SCC 47 at para 36.

[457] [I]ndividuals may attract liability for hyperlinking if the manner in which they have referred to content conveys defamatory meaning; not because they have created a reference, but because, understood in context, they have actually *expressed* something defamatory." *ibid* at para 40 (emphasis in original).

[458] *Ibid* at para 42.

[459] Emily B Laidlaw & Hilary Young, "Internet Intermediary Liability in Defamation" (2018) 56:1 Osgoode Hall Law Journal at 118.

[460] *Ibid*.

[461] *Ibid* at 122.

[462] *Ibid*.

[463] *Google Inc v Equustek Solutions Inc*, 2017 SCC 34.

[464] *Ibid*.

> from unknown locations outside Canada. Equustek has made efforts to locate Datalink with limited success. Datalink is only able to survive—at the expense of Equustek's survival—on Google's search engine which directs potential customers to its websites. In other words, Google is how Datalink has been able to continue harming Equustek in defiance of several court orders.
>
> This does not make Google liable for this harm. It does, however, make Google the determinative player in allowing the harm to occur. On balance, therefore, since the interlocutory injunction is the only effective way to mitigate the harm to Equustek pending the resolution of the underlying litigation, the only way, in fact, to preserve Equustek itself pending the resolution of the underlying litigation, and since any countervailing harm to Google is minimal to non-existent, the interlocutory injunction should be upheld.[465]

The SCC additionally took into account Google's control over its search engine results in context of the company's self-characterization as a "content neutral" intermediary. This has potential implications for analyzing social media platforms and their content moderation practices with respect to TFGBV:

> … I have trouble seeing how [the de-indexing order] interferes with what Google refers to as its content neutral character. The injunction does not require Google to monitor content on the Internet, nor is it a finding of any sort of liability against Google for facilitating access to the impugned websites. … [Google] acknowledges, fairly, that it can, and often does, exactly what is being asked of it in this case, that is, alter search results. It does so to avoid generating links to child pornography and websites containing "hate speech". It also complies with notices it receives under the US *Digital Millennium Copyright Act* … to de-index content from its search results that allegedly infringes copyright, and removes websites that are subject to court orders.[466]

Courts have also issued interim injunctions to Google in the defamation context. For example, in *Canadian National Railway Company v Google Inc*, the company was ordered to take down an allegedly defamatory blog, which Google did not author but hosted on its blogging platform, Blogspot.[467]

In *Niemala v Malamas*, the plaintiff both sued Google for defamation and "sought an interlocutory injunction compelling Google Inc. to block from its global search results 146 universal resource locators ('URLs') for websites containing defamatory comments about the plaintiff."[468] The plaintiff considered Google liable for defamation due to the company "publishing snippets [excerpts] containing defamatory material on its search pages, and for publishing the hyperlinks to websites containing the defamatory postings".[469] First, the court denied the injunction because the plaintiff had not met the bar for "irreparable harm", given Google had already voluntarily removed the URLs from its Canadian search results and given the plaintiff had waited two years to request the injunction. With respect to whether Google is liable for defamation by virtue of its search results linking to defamatory content and

---

[465] *Ibid* at paras 52-53 (footnotes omitted).

[466] *Ibid* at paras 49-50.

[467] *Canadian National Railway Company v Google Inc*, 2010 ONSC 3121.

[468] *Pritchard v Van Nes*, 2016 BCSC 686.

[469] *Niemala v Malamas*, 2015 BCSC 1024 at para 7.

providing automatically generated snippets from linked webpages, Fenlon J. (as she then was) concluded, after reviewing both *Crookes* and *CAIP*:

> Google programs its search algorithm so that it locates URLs likely to relate to a user's search query. It is not aware of the snippets and hyperlinks produced, nor can it be, realistically. In the words of Eady J. in *Metropolitan*, Google does not authorize the appearance of the snippets on the user's screen "in any meaningful sense" but "has merely, by the provision of its search service, played the role of a facilitator": at para. 51. In summary on this issue, I conclude that Google is a passive instrument and not a publisher of snippets [and thus not liable for defamatory content contained within such snippets].[470]

Canadian courts have established jurisdiction over platform companies even if a company is based in another country, provided the facts of the case support a "real and substantial connection" to Canada or to the province where a lawsuit is based.[471] For example, in *Giustra v Twitter, Inc*, the plaintiff sued Twitter for "damages and an injunction for defamatory tweets authored by others and relayed on Twitter's internet platform."[472] Twitter argued the case should be heard in California, where the company is based and where the plaintiff also had connections and a reputation. The court found the plaintiff and the tweets had a real and substantial connection to BC and held that it thus had jurisdiction over the case,[473] following the authority of *Haaretz.com v Goldhar*, a Supreme Court of Canada case that addressed online defamation with extraterritorial elements.[474]

The Supreme Court of British Columbia in *Giustra* had the discretion to decline actually hearing the case even if it had jurisdiction, on the basis of *forum non conveniens*, if Twitter could show that California was the "*clearly* more appropriate" venue for the lawsuit.[475] Notably, in rejecting Twitter's arguments, the court took into account the broad immunity for online platforms under US law, highlighting access-to-justice implications:

> One of the significant factors in this case is that both parties acknowledge that under the law of the United States, Twitter would have no liability to Mr. Giustra pursuant to

---

[470] *Ibid* at paras 106-107.

[471] See e.g., *Giustra v Twitter, Inc*, 2021 BCSC 54 at para 34.

[472] *Ibid* at para 2.

[473] "[T]here can be no dispute that Mr. Giustra has a significant reputation in British Columbia. He also has strong ties to the province. The fact that he has a reputation in or connections to other jurisdictions does not detract from that. Giustra is not, as implied by Twitter, relying on his mere residence in British Columbia; rather he is relying on his reputation here. … [T]he notice of civil claim alleges that each tweet has been read by many people in BC. There is no evidence as to the number of people in British Columbia who read the tweets but it appears there [are] at least 500,000 twitter users in the province. In my view for the purposes of a jurisdictional challenge (where pleaded facts are taken to be correct unless challenged by evidence adduced by the defendant) Giustra has gone far enough in demonstrating damage to his reputation here. […] ": *ibid* at paras 50-51.

[474] "The governing authority regarding jurisdiction over internet defamation cases is the Supreme Court of Canada's decision in *Haaretz.com v Goldhar*, 2018 SCC 28. At para 36, the majority reiterated that the tort of internet defamation takes place where the defamatory statements are read, accessed or downloaded by a third party. Mr. Giustra's unchallenged allegation that the defamatory statements were read by persons in BC is therefore sufficient to establish the presumption of jurisdiction simpliciter. It is, then, up to Twitter to rebut the presumption." *Giustra v Twitter, Inc*, 2021 BCSC 54 at para 38. (Twitter did not successfully rebut the presumption.) See also *Douez v Facebook, Inc*, 2017 SCC 33.

[475] *Giustra v Twitter, Inc*, 2021 BCSC 54 at para 100 (emphasis in original).

> the freedom of speech protection in the First Amendment to the United States Constitution and two other statues [one of which is Section 230 of the *Communications Decency Act*]. […]
>
> The simple and obvious point here is that California cannot be an alternative forum at all much less the clearly more appropriate forum when the plaintiff would have no cause of action there for tweets published in British Columbia and harm suffered in B.C. to which B.C. law would apply under our conflict rules. […]
>
> Twitter argues that there is a preferable forum which has a cause of action for defamation; it is just that Mr. Giustra will lose because U.S. law does not recognise any liability of Twitter for this type of defamation claim. I think that is overly simplistic. It is somewhat analogous to what Brown J. observed in *Uber Technologies Inc. v. Heller* [...], a case dealing with the enforceability of an arbitration clause:
>
> > "[113] . . . there is no good reason to distinguish between a clause that *expressly* blocks access to a legally determined resolution and one that has the ultimate *effect* of doing so." [Emphasis in original.]
>
> As Brown J. also noted at para. 115, public policy requires access to justice and that is not merely access to a resolution. These comments are not inapt to the *forum conveniens* issues here."[476]

At time of writing this report (in early April 2021), the case has yet to be decided on its merits as to whether Twitter is liable for the tweets concerned.

Canadian courts may also issue *Norwich* orders to digital platform companies, where a potential defendant has remained anonymous. The Supreme Court of Nova Scotia outlined relevant factors to issuing such an order, in *Olsen v Facebook Inc*:

> It is clear from all of these authorities that a *Norwich* order may be granted to require production of identifying information with respect to persons who post anonymous comments online. Whether to do so will depend on the particular circumstances and necessitate a balancing of competing interests. The five factors identified in *York University* will govern the determination. I would expect that in many cases the application will be resolved by deciding whether the interests of justice favour the disclosure which involves consideration of the strength of the plaintiff's potential claim and the interests of privacy and freedom of expression. Whether the allegedly defamatory comments relate to a matter of public interest or are limited to a dispute between private persons is also relevant.[477]

Some general principles may be extrapolated from the cases above. Courts have generally been reluctant to hold online intermediaries liable for user expression or conduct, without something more

---

[476] *Ibid* at paras 5, 101, 113 (in-text citations omitted).

[477] *Olsen v Facebook Inc,* 2016 NSSC 155 at para 11. *York University* refers to *York University v Bell Canada Enterprises*, 99 OR (3d) 695 (SCJ). See also *Ville de Rivière-Rouge c Facebook inc*, 2020 QCCS 4300 at paras 7-8 (ordering Facebook to provide "basic subscriber information", including (translated) "vanity; account closure date, if applicable; name and e-mail address (es) and / or telephone number (s) at the time of production; date, time, and IP address of registration; and date, time, and IP addresses for recent logins and logouts, of" administrators of certain Facebook pages).

to justify attaching liability to one party for another party's wrongdoing. This is particularly the case where the intermediary is a "mere conduit" and plays a key infrastructural role, rooted in the concept of common carriage. However, Canadian law may be more amenable to applying liability for unlawful speech or harmful conduct by users if someone has been substantively personally harmed, and the platform had specific knowledge about it but took no action to remove or disable access to the user's expression. The degree of liability rises the more the platform is involved, up to direct liability where the platform has essentially abandoned its "intermediary" role in producing content that constitutes a civil or criminal offence.

Even if a platform company is not party to a case and confirmed to have no liability for the harmful content in question, the company may be expected to comply with a range of injunctions and court orders or statutory obligations such as forwarding a notice to the offending user; removing, deindexing, or disabling access to content; or releasing information to help identify an anonymous user engaging in abuse. These orders and obligations are based on ensuring access to justice and practical remedies for victims, in a way that recognizes the realities of the Internet where relevant to a dispute. Providing platforms with explicit liability shields acknowledges the particular role of platforms in the Internet ecosystem and with respect to specific harms—namely, a dominant and facilitative role which justifies accountability and responsibility for assisting in the remedy, but does not warrant imposing liability for the wrongdoing itself (absent specific knowledge or involvement).

4.2. Federal Legislation Currently in Force

Digital platforms and other kinds of online intermediaries currently may face liability or have legal obligations regarding certain kinds of user-generated content under several federal statutes in Canada. First, under the Canadian *Copyright Act*, platforms have a legal obligation to forward to users notices containing allegations of copyright infringement, enforced through limited statutory damages. Additionally, platforms may be liable for users' copyright infringement if they meet the test for 'enabling' such infringement based on a six-factor test in the *Copyright Act*. Second, some *Criminal Code* provisions for types of TFGBV, in particular those addressing 'hate propaganda' and non-consensual distribution of intimate images (NCDII) are drafted in such a way that could potentially capture digital platforms, provide all the elements of the offence were met. Platforms may also be liable for criminal corporate negligence, or for being an organization party to an offence committed by a user. Third, the *Canada-United States-Mexico Agreement* (CUSMA), although not strictly federal legislation, includes a provision that would appear to bind Canada to an intermediary liability regime that provides some baseline level of protection from liability where harmful or illegal activity from a platform's users is concerned. Each of these laws will be discussed in turn.

4.2.1. *Copyright Act*

The federal *Copyright Act* provides a legislative regime for Internet intermediaries—including both ISPs and digital platforms such as online hosts and search engines—that applies to copyright infringement specifically. Under subsection 2.4(1) of the federal *Copyright Act*, online platforms that maintain an

intermediary role closer to that of a "mere conduit" are protected by an explicit limitation of liability for copyright infringement by users; this is known as the common carrier exception.[478]

Where a platform user is alleged to have engaged in copyright infringement, the platform has a regulatory obligation known as "notice-and-notice". If an intermediary receives a notice that claims one of their users engaged in copyright infringement, and the notice complies with criteria that the *Act* sets out,[479] then the intermediary must forward that notice to the user whose account is tied to the alleged infringement, "as soon as feasible".[480] If a platform fails to forward the notice, they are not considered to be liable for copyright infringement; rather, the platform would be liable for statutory damages between $5,000 and $10,000.[481] The platform company that receives such a notice is additionally required to retain records that can be used to identify the person attached to the account tied to the alleged copyright infringement.[482] This notice-and-notice framework does not apply, of course, if the platform company engages in copyright infringement itself, in which case standard copyright liability applies. Notice-and-notice governs copyright infringement by the platform's users only.

A separate provision of the *Copyright Act*, known as the "enabler" provision, does establish platform liability for copyright infringement by their users. Under subsection 27(2.3), a platform is directly liable for copyright infringement if they have "enabled" user infringement, according to a six-factor test established in the *Act*, under subsection 27(2.4).[483] The six factors are:

> (a) whether the person expressly or implicitly marketed or promoted the service as one that could be used to enable acts of copyright infringement;
>
> (b) whether the person had knowledge that the service was used to enable a significant number of acts of copyright infringement;
>
> (c) whether the service has significant uses other than to enable acts of copyright infringement;
>
> (d) the person's ability, as part of providing the service, to limit acts of copyright infringement, and any action taken by the person to do so;
>
> (e) any benefits the person received as a result of enabling the acts of copyright infringement; and
>
> (f) the economic viability of the provision of the service if it were not used to enable acts of copyright infringement.[484]

---

[478] Paragraph 2.4(1)(b) states: "[A] person whose only act in respect of the communication of a work or other subject-matter to the public consists of providing the means of telecommunication necessary for another person to so communicate the work or other subject-matter does not communicate that work or other subject-matter to the public" (where "communication of a work to the public" can otherwise constitute an act of copyright infringement): *Copyright Act*, RSC 1985, c C-42. See also the discussion on *Society of Composers, Authors and Music Publishers of Canada v Canadian Assn of Internet Providers*, 2004 SCC 45 above.

[479] *Copyright Act*, RSC 1985, c C-42, ss 41.25(2) and 41.25(3).

[480] *Ibid*, s 41.26.

[481] *Ibid*, s 41.26(1)(a).

[482] *Ibid*, s 41.26(1)(b).

[483] *Ibid*, s 27.

[484] *Ibid*, s 27(2.4).

Andrea Slane and Ganaele Langlois have suggested adapting these factors "to determine whether an online business operates primarily for the purpose of enabling users to violate the prohibition on distributing sexual images without consent."[485] After revising each factor to apply to the context of NCDII, Slane and Langlois generalize their test into "two broader factors", specifically:

> (1) the site expressly solicits and promotes users to post sexual images of others as a means to attract users to their service and (2) the site does not meaningfully require proof of consent of those whose images are posted. If these two factors apply to a business, then this business should be prosecuted under Canada's non-consensual distribution of intimate images offence.[486]

A report by the Citizen Lab at the University of Toronto has also suggested adapting the subsection 27(2.4) "enabler" test to address another form of TFGBV, specifically to hold liable developers and vendors of stalkerware apps.[487] Stalkerware apps, which are closely tied to intimate partner violence, are widely available mobile spyware apps used to covertly track and monitor the text messages, calls, locations, and private social media activities of intimate, former, or dating partners.[488] The authors demonstrate how each of the six "enabler" factors could be applied to determine intermediary liability in the stalkerware context.[489]

Taken together, the two publications above demonstrate that a version of the enabler provision may be welcome as a meaningful approach to platform liability for at least some forms of TFGBV. The most effective frameworks would likely target a specifically defined and easily identifiable form of TFGBV with clear and substantive harms, a standard that both NCDII and use of stalkerware apps meet.

ISSUE SPOTLIGHT NO. 1
Copyright, Intermediary Liability, and Safeguarding Human Rights in Context

Examining the history and development of intermediary liability in Canadian copyright law, including its deviation from the United States in this area, serves as relevant context for determining how best to apply intermediary liability in a way that most advances and upholds human rights, across different issue areas with widely varied contexts and equities.

The notice-and-notice regime described above is generally considered a uniquely Canadian example of providing copyright claimants with a satisfactory remedy while respecting users' fair dealing rights under copyright law and without unduly intruding upon or causing disproportionate collateral

[485] Andrea Slane & Ganaele Langlois, "Debunking the Myth of 'Not My Bad': Sexual Images, Consent, and Online Host Responsibilities in Canada" 30:1 Canadian Journal of Women and the Law 42 at 57.

[486] *Ibid* at 58.

[487] Cynthia Khoo, Kate Robertson & Ronald Deibert, "Installing Fear: A Canadian Legal and Policy Analysis of Using, Developing, and Selling Smartphone Spyware and Stalkerware Applications" (June 2019) at 146-149, online (pdf): *Citizen Lab* <https://citizenlab.ca/docs/stalkerware-legal.pdf>.

[488] *Ibid* at 10-13.

[489] *Ibid* at 148-49 (Table 1).

damage to users' human rights such as the right to privacy and freedom of expression.[490] This is particularly the case relative to an alternative approach, notice-and-takedown, which the United States applies under section 512 of the *Digital Millennium Copyright Act* (DMCA).[491]

The notice-and-takedown regime in the DMCA requires digital platforms that receive claims of copyright infringement to "expeditiously […] remove, or disable access to, the material" that is alleged to be infringing—with little to no due process[492] or notice to the impacted user to provide an opportunity to respond, and ascertain if the material is in fact infringing or whether the infringement claim is in good faith.[493] This resulted in widely documented proliferation of overzealous removals, erroneous automated takedowns, and "collateral censorship",[494] including bad-faith takedown claims—such as those aimed at competitors, bad reviews, satire and parodies, or political commentary—and chilling effects on Internet users online, including those speaking from and targeted for their perspectives as members of marginalized communities.[495]

Notice-and-notice was intended to mitigate the risk that similar results would occur in Canada.[496] Even with notice-and-notice, however, abuses quickly accumulated through 'copyright trolls' sending to individuals threatening notices with false infringement claims and offers to settle, which intermediaries were required to forward.[497] The notice-and-notice provision was amended in 2019 to

---

[490] See e.g., Michael Geist, "Notice the Difference? New Canadian Internet Copyright Rules for ISPs Set to Launch" (22 December 2014), online: *Michael Geist* <http://www.michaelgeist.ca/2014/12/notice-difference-new-canadian-internet-copyright-rules-isps-set-launch/>.

[491] *Digital Millennium Copyright Act,* 17 USC § 512 (2010).

[492] Annemarie Bridy and Daphne Keller note a similar drawback to the counter-notice process for alleged copyright infringement in the US *Digital Millennium Copyright Act* (DCMA) in "US Copyright Office Section 512 Study: Comments in Response to Second Notice of Inquiry"(21 February 2017) at 29-30 [unpublished], online: *SSRN* <https://ssrn.com/abstract=2920871>; and Jennifer M Urban, Joe Karaganis & Brianna L Schofield, "Notice and Takedown in Everyday Practice" (2016) at 126-139, online (pdf): *Illusion of More* <https://illusionofmore.com/wp-content/uploads/2016/04/Berkeley_Columbia-on-512-takedown.pdf>.

[493] *Digital Millennium Copyright Act,* 17 USC § 512 (2010).

[494] Sue Gratton, "Defamation Law in the Internet Age" (March 2020) at 75, online (pdf): *Law Commission of Ontario* <https://www.lco-cdo.org/wp-content/uploads/2020/03/Defamation-Final-Report-Eng-FINAL-1.pdf>.

[495] See e.g., Jon Penney, "Privacy and Legal Automation: The DMCA as a Case Study" (2019) 22:2 Stanford Technology Law Review 412 at 470 ("Statistical analysis of survey findings for both DMCA scenarios set out in Tables 1 and 2 showed a gender effect—female respondents were consistently more likely to be chilled across a range of online activities. That is, female participants, in both Google Blogger and Twitter DMCA scenarios, were statistically less likely to speak or write online in certain contexts, less likely to share personally created content, less likely to engage with social media, and would be more careful in their online search activities both when they were personally targeted by a notice, and when a friend was targeted."); "About Us" (2017), online: *Lumen Database* <https://lumendatabase.org/pages/about>; "Using the DMCA to Censor-Options for Dealing with Abusive Notices", online: *Helbraun Law Firm* <https://www.helbraunlaw.com/using-the-dmca-to-censor-options-for-dealing-with-abusive-notice.html>; Nate Anderson, "DMCA takedowns: trampling on free speech rights?" (6 April 2010), online: *Ars Technica* <https://arstechnica.com/tech-policy/2010/04/dmca-takedowns-a-free-speech-killer/>; Joel D Matteson, "Unfair Misuse: How Section 512 of the DMCA Allows Abuse of the Copyright Fair Use Doctrine and How to Fix It" (2018) 35:2 Santa Clara High Technology Law Journal 1.

[496] See e.g., Canada, Parliament, House of Commons, Standing Committee on Industry, Science and Technology, *Statutory Review of the Copyright Act*, 42nd Parl, 1st Sess (June 2019) (Chair: Dan Ruimy) at 82; and Law Commission of Ontario, "Defamation Law in the Internet Age" (March 2020) at 85, online (pdf): <https://www.lco-cdo.org/wp-content/uploads/2020/03/Defamation-Final-Report-Eng-FINAL-1.pdf>.

[497] See e.g., Sophia Harris, "'Feels like blackmail': Canada needs to take a hard look at its piracy notice system", *CBC News* (2 November 2016), online: <https://www.cbc.ca/news/business/copyright-infringement-notice-canada-piracy-1.3831492>;

exclude such notices from being forwarded, after years of warnings and complaints of such abuses since their original enactment in 2015.[498]

Copyright infringement is the sole category of illegal activity for which the federal Canadian government has legislated a direct path to online platform liability for user activity, through the enabler provision, enacted in 2012. This seeming selective attentiveness is also the case in the United States, which had made intermediary safe harbour conditional for copyright alone, and unconditional for all other civil offences, until the controversial and fiercely contested 2018 enactment of the *Allow States and Victims to Fight Online Sex Trafficking Act* and *Stop Enabling Sex Traffickers Act* (collectively FOSTA-SESTA).[499]

Both the US and federal Canadian approaches to platform liability have appeared in contrast to the EU E-Commerce Directive[500] and Quebec's *Act to establish a legal framework for information technology*.[501] The latter two regimes make intermediary safe harbour conditional on expeditious action regarding all forms of illegal content and activity by users, not just copyright infringement alone.

For many years in Canada, no platform liability issue seemed to concern federal Parliamentarians as early and as strongly as that of copyright infringement, despite users having engaged in other illegal and arguably more harmful activities such as hate speech, criminal harassment, malicious impersonation, invasion of privacy, and NCDII on digital platforms for just as long. This early prioritizing of copyright infringement, as a user activity for which to consider holding platforms liable, is likely a reflection of the immense political power and financial resources that those most impacted by copyright infringement in particular brought to bear upon the Canadian government, lawmakers, and politicians, including industry lobbyists representing the film, music, and legacy broadcasting industries.[502] In addition, pressure from the United States and their copyright industry

---

Rosa Marchitelli, "'Shocked' grandmother on hook for illegal mutant game download", *CBC News* (31 October 2016), online: <https://www.cbc.ca/news/canada/ottawa/notice-and-notice-system-internet-copyright-enforcement-settlement-1.3823986>; Jane Wakefield, "Canadian grandmother accused of pirating zombie game", *BBC* (1 November 2016), online: <https://www.bbc.com/news/technology-37834766>; Karl Bode, "Canada Eyes Ban On Shady Piracy Warnings That Demand Payment" (1 November 2018), online: *Vice* <https://www.vice.com/en_us/article/kzj49v/canada-bill-c-86-ban-copyright-notices-that-demand-settlement>; Garett Williams, "U of M forwards 8,000 emails regarding illegal downloads Copyright office likens threatening notices to extortion", *The Manitoban* (7 September 2016), online: <www.themanitoban.com/2016/09/u-of-m-forwards-8000-emails-regarding-illegal-downloads/28934/>.

[498] *Copyright Act*, RSC 1985, c C-42, s 41.25(c); Nick Kirmse, "Copyright notices can no longer demand payment for alleged piracy" , CTV News (27 January 2019), online: <https://www.ctvnews.ca/sci-tech/copyright-notices-can-no-longer-demand-payment-for-alleged-piracy-1.4271132>.

[499] *Allow States and Victims to Fight Online Sex Trafficking Act of 2017*, Pub L No 115-164, 132 Stat 1253 (2018).

[500] EC, *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market*, [2000] OJ, L 178/1.

[501] CQLR, c C-1.1, s 22.

[502] Michael Geist has documented instances of such lobbying and pressure throughout the years: see e.g., "The Power of Backroom Lobbying: How the Recording Industry Got Their Copyright Term Extension" (28 April 2015), online: *Michael Geist* <www.michaelgeist.ca/2015/04/the-power-of-backroom-lobbying-how-the-recording-industry-got-their-copyright-term-extension/>; "One-Sided Story: Lobbyist Data Shows Music, Movie and Publisher Groups Account For 80 Per cent of Registered Copyright Meetings in Canada Since 2015 Election" (14 August 2018), online: *Michael Geist* <www.michaelgeist.ca/2018/08/onesidedstory/>; "Canadian Heritage Minister Joly Hints Many Cultural Groups Don't Comply With Lobbyist Reporting Rules" (3 November 2017), online: *Michael Geist* <www.michaelgeist.ca/2017/11/canadian-heritage-minister-joly-hints-many-cultural-groups-dont-comply-lobbyist-reporting-rules/>; "Movie Industry Denies Lawsuit Strategy

representatives, such as through trade agreement negotiations further contributed to the heavy focus on copyright infringement.[503]

Issues central to copyright law and its key stakeholders have thus driven much of the early historical context of Canadian intermediary liability law, to the extent of law and related literature explicitly categorized under the umbrella of "intermediary liability" (as opposed to other laws of general application that may happen to apply to an intermediary platform as they would to any other business or organization). This history is relevant to discussions of platform liability or accountability for TFGBV, because the rights and interests at stake are substantially different from those of, for example, licensing fees and copyright royalties.

In fact, as Rebecca Katz examines, the historical prioritization of addressing copyright infringement above all other illegal or more harmful activities by platform users in the intermediary liability space has already led to copyright law being "touted elsewhere as an unexpected addition to [NCDII] victims' legal toolkits"[504] and proposals to "expand copyright to address NCDII subjects (not just creators)".[505] Such proposals "have been criticized for potentially destabilizing this branch of the law, or setting a precedent for other, more powerful [copyright] rightsholders to demand further rights expansions, without necessarily preventing [NCDII]."[506] This illustrates the dangers and pitfalls of applying an intermediary liability framework established for one particular objective (copyright) to address another objective (NCDII) that requires a greatly divergent analysis in assessing the proportionality of robust or severe legal responses.

What is considered a proportionate approach to platform liability in the copyright context is unlikely to be an equally appropriate calibration when it comes to TFGBV, which involves attacking the fundamental human rights and well-being of historically marginalized and systemically oppressed individuals and communities. 'Classical' digital rights debates regarding copyright, intermediary liability, and Internet users' right to fair dealing and freedom of expression may not transpose well, and should not be transposed directly, to discussions where the central objective is protecting marginalized users and their physical and psychological safety, as well as their often disproportionately violated rights to equality, freedom of expression, and privacy.

Despite Proliferation of Legal Actions and Settlement Demands Against Thousands of Canadians" (9 July 2018), online: *Michael Geist* <www.michaelgeist.ca/2018/07/movie-industry-denies-lawsuit-strategy-despite-proliferation-of-legal-actions-and-settlement-demands-against-thousands-of-canadians/>; and Patrick O'Rourke, "CRTC denies Bell-led FairPlay Canada coalition on 'jurisdictional grounds'", *MobileSyrup* (2 October 2018), online: <https://mobilesyrup.com/2018/10/02/crtc-denies-fairplay-canada-coalition-jurisdictional-grounds/>.

[503] See e.g., Alex Boutilier, "Canada capitulates on copyright in new USMCA deal, experts say", *Toronto Star* (1 October 2018), online: <https://www.thestar.com/news/canada/2018/10/01/canada-capitulates-on-copyright-in-new-usmca-deal.html>; Michael Geist, "U.S. Lobby Groups Take Aim At Canadian Copyright Law in NAFTA Comments: No Balance, No Fair Use, & No Cultural Exception" (22 June 2017), online: *Michael Geist* <www.michaelgeist.ca/2017/06/naftacopyrightcomments/>; Margot E Kaminski, "The Capture of International Intellectual Property Law through the U.S. Trade Regime" (2014) 87 Southern California Law Review 977; Ross Bagley, "USMCA Set To Export U.S. Copyright Law to North American Neighbors" (29 January 2020), online: *IP Watchdog* <https://www.ipwatchdog.com/2020/01/29/usmca-set-export-u-s-copyright-law-north-american-neighbors/id=118269/>.

[504] Rebecca Katz, "Takedowns and Trade-Offs: Can Copyright Law Assist Canadian Victims of Non-Consensual Intimate Image Distribution?" (2020) 29 Education & Law Journal 169 at 169.

[505] *Ibid* at 181.

[506] *Ibid*.

Conversely, where equality rights and harm reduction objectives may call for moving the needle towards greater liability and accountability for digital platforms, stakeholders invested in the issue from a business perspective—such as legacy publishing and broadcasting industries—should not be permitted to opportunistically 'piggyback' on or exploit gender equality and racial justice advocacy as a way to strengthen laws imposing platform liability across the board, to further serve their own commercial interests.[507]

It is both justified and necessary to impose different levels and types of obligations and liability on digital platforms, depending on the specific context and area of law concerned. This does not result in inconsistency, but rather reflects principled recognition that different objectives and issue areas are, correctly, informed and shaped by different underlying legal principles and values, rationales, case law, and protected rights or interests specific to each particular issue. Legislators, policy analysts, and other relevant decision-makers must maintain such nuances and distinctions. Context-sensitivity is required to increase the chances that the appropriate balance will be struck when it comes to determining platform liability in one particular context, without distorting the balance to be struck in an altogether different context, to reach solutions that are considered proportionate relative to any respective trade-offs.

A more in-depth discussion of proportionality in the context of Canadian constitutional law, platform liability for TFGBV, and upholding the rights to equality and freedom of expression is presented in Part 6 ("Constitutional and Critical Analysis of Platform Liability for TFGBV").

4.2.1. *Criminal Code*

Some provisions in the Canadian *Criminal Code* could be interpreted to give rise to platform liability for TFGBV in certain circumstances.[508] Specifically, these are offenses concerning NCDII and child sexual abuse material (CSAM), in sections 162.1 and 163.1(3), respectively, and offences where an organization has been criminally negligent or a party to an offence, in sections 22.1 and 22.2, respectively.

Subsection 162.1(1) of the *Criminal Code* makes it a hybrid (i.e., either a summary or indictable) offence if someone "knowingly publishes, distributes, transmits, sells, makes available or advertises" an intimate image of someone without their consent.[509] It is possible that this provision could be interpreted to hold a digital platform criminally liable for distributing, transmitting, or making available an intimate image without consent, if it could be shown the platform company acted "knowingly" in a particular case.[510]

---

[507] See e.g. Cynthia Khoo, "Crafting Internet policy with nuance, not kneejerks" (16 May 2018), online: *Policy Options* <https://policyoptions.irpp.org/magazines/may-2018/crafting-internet-policy-nuance-not-kneejerks/>.

[508] For certain offences such as hate speech and CSAM, there are provisions that do not address platform liability, but instead grant judges the power effectively to order a digital platform to "ensure that the material is no longer stored on and made available" through the platform's system: *Criminal Code*, RSC 1985, c C-46, ss 164.1(1)(b) and 320.1(1)(b).

[509] *Ibid*, s 162.1(1).

[510] For elaboration on how section 162.1 of the *Criminal Code* could be applied to platform intermediaries, including proposed amendments, see Andrea Slane & Ganaele Langlois, "Debunking the Myth of 'Not My Bad': Sexual Images, Consent, and Online Host Responsibilities in Canada" 30:1 Canadian Journal of Women and the Law 42. As of 2020, Rebecca Katz noted that "this approach has not yet been tried in Canada", and this appears to remain the case at time of writing of this

Subsection 163.1(3) makes it a criminal offence if someone "transmits, makes available, distributes, sells, advertises, imports, [or] exports…" CSAM.[511] Unlike in the NCDII offence above, no *mens rea* (fault element) is specified. However, the courts have interpreted subsection 163.1(3) to include a presumption that knowledge is required.[512] In *R v Spencer*, which involved CSAM being distributed online through the peer-to-peer file-sharing program LimeWire, the SCC established that there is no requirement that the accused "knowingly, by some positive act, facilitate the availability of the material."[513] It is enough to have demonstrated wilful blindness to the fact that one may be making CSAM available,[514] such as, in the case of *Spencer*, knowledge of certain facts about how LimeWire worked and its potential consequences.[515] Thus, this offence could potentially apply to an online platform if the company were shown to have had knowledge or wilful blindness in a particular case.

Paragraph 320.1(1)(b) of the *Criminal Code*, which concerns "hate propaganda" (as defined),[516] does not expose an intermediary platform to direct liability, but recognizes that such platforms may be involved and have corresponding obligations, to the extent a platform is a "computer system" (as defined).[517] The provision states that if a judge finds there are "reasonable grounds to believe" that "hate propaganda" is "stored on and made available to the public through a computer system", then the judge may order the "custodian of the computer system to … ensure that the material is no longer stored on and made available through the computer system".[518] Paragraph 320.1(1)(b) could potentially be used to compel a platform company to remove or terminate the availability of some instances of TFGBV on its platform, where the content meets the definition of "hate propaganda".

---

report. Rebecca Katz, "Takedowns and Trade-Offs: Can Copyright Law Assist Canadian Victims of Non-Consensual Intimate Image Distribution?" (2020) 29 Education & Law Journal 169 at 175.

[511] *Criminal Code*, RSC 1985, c C-46, s 163.1(3).

[512] *R v Branco*, 2019 ONSC 1026 at paras 18-19.

[513] *R v Spencer*, 2014 SCC 43 at paras 82-86.

[514] "There is no dispute that the accused in a prosecution under s. 163.1(3) of the *Criminal Code* must be proved to have had knowledge that the pornographic material was being made available. This does not require, however, as the trial judge suggested, that the accused must knowingly, by some positive act, facilitate the availability of the material. I accept Caldwell J.A.'s conclusion that the offence is complete once the accused knowingly makes pornography available to others. As he put it, '[i]n the context of a file sharing program, the *mens rea* element of making available child pornography under s. 163.1(3) requires proof of the intent to make computer files containing child pornography available to others using that program *or actual knowledge that the file sharing program makes files available to others*.' [para. 87]" (emphasis added): *ibid* at para 83.

[515] "The evidence calling for consideration of wilful blindness included, for example, evidence that in Mr. Spencer's statement to police he acknowledged that LimeWire is a file sharing program; that he had changed at least one default setting in LimeWire; that when LimeWire is first installed on a computer, it displays information notifying the user that it is a file sharing program; that at the start of each session, LimeWire notifies the user that it is a file sharing program and warns of the ramifications of file sharing; and that LimeWire contains built-in visual indicators that show the progress of the uploading of files by others from the user's computer: paras. 88-89." *ibid* at para 85.

[516] "Hate propaganda" is defined in section 320(8) of the *Criminal Code,* RSC 1985, c C-46: "hate propaganda means any writing, sign or visible representation that advocates or promotes genocide or the communication of which by any person would constitute an offence under section 319".

[517] "Computer system" is defined in subsection 342.1(2) of the *Criminal Code*, RSC 1985, c C-46: "computer system means a device that, or a group of interconnected or related devices one or more of which, (a) contains computer programs or other computer data, and (b) by means of computer programs, (i) performs logic and control, and (ii) may perform any other function".

[518] *Ibid*, s 320.1(1)(b).

Section 22.1 of the *Criminal Code* sets out the criteria to demonstrate that an organization was criminally negligent with respect to a particular offence. Specifically, a digital platform would have committed this offence if two conditions are met. First, one or more of the platform's representatives, while "acting within the scope of their authority", engages in conduct "by act or omission" that made them party to the offence.[519] Second, a senior officer responsible for activities relevant to the offence "departs […] markedly from the standard of care" that "could reasonably be expected" to have prevented the representative(s) from becoming party to the offence in question.[520]

Section 22.2 sets out the criteria to demonstrate that an organization was at fault beyond negligence (i.e., that the organization knowingly participated in the offence, or otherwise had the requisite intent):

> [A]n organization is a party to the offence if, with the intent at least in part to benefit the organization, one of its senior officers
>
> > (a) acting within the scope of their authority, is a party to the offence;
> >
> > (b) having the mental state required to be a party to the offence and acting within the scope of their authority, directs the work of other representatives of the organization so that they do the act or make the omission specified in the offence; or
> >
> > (c) knowing that a representative of the organization is or is about to be a party to the offence, does not take all reasonable measures to stop them from being a party to the offence.[521]

At time of writing, there do not appear to have been attempts to hold a platform company criminally liable for an offence committed by its users, involving TFGBV or not, through either sections 22.1 or 22.2.[522] However, both of these provisions would only apply to the narrow and discrete acts of TFGBV that already constitute criminal offences in their own right, and which can be difficult to prosecute successfully.[523] That limitation leaves out much expression-based TFGBV in particular. This is not to say that all forms of TFGBV should be criminalized for the purpose of being able to apply sections 22.1 and 22.2, but rather, to highlight that meaningfully addressing the full spectrum of TFGBV requires looking beyond the criminal law, even while relying on the latter as one among many tools to use where needed.

---

[519] *Criminal Code*, s 22.1(a)

[520] *Ibid*, s 22.1.

[521] *Ibid*, s 22.2.

[522] For further discussion on determining the presence or absence of "intent" and "knowledge" on the part of an online intermediary in the context of NCDII, see Hilary Young & Emily Laidlaw, "Creating a Revenge Porn Tort for Canada" (2020) 96 Supreme Court Law Review 147 at 180-84. For more details on corporate criminal liability in Canada generally, outside of but possibly applicable to the context of TFGBV, and how to establish that a company had the required *mens rea* for a particular offence, see e.g., Theo Milosevic, "Corporate Criminal Liability for Algorithmic Price-Fixing in Canada" (2018) 16 Canadian Journal of Law & Technology 417 at 428-30; Erin Sheley, "Victim Impact Statements at Canadian Corporate Sentencing" (2020) 43:3 Manitoba Law Journal 421 at 425-431; and Allens Arthur Robinson, "'Corporate Culture' as a Basis for the Criminal Liability of Corporations" (February 2008) at 24-28, online (pdf): *Business & Human Rights Resource Centre* <https://media.business-humanrights.org/media/documents/f72634fd87adfd3d31a22f5f4b93150267b8a764.pdf>.

[523] See e.g., the hate speech offences in the *Criminal Code*, RSC 1985, c C-46, ss 318-319.

4.2.3. *Canada-United States-Mexico Agreement* (CUSMA)

In April 2020, the Canadian government ratified the *Canada-United States-Mexico Agreement* (CUSMA),[524] which replaced the former *North American Free Trade Agreement* (NAFTA). Many have compared Article 19.17 of CUSMA to section 230 of the US *Communications Decency Act*. For that reason, and because it is a treaty signed at the federal level, CUSMA is examined alongside federal legislation, though there may be a division of powers question at the implementation stage. Article 19.17(2) states:

> [N]o Party shall adopt or maintain measures that 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.[525]

Some have interpreted this provision, in conjunction with Article 19.17(4),[526] to mean that online platforms cannot be held civilly liable for any legal wrongdoing in Canada, except for intellectual property infringement, without violating CUSMA.[527] However, others have pointed out that "[w]hile Article 19.17.2 of the USMCA imports principles similar to those of paragraph 230(c)(1), it does not necessarily import *the judicial interpretations of those principles*, which has resulted in the broad immunity provided in the U.S. to websites, search engines, ISPs, and service providers."[528]

Additionally, Article 19.17 does not mandate complete restrictions on placing any legal obligations whatsoever on digital platforms, separate from the question of determining liability. This "leaves open the possibility of equitable remedies" even in the absence of direct liability for a given harm, such as issuing an injunction,[529] as well as the possibility of regulatory obligations that do not place digital platforms in the same legal position as a direct perpetrator of TFGBV, but rather takes into account the platform's own particular role with respect to the harm caused.

---

[524] *Canada-United States-Mexico Agreement Implementation Act*, SC 2020, c 1.

[525] *Protocol replacing the North American Free Trade Agreement with the Agreement between Canada, the United States of America and Mexico*, 30 November 2018, CAN TS 2020 No 6, art 19.17(2) (entered into force in Canada 1 July 2020).

[526] Article 19.17(4) exempts intellectual property law, criminal law, and requirements to comply with "a specific, lawful order of a law enforcement authority". *Ibid*, art 19.17(4).

[527] See e.g., Ethan Phillips, "Revised NAFTA agreement imposes U.S. internet rules on Canada (21 December 2019), online: *Canada Fact Check* <https://canadafactcheck.ca/2019/12/21/revised-nafta-agreement-imposes-u-s-internet-rules-on-canada/>; Patrick Leblond, "Digital Trade at the WTO: The CPTPP and CUSMA Pose Challenges to Canadian Data Regulation" (October 2019), CIGI Papers No 227 at 6, online: *Centre for International Governance Innovation* <https://www.cigionline.org/sites/default/files/documents/no.227.pdf> ("CUSMA's article 19.17 will likely make it harder for Canadian governments to develop measures to protect individuals and consumers of social media, search engines and other user-generated content providers from the consequences of disinformation (for example, "fake news")."

[528] "Demystified: USMCA's Digital Trade Provisions on ISP Liability in Canada" (14 November 2018), online: *Blakes* <https://blakes.com/insights/bulletins/2018/demystified-usmcas-digital-trade-provisions-on-isp> (emphasis added). See also Danielle Keats Citron and Benjamin Wittes, "The Problem Isn't Just Backpage: Revising Section 230 Immunity" (2018) 2 Georgetown Law Technology Review 453 at 459, 462.

[529] An equitable remedy is a "non-monetary forms of relief granted by courts when other legal remedies will not adequately redress an injury": Vivek Krishnamurthy & Jessica Fjeld, "CDA 230 Goes North American? Examining the Impacts of the USMCA's Intermediary Liability Provisions in Canada and the United States" (July 2020) at 7, online (pdf): *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3645462>.

4.3. Federal Legislation Announced and Parliamentary Studies

As of late 2019, the Canadian federal government began signalling to the public that there would be forthcoming legislation to address certain categories of harmful content on social media platforms.[530] Prime Minister Justin Trudeau's mandate letter to Steven Guilbeault, the Minister of Canadian Heritage, tasked the minister with creating "new regulations for social media platforms, starting with a requirement that all platforms remove illegal content, including hate speech, within 24 hours or face significant penalties. This should include other online harms such as radicalization, incitement to violence, exploitation of children, or creation or distribution of terrorist propaganda."[531] Similarly, the mandate letter and supplementary mandate letter to the Minister of Justice and Attorney General of Canada, David Lametti, instructed him to "[d]evelop options for legal remedies for victims of hate speech" and work with the Ministers of Canadian Heritage, Public Safety and Emergency Preparedness, and Diversity and Inclusion and Youth to "take action on combatting hate groups and online hate and harassment, ideologically motivated violent extremism and terrorist organizations".[532]

During a meeting of the House of Commons Standing Committee on Canadian Heritage in January 2021, Minister Guilbeault shared that his department had been working "with the Department of Justice, the Department of Public Safety and Emergency Preparedness, and the Department of Innovation … to bring forward a bill that will set out a regulatory framework to control hate speech, child pornography, incitement to violence, incitement to terrorism and the non-consensual disclosure of images",[533] projected to arrive in spring of 2021. The legislation reportedly will also "create a new government regulator with the power to monitor social media platforms and levy fines on social media companies

[530] See e.g., "Online Hate Speech, Exploitation and Harassment Online", online: *Liberal Party of Canada*, <https://liberal.ca/our-platform/online-hate-speech-exploitation-and-harassment-online/>; Janet E Silver, "Regulation of online hate speech coming soon, says minister" (29 January 2021), online: *iPolitics* <https://ipolitics.ca/2021/01/29/regulation-of-online-hate-speech-coming-soon-says-minister/>; Elizabeth Thompson, "Canada not exempt from social media forces that created U.S. Capitol riot, heritage minister says", *CBC News* (29 January 2021), online: <https://www.cbc.ca/news/politics/facebook-twitter-canada-regulation-1.5894301>.

[531] Rt Hon Justin Trudeau, PC, MP, Prime Minister of Canada, "Minister of Canadian Heritage Mandate Letter" (13 December 2019), online: *Prime Minister of Canada* <https://pm.gc.ca/en/mandate-letters/2019/12/13/minister-canadian-heritage-mandate-letter>.

[532] Rt Hon Justin Trudeau, PC, MP, Prime Minister of Canada, "Minister of Justice and Attorney General of Canada Mandate Letter" (13 December 2019), online: *Prime Minister of Canada* <https://pm.gc.ca/en/mandate-letters/2019/12/13/minister-justice-and-attorney-general-canada-mandate-letter>; and Rt Hon Justin Trudeau, PC, MP, Prime Minister of Canada, "Minister of Justice and Attorney General of Canada Supplementary Mandate Letter" (15 January 2021), online: *Prime Minister of Canada* <https://pm.gc.ca/en/mandate-letters/2021/01/15/minister-justice-and-attorney-general-canada-supplementary-mandate>; see also Rt Hon Justin Trudeau, PC, MP, Prime Minister of Canada, "Minister of Heritage Supplementary Mandate Letter" (15 January 2021), online: *Prime Minister of Canada* <https://pm.gc.ca/en/mandate-letters/2021/01/15/minister-canadian-heritage-supplementary-mandate-letter>.

[533] Canada, Parliament, House of Commons, Standing Committee on Canadian Heritage, *Evidence*, 43rd Parl, 2nd Sess, No 12 (29 January 2021).

that allow things like hate speech to remain on their platforms".[534] It would be up to the regulator itself to determine specifically how it will apply the legal framework created by the legislation.[535]

According to media interviews with Minister Guilbeault, the regulator may also have powers to enforce transparency requirements[536] and audit social media platforms' algorithms—albeit not powers to "go after proprietary information",[537] which would seem to render the power pointless given the likely proprietary nature of such platforms' most relevant algorithms.[538] Other elements of the new regulatory scheme being considered at time of writing are a complaint process for users, and an independent appeal process for removed content.[539] It is unknown whether this appeal process would also be available for content that has been left up on a platform, which is often the greater concern for those targeted by TFGBV.

One of the prevailing concerns with regulating user speech through intermediary platforms is with how different types of user speech and expression is defined for the purpose of making content moderation decisions.[540] This makes particularly notable that the advertised bill from the Department of Canadian Heritage, in conjunction with the Department of Justice, may introduce "a new statutory definition of hate … based on previous court decisions and how the Supreme Court has defined hate", including the landmark case *Saskatchewan (Human Rights Commission) v Whatcott*.[541] Again, it is left to the new

---

[534] Elizabeth Thompson, "Canada not exempt from social media forces that created U.S. Capitol riot, heritage minister says", *CBC News* (29 January 2021), online: <https://www.cbc.ca/news/politics/facebook-twitter-canada-regulation-1.5894301>.

[535] Anja Karadeglija, "New definition of hate to be included in Liberal bill that might also revive contentious hate speech law", *National Post* (3 March 2021), online: <https://nationalpost.com/news/politics/new-definition-of-hate-to-be-included-in-liberal-bill-that-might-also-revive-contentious-hate-speech-law>.

[536] *Ibid*.

[537] Kieran Leavitt, "Ottawa ready to give police more powers to go after social media companies and the people who use them", *Toronto Star* (27 January 2021), online: <https://www.thestar.com/politics/federal/2021/01/27/ottawa-ready-to-give-police-more-powers-to-go-after-social-media-companies-and-the-people-who-use-them.html>.

[538] See e.g., Jeremy B Merrill & Ariana Tobin, "Facebook Moves to Block Ad Transparency Tools — Including Ours", *ProPublica* (28 January 2019), online: <https://www.propublica.org/article/facebook-blocks-ad-transparency-tools>; and Paddy Leerssen, "The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems" (2020) 11:2 European Journal of Law and Technology at 13.

[539] Kieran Leavitt, "Ottawa ready to give police more powers to go after social media companies and the people who use them", *Toronto Star* (27 January 2021), online: <https://www.thestar.com/politics/federal/2021/01/27/ottawa-ready-to-give-police-more-powers-to-go-after-social-media-companies-and-the-people-who-use-them.html>.l

[540] As just one example, Lex Gill points out that different phrases used to describe offensive or hurtful expression may evoke wider or narrower swaths of expression, with potential legal implications: "This report generally uses the language of 'hate speech' (which tends to be the common term used by Canadian courts), or 'hate propaganda' when describing the *Criminal Code* offence. However, [...] certain authors choose different language, some of which may encompass expression that is lawful in Canada (e.g., 'harmful speech'), while others suggest a more narrow scope than the Canadian legal definition of hate speech would tend to include (e.g., 'violent' or 'dangerous' speech)" Lex Gill, "The Legal Aspects of Hate Speech in Canada" (June 2020), at 6, online: *Public Policy Forum* <https://ppforum.ca/wp-content/uploads/2020/07/1.DemX_LegalAspects-EN.pdf> (footnotes omitted). This variance in terminology would have practical implications if a platform were legally required to prohibit, for example, 'hate propaganda' as opposed to 'harmful speech'. The application of such terms to specific user posts would be additionally influenced by how the platform company or its content moderators interpret such terms for themselves.

[541] Anja Karadeglija, "New definition of hate to be included in Liberal bill that might also revive contentious hate speech law", *National Post* (3 March 2021), online: <https://nationalpost.com/news/politics/new-definition-of-hate-to-be-included-in-liberal-bill-that-might-also-revive-contentious-hate-speech-law>.

regulator "to give clarity on the new definition" to digital platforms that fall under the new law.[542] This also leaves open the question of how each of the other four categories of content—child pornography, incitement to violence, incitement to terrorism and NCDII—will be defined for the purposes of enforcing the new legislation, whether hewing to pre-existing legal definitions or creating new statutory ones for this specific context. However, the Heritage Minister's office has stated that the legislation will not "expand the definition of illegal content beyond what's already in the *Criminal Code*".[543]

Additionally, the forthcoming legislation may include new law enforcement mechanisms that would require or permit social media platforms to pass certain content moderation cases or related information to the police.[544] According to Minister Guilbeault, "Law enforcement will have the ability to get information from the platforms to prosecute the individuals or groups of individuals in question."[545]

It is also possible that the federal government may reinstate a version of section 13 of the *Canadian Human Rights Act*,[546] which established that it is unlawfully discriminatory for an individual or a group to communicate online "any matter that is likely to expose a person or persons to hatred or contempt by reason of the fact that that person or those persons are identifiable on the basis of a prohibited ground of discrimination".[547] The provision was repealed in 2013 after being subjected to critiques of overbreadth and online censorship, despite having been found constitutional by the Supreme Court of Canada and the Federal Court of Appeal. While the former section 13 applies only to the direct speaker, and could be interpreted to specifically exclude owners and operators of online intermediaries,[548] it remains to be seen whether legislators will attempt to bring digital platforms within the ambit of a newly revived section 13, given the evolution of context around digital platforms and their role in society, industry and sociopolitical developments, and sensibilities and attitudes towards platform companies over the intervening years since the provision's repeal.

The above-described legislation in part builds on, or is progressing in parallel to, several Parliamentary studies on TFGBV and related issues. These include the following studies:

- *Taking Action to End Violence against Young Women and Girls in Canada* (March 2017), by the House of Commons Standing Committee on the Status of Women (FEWO);[549]

---

[542] *Ibid*.

[543] *Ibid*.

[544] Kieran Leavitt, "Ottawa ready to give police more powers to go after social media companies and the people who use them", *Toronto Star* (27 January 2021), online: <https://www.thestar.com/politics/federal/2021/01/27/ottawa-ready-to-give-police-more-powers-to-go-after-social-media-companies-and-the-people-who-use-them.html>.

[545] *Ibid*.

[546] Anja Karadeglija, "New definition of hate to be included in Liberal bill that might also revive contentious hate speech law", *National Post* (3 March 2021), online: <https://nationalpost.com/news/politics/new-definition-of-hate-to-be-included-in-liberal-bill-that-might-also-revive-contentious-hate-speech-law>.

[547] *Canadian Human Rights Act*, RSC 1985, c H-6, s 13 [repealed].

[548] "For the purposes of this section, no owner or operator of a telecommunication undertaking communicates or causes to be communicated any matter described in subsection (1) by reason only that the facilities of a telecommunication undertaking owned or operated by that person are used by other persons for the transmission of that matter." *ibid*, s 13(3).

[549] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada*, 42nd Parl, 1st Sess (March 2017) (Chair: Marilyn Gladu).

- *Taking Action to End Online Hate* (June 2019), by the House of Commons Standing Committee on Justice and Human Rights (JUST), which informed the Department of Justice's subsequent involvement in the forthcoming legislation;[550] and
- *Protection of Privacy and Reputation on Platforms such as Pornhub* (February 2021), a currently ongoing study by the House of Commons Standing Committee on Access to Information, Privacy and Ethics (ETHI), which has not yet resulted in a published report.[551]

Moreover, some of these committees have introduced motions to undertake further relevant studies, but these have not yet begun or have been set aside for the time-being in order to pursue other priorities. The Standing Committee on Canadian Heritage (CPHC), in response to the JUST report on online hate and Minister Guilbeault's December 2019 mandate letter, introduced in February 2020 and reintroduced in October 2020 a motion to

> study the creation and implementation of new regulations for online media platforms and Internet service providers requiring them to monitor, address and remove content that constitutes hate speech and remove any other content which is illegal in Canada or prohibited by the Criminal Code, such as incitement of violence, incitement of genocide, creation or distribution of terrorist propaganda and exploitation of children, in a timely manner; that the committee hold at least no less than four meetings on this subject with relevant witnesses; and that the Committee report its findings to the House of Commons.[552]

In December 2020, FEWO introduced, but has not since followed up on (as of March 2021), a motion to "undertake a study on the sexual violence and exploitation experienced by women" resulting from pornography and CSAM distributed online for commercial purposes, in Canada with seeming impunity and "under no Canadian legislative framework".[553] The motion specifically identifies Pornhub and its Canadian owner, MindGeek, as a focus of investigation, in addition to the "devastating psychological

---

[550] Canada, Parliament, House of Commons, *Taking Action to End Online Hate: Report of the Standing Committee on Justice and Human Rights*, 42nd Parl, 1st Sess (June 2019) (Chair: Anthony Housefather).

[551] Canada, Parliament, House of Commons, Standing Committee on Access to Information, Privacy and Ethics, *Protection of Privacy and Reputation on Platforms such as Pornhub*, 43rd Parl, 2nd Sess, online: *Parliament of Canada* <https://www.ourcommons.ca/Committees/en/ETHI/StudyActivity?studyActivityId=11088039>; Canada, Parliament, House of Commons, Standing Committee on Access to Information, Privacy and Ethics, *Minutes of Proceedings*, 43rd Parl, 2nd Sess, No 16 (11 December 2020) ("On motion of Nathaniel Erskine-Smith, it was agreed, — That the committee invite to appear representatives of Pornhub / Mindgeek, namely Feras Antoon and David Tassillo, to explain the company's failure to prohibit rape videos and other illegal content from its site, and what steps it has taken and plans to take to protect the reputation and privacy of young people and other individuals who have never provided their consent.")

[552] Canada, Parliament, House of Commons, Standing Committee on Canadian Heritage, *Minutes of Proceedings*, 43rd Parl, 1st Sess, No 1 (19 February 2020) (notice of motion given); Canada, Parliament, House of Commons, Standing Committee on Canadian Heritage, *Evidence*, 43rd Parl, 2nd Sess, Number 2 (23 October 2020) (motion follow-up). As of December 10, 2020, the motion had not yet been adopted: "Regulation of social media Platforms" (10 December 2020), online: *Government of Canada* <https://search.open.canada.ca/en/qp/id/pch,PCH-2020-QP-00084>.

[553] Canada, Parliament, House of Commons, Standing Committee on the Status of Women (FEWO), *Minutes of Proceedings*, 43rd Parl, 2nd Sess, No 10 (10 December 2020).

effects on victims of sex crimes and the effects on the lives of women" who appear in NCDII and potential legislative measures to "prevent the production or distribution" of NCDII and CSAM.[554]

Lastly, the federal government has published two other reports that are not about TFGBV or online hate on digital platforms, but may have adjacent repercussions when it comes to issues of platform regulation and platform liability. These reports are: *Democracy Under Threat: Risks and Solutions in the Era of Disinformation and Data Monopoly* (December 2018),[555] by the House of Commons Standing Committee on Access to Information, Privacy and Ethics (ETHI), undertaken in response to the Facebook and Cambridge Analytica scandal; and *Canada's Communications Future: Time to Act* (January 2020),[556] a report by the federal Broadcasting and Telecommunications Legislative Review (BTLR) Panel making a series of recommendations to overhaul Canada's telecommunications and broadcasting legal regimes.

4.4. Provincial Legislation

Relevant provincial legislation that applies to platform liability for TFGBV is discussed in two sections below. Section 4.4.1 will present a general intermediary liability framework enacted in Quebec and Section 4.4.2 will discuss legislation addressing NCDII, which six provinces have enacted to date.

4.4.1. Quebec Intermediary Liability Provision

At the provincial level, only Quebec has a general intermediary liability regime, enacted in 2001. Section 22 of the *Act to establish a legal framework for information technology*[557] establishes a form of safe harbour for certain kinds of online intermediaries, similarly to that in the EU E-Commerce Directive.[558] As the default position, platform companies that fall under section 22 are not responsible for the activities their users engage in through the companies' services. The companies lose this shield if they fail to act promptly upon becoming aware of illicit user activity.

Specifically, intermediaries that provide "document storage services on a communication network" risk incurring liability if, "upon becoming aware that the documents are being used for an illicit activity, or of circumstances that make such a use apparent, the service provider does not act promptly to block access to the documents or otherwise prevent the pursuit of the activity."[559] Intermediaries that provide "technology-based documentary referral services, such as an index, hyperlinks, directories or search

---

[554] *Ibid*.

[555] Canada, Parliament, House of Commons, Standing Committee on Access to Information, Privacy and Ethics, *Democracy under Threat: Risks and Solutions in the Era of Disinformation and Data Monopoly*, 42nd Parl, 1st Sess (December 2018) (Chair: Bob Zimmer).

[556] Canada, Broadcasting and Telecommunications Legislative Review Panel, *Canada's Communications Future: Time to Act* (Ottawa: Government of Canada, 2020), online (pdf): *Innovation, Science and Economic Development Canada* <https://www.ic.gc.ca/eic/site/110.nsf/eng/00012.html>.

[557] *Act to establish a legal framework for information technology*, CQLR c C-1.1, s 22.

[558] EC, *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market*, [2000] OJ, L 178/1, arts 12-14.

[559] *Act to establish a legal framework for information technology*, CQLR c C-1.1, s 22.

tools" lose protection "if, upon becoming aware that the services are being used for an illicit activity, the service provider does not act promptly to cease providing services to the persons known by the service provider to be engaging in such an activity."[560]

The above descriptions capture, at the very least, intermediaries such as social media platforms, online forums, image- and video-sharing websites ("document" is defined to include all manner of multimedia[561]), and search engines. However, at time of writing, there have been no published court decisions that involve using section 22 to hold a platform company liable for a user's wrongdoing.[562]

4.4.2. Provincial NCDII Statutes

Six provinces—Nova Scotia, Saskatchewan, Manitoba, Alberta, Prince Edward Island (PEI), and Newfoundland and Labrador—currently have legislation in place that specifically outlaws and provides statutory remedy for NCDII, with varying levels of attention to the role of online platforms.[563] The NCDII provisions of five provinces are silent on intermediary liability,[564] which suggests that platform companies could, in theory, be held liable under such legislation, so long as they met all other elements of the relevant NCDII offence (such as intent, or knowledge).[565]

Only PEI's *Intimate Images Protection Act* (IIPA)[566] explicitly addresses the issue of intermediary liability, based on a uniform statute developed by Laidlaw and Young, and adopted in December 2020 by the ULCC.[567] Specifically, PEI's IIPA states, "No application or claim may be brought against an internet intermediary if the internet intermediary has taken reasonable steps to address unlawful distribution of intimate images in the use of its services."[568]

---

[560] *Ibid*.

[561] *Ibid*, ss 3, 4, and 71.

[562] A news article from 2016 suggests that one website may have shut down in response to a letter threatening legal action based in part on claimed violation of section 22: Robert Frank, "LBPSB demanded web site delete Chinese students' complaints", *The Suburban* (21 December 2016), online: <https://www.thesuburban.com/news/west_island_news/lbpsb-demanded-web-site-delete-chinese-students-complaints/article_746722e3-04b9-5eb6-9869-c881fbeb2a3e.html>.

[563] *Intimate Images and Cyber-protection Act,* SNS 2017, c 7; *The Privacy Act*, RSS 1978, c P-24; *Intimate Image Protection Act*, CCSM c 187; *Protecting Victims of Non-consensual Distribution of Intimate Images Act*, RSA 2017, c P-26.9; *Intimate Images Protection Act*, RSPEI 1988, c I-9.1; *Intimate Images Protection Act*, RSNL 2018, c I-22.

[564] *Intimate Images and Cyber-protection Act,* SNS 2017, c 7; *Privacy Act*, RSS 1978, c P-24; *Intimate Image Protection Act*, CCSM, c 187; *Protecting Victims of Non-consensual Distribution of Intimate Images Act*, RSA 2017, c P-26.9; *Intimate Images Protection Act*, RSNL 2018, c I-22.

[565] "As for intent to distribute or publish, of the Canadian jurisdictions that have legislated in this area, only Manitoba and Alberta require such intent. The Saskatchewan, Newfoundland and Labrador and Nova Scotia statutes and the U.S. Draft Intimate Images Act tort do not. […] Existing Canadian NCDII torts tend to define 'distribution' to mean *knowing* distribution. Presumably this means something like distribution with knowledge of that image and that it is being distributed. The defamation experience suggests more clarity in the statutory language may be desirable, both as to the knowledge requirement between the plaintiff and individual defendant, and because intermediaries can be captured, depending on how widely a provision is drafted." (footnotes omitted, emphasis in original): Hilary Young & Emily Laidlaw, "Creating a Revenge Porn Tort for Canada" (2020) 96 Supreme Court Law Review 147 at 181-82.

[566] *Intimate Images Protection Act*, RSPEI 1988, c I-9.1.

[567] For more details on the UNCDII Act, see Section 4.5.1 ("*Uniform Non-Consensual Disclosure of Intimate Images Act* (2021)").

[568] *Intimate Images Protection Act*, RSPEI 1988, c I-9.1, s 5.3(1).

In terms of platform obligations, the PEI's IIPA grants judges the power to "order an internet intermediary to make every reasonable effort to remove or de-index the intimate image";[569] and Nova Scotia's *Intimate Images and Cyber Protection Act* grants judges the power to order "any person"—which, as defined in the Act, includes platform companies—to "take down or disable access to an intimate image" or certain defined instances of TFGBV.[570] Both Nova Scotia's and Manitoba's statutes empower judges to order a company to assist in identification where the perpetrator has remained anonymous.[571] Saskatchewan's *Privacy Act* and Newfoundland and Labrador's *Intimate Images Protection Act* both provide that courts may issue an injunction on any terms and conditions considered appropriate, though it is not clear if "to any person" extends beyond the defendant to include relevant platform companies.[572] Laidlaw and Young have noted, "While there is overlap between [the five provincial NCDII statutes in force prior to PEI's], there is no uniform approach. Rather than simply recommending one as a model for other Canadian jurisdictions, we believe they can all be improved."[573]

At time of writing, one reported case (as in a court decision issued and reported in a legal database)[574] grounded in any of the provincial NCDII statutes was found, in Nova Scotia.[575] Although the case involves a substantial amount of content posted by both parties on Facebook, the platform company itself was not brought into the case, nor does the decision refer to any court orders or takedown or other requests issued to Facebook.

4.5. Uniform and Model Legislation

This section of the report will examine two prominent and substantively developed Canadian law reform proposals that specifically address platform liability for harmful expression by users, which can apply directly to, or inform additional proposed legislation aimed at, TFGBV. Section 4.5.1 will discuss the Uniform Non-Consensual Disclosure of Intimate Images Act (2021), as adopted by the Uniform Law Conference of Canada (ULCC) and which is based on a tort law framework created by Emily Laidlaw and Hilary Young. Section 4.5.2 will discuss *Defamation Law in the Internet Age*, an extensive project by the Law Commission of Ontario, led by Sue Gratton, the final report of which proposes a detailed regulatory regime to govern intermediary liability in the context of Ontario defamation law.

---

[569] *Ibid*, s 5.2(2).

[570] *Intimate Images and Cyber-protection Act,* SNS 2017, c 7, s 6(2)(b).

[571] *Ibid,* s 6(2)(a); *Intimate Image Protection Act*, CCSM c 187, s 6.

[572] *The Privacy Act*, RSS 1978, c P-24, s 7.7(1)(c); *Intimate Images Protection Act*, RSNL 2018, c I-22, s 9(1)(c).

[573] Hilary Young & Emily Laidlaw, "Creating a Revenge Porn Tort for Canada" (2020) 96 Supreme Court Law Review 147 at 150. For more detailed discussion on the role and impact of digital platforms and intermediary liability with respect to NCDII in Canada, and related legal and policy considerations, see *ibid* at 150-53, 156-58, and 183-84.

[574] A second case, *Doucet v The Royal Winnipeg Ballet (The Royal Winnipeg Ballet School)*, 2019 ONSC 6982, is ongoing at time of writing. *Doucet* is a certified class action case by students at a ballet school who were sexually assaulted by an instructor and photographer at the school, and one of the common issues established for the class is whether the defendant violated subsection 11(1) of the Manitoba *Intimate Image Protection Act*, CCSM c 187 (i.e., committed NCDII).

[575] *Candelora v Feser*, 2019 NSSC 370.

4.5.1. *Uniform Non-Consensual Disclosure of Intimate Images Act (2021)*

In December 2020, the Uniform Law Conference of Canada (ULCC)[576] adopted uniform legislation for two new statutory torts to address NCDII.[577] The main substance of the *Uniform Non-Consensual Disclosure of Intimate Images Act (2021)* (UNDCII Act)[578] was developed by Laidlaw and Young, who also published their proposal in a paper, "Creating a Revenge Porn Tort for Canada".[579] That the ULCC adopted Laidlaw and Young's proposal as a uniform statute, rather than a model statute, means that the organization actively recommends that all relevant governments across Canada implement the legislation (as opposed to simply making it available for use with no position attached).[580]

The UNCDII Act makes it a tort to distribute or threaten to distribute an intimate image of someone,[581] and creates two paths of redress: a fast-track application that prioritizes harm reduction, based on strict liability, and a more traditional fault-based tort action, with associated process, defences, and damages. Crucially in the context of TFGBV, the Act allows for circumstances where consent to create or distribute an image may have been granted initially, and was later revoked after creation or distribution—such as in a situation where an intimate relationship has ended or become abusive.[582] The definition of "intimate image" also differs from that in subsection 162.1 of the *Criminal Code* and some provincial NCDII statutes in two key ways which take into account the realities of TFGBV.

First, "intimate image" is defined to include altered images.[583] This inclusion provides remedy for women targeted by deep fakes or cheap / shallow fakes,[584] since the harm inflicted is sufficiently similar to, if not precisely identical to, NCDII of the targeted individual, even if the photo or video is not authentic. Ensuring tort liability for NCDII through altered images recognizes that (as of 2018) "96% of publicly facing deepfakes were sexual deepfakes made almost exclusively of women without their

---

[576] The ULCC collaborates with lawyers and policy analysts in academia, private practice, government, public policy, and law reform to develop fully drafted Acts that are made available to be adopted, as is or with modifications, by each province and territory or federally in Canada. The aim is to provide legislation that addresses gaps or jurisdictional inconsistency in given areas of law, including modernizing or harmonizing legal approaches to a particular issue across all provinces and territories. See "What We Do" (2019), online: *Uniform Law Conference of Canada*, <https://www.ulcc.ca/en/about-us-en-gb-1/what-we-do>.

[577] Uniform Law Conference of Canada, *Uniform Non-Consensual Disclosure of Intimate Images Act (2021)*, adopted January 1, 2021, available online: <http://staging11.airwhistle.com/ULCC/media/EN-Uniform-Acts/Uniform-Non-consensual-Disclosure-of-Intimate-Images-Act-(2021).pdf>; "Professor contributes to Non-consensual Disclosure of Intimate Images tort adopted by the ULCC" (26 January 2021), online: University of Calgary <https://news.ucalgary.ca/news/professor-contributes-non-consensual-disclosure-intimate-images-tort-adopted-ulcc>.

[578] Uniform Law Conference of Canada, "Uniform Non-consensual Disclosure of Intimate Images Act (2021)" (1 January 2021), online (pdf): *Uniform Law Conference of Canada* <https://cdn-res.keymedia.com/cms/files/ca/126/0299_637504690287791552.pdf>.

[579] Hilary Young & Emily Laidlaw, "Creating a Revenge Porn Tort for Canada" (2020) Supreme Court Law Review 147.

[580] "What We Do" (2019), online: *Uniform Law Conference of Canada*, <https://www.ulcc.ca/en/about-us-en-gb-1/what-we-do>.

[581] Uniform Law Conference of Canada, *Uniform Non-Consensual Disclosure of Intimate Images Act (2021)*, adopted January 1, 2021, s 3 ("A person who distributes or threatens to distribute an intimate image commits a tort that is actionable without proof of damage.")

[582] *Ibid*, s 11; Commentary at 6 (attached to s 2) and 14 (attached to s11).

[583] *Ibid*, s 1 ("intimate image").

[584] *Ibid*, General Commentary at 2.

consent"[585]—even as much deepfake-related policy discourse tends to focus on political manipulation, national security, and electoral integrity.[586] As tools to create fake yet realistic photos and videos become ever cheaper, more available, and easier to use, women and girls will become even more vulnerable to abusers creating intimate images of them from whole cloth, or altering publicly available non-intimate images into intimate ones,[587] for the same purpose as violative distribution of intimate images that are real.

The second difference is that the two torts are available to an individual even if she is not identifiable in the intimate image (by people other than herself, whether through the person's face or other physical features, or background information in the image, such as documents or a bedroom) — so long as the targeted individual can show the court that she is the one in the image. The Working Group provided the following rationale for not limiting the definition to images where the person is identifiable, demonstrating considerations that should inform other legal reforms to address TFGBV:

> Such a narrow definition would mean certain harmful scenarios would not be captured by the torts. For example, a person takes a selfie of intimate parts of her body and shares it with a partner, who distributes it to others without consent. The person knows it is her body even if no one else knows it is her. Further, the person may live in fear that she will be identifiable at some point in the future, whether because someone pieces together it is her, or the person who posted the image identifies it as her. […]
>
> Th[e] anchoring of the cause of action [under pre-existing Canadian NCDII laws] to the concept of privacy more readily enables an interpretation of intimate image that includes non-identifiable recordings, because the right to dignity captured by privacy is most readily implicated in this type of disclosure. […]
>
> The Working Group concluded that a NCDII victim should be able to seek relief under the act without having to wait until they are identifiable and the worst damage possible is inflicted. […]
>
> Explicitly including unidentifiable persons within the scope of the act enables a cause of action for both reputational harms and invasions of privacy. It also recognizes that sexual identity and sexual objectification are at issue regardless of whether a victim is identifiable. There is no reason in principle to protect identifiable victims over /

---

[585] Suzie Dunn, "Identity Manipulation: Responding to advances in artificial intelligence and robotics" (Paper delivered at We Robot 2020, Ottawa, 2 April 2020), at 10 [unpublished].

[586] Suzie Dunn, "Women, Not Politicians, Are Targeted Most Often by Deepfake Videos" (3 March 2021), online: *Centre for International Governance Innovation* <https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos>; and "Danielle Citron on Feminism and National Security" (28 December 2019), online (podcast): *Lawfare Podcast* <https://www.lawfareblog.com/lawfare-podcast-danielle-citron-feminism-and-national-security>.

[587] See e.g., "Over 680,000 women have no idea their photos were uploaded to a bot on the messaging app Telegram to produce photo-realistic simulated nude images without their knowledge or consent, according to tech researchers. The tool allows people to create a deepfake, a computer-generated image, of a victim from a single photo." Jane Lytvynenko & Scott Lucas, "Thousands Of Women Have No Idea A Telegram Network Is Sharing Fake Nude Images Of Them" (20 October 2020), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/janelytvynenko/telegram-deepfake-nude-women-images-bot>; Jane Lytvynenko, "Writing this turned my blood cold. A Telegram bot allows men to create fake nude images of women from a single clothed photo. Over 680,000 women have been affected with about 104,000 images shared publicly, per new to research from @sensityai." (20 October 2020), online: *Twitter* <https://twitter.com/janelytv/status/1318554224310030343?lang=en>.

> unidentifiable victims since both can experience severe emotional distress from distribution of such an image.[588]

While the UNCDII Act places liability only on the person who distributed the image or made the threat, the proposed law provides that courts may "order an internet intermediary or other person or organization to make every reasonable effort to remove or de-index the intimate image".[589] The UNCDII Act deliberately does not mandate specificities of orders directed at intermediary platforms, such as setting a required time period within which content must be taken down.[590] Further, the UNCDII Act explicitly shields intermediary platforms from being sued under this law if they have "taken reasonable steps to address unlawful distribution of intimate images" on their respective platforms.[591] In considering whether to specify what constitutes "reasonable steps", the Working Group "concluded that courts are familiar with the 'reasonable steps' standard and it allows the necessary flexibility to adapt to technological change and evolving business models."[592]

The provision prohibiting direct liability for intermediaries under this tort was to prevent the law from capturing online platforms under the definition of "distribute".[593] The Working Group determined that making intermediaries liable for distribution under the UNCDII Act would raise significant freedom of expression implications while distracting from the central purpose of the legislation (i.e., providing quick and accessible relief to victims of NCDII).[594] At the same time, the Working Group concluded that "it is important that intermediaries be responsible as third parties to carry out takedown and de-indexing orders issued under the act".[595]

However, the definition of "internet intermediary" refers only to organizations with "the ordinary function of bringing together or facilitating transactions among third parties on an internet platform", and does not protect individuals who "host or index third party content".[596] This suggests that if an individual personally set up a website which primarily hosts or indexes NCDII, they would (or should) not be protected under the intermediary liability shield and could be sued under this Act, even if it was the website's users providing NCDII, and not the person owning and operating the platform.

---

[588] Uniform Law Conference of Canada, *Uniform Non-Consensual Disclosure of Intimate Images Act (2021)*, adopted January 1, 2021, Commentary at 4-5.

[589] *Ibid*, ss 4(2)(e), 5(2)(e); Commentary at 8 adds: "A court order under section 4 will allow an applicant or respondent, depending on the order, to seek takedown or de-indexing of the content directly from internet intermediaries hosting the content, by sending the order to counsel at the intermediary's corporate office."

[590] *Ibid*, Commentary at 9: "The Working Group concluded that the act should not mandate particular details to be included in injunctive orders (such as a time frame for compliance by a respondent or intermediary). It was agreed that remedial flexibility should be left to the court."

[591] *Ibid*, s 8(1).

[592] *Ibid*, Commentary at 12.

[593] *Ibid*.

[594] *Ibid*; see also Commentary at 1: "The primary objective of the act is to create an inexpensive, fast-track proceeding for victims to have NCDII removed from the internet. … The goal is to give victims what they most want: the destruction, removal or de-indexing of the intimate image as cheaply and quickly as possible."

[595] *Ibid*, Commentary at 12.

[596] *Ibid*, Commentary at 3.

4.5.2. *Defamation Law in the Internet Age*

In March 2020, the Law Commission of Ontario published the final report of a four-year long project, *Defamation Law in the Internet Age*, led and authored by Sue Gratton. The report provides a comprehensive and thorough legal and policy analysis of key issues at the intersection of Ontario defamation law and related impacts of the Internet and the rise of social media and other digital platforms. These issues include, for example, privacy and reputation, public interest and freedom of expression, implications for media and journalism, the fundamental distinction between publishers and platforms, and intermediary liability for defamatory content posted by a platform's users.

Defamation and the related wrong of impersonation are often a part of TFGBV.[597] To the extent this is the case, and to the extent that the LCO report addresses, substantively and at length, the issue of platform liability in Canadian law for a certain category of user expression that can involve speech-based abuse towards women and members of intersecting marginalized groups, *Defamation Law in the Internet Age* provides relevant context, principles, and analysis that could inform Canadian law and policy considerations regarding platform liability for TFGBV. Such insights include:

- examining—and rejecting for adoption in Canadian law—platform liability regimes in other jurisdictions in the defamation context (specifically, CDA 230 in the United States; the European Union's notice-and-takedown model; and a complex hybrid regime under the United Kingdom's *Defamation Act*);[598]
- distinguishing between privacy violations and defamation as distinct legal harms[599]—which speaks to the difficulties of attempting to apply one platform liability regime to all types of "online harms" across the board;
- establishing that digital platforms and other online intermediaries are "significantly different" from publishers such as newspapers, and should not be characterized or defined as publishers for the purposes of defamation law;[600] and
- demonstrating the complexities in weighing a matrix of overlapping and potentially conflicting legal and policy considerations against each other in the process of creating a regulatory

---

[597] See e.g. (in the context of intimate partner violence and technology-facilitated coercive control): "Abusers also sought to intimidate, harass, and humiliate women or challenge women's accounts of abuse via their networks. Jia's ex 'mentioned that he had all - like, most of my friends' contact information. She worried that her abuser would 'say bad things to my friends about me, to my friends' or impersonate her online if she refused demands to meet with him. Similarly, post-separation, Charlotte felt she had to 'engage in reputational management' after her abuser reached out to various family members and sent several text messages to her warning that he would tell everyone his version of the story. Josie's former partner stole her phone and wrote to all her Facebook friends to inform them she had left him, adding 'I don't know what happened to her. She's not mentally okay.' … Our participants indicated that abusers impersonated survivors and others online. Some reported that abusers had used their devices or logged into their social media accounts in order to impersonate them. This happened during relationships and postseparation." Molly Dragiewicz et al, *Domestic violence and communication technology: Survivor experiences of intrusion, surveillance, and identity crime* (Sydney: Australian Communications Consumer Action Network, 2019) at 24-26.

[598] Sue Gratton, "Defamation Law in the Internet Age" (March 2020) at 84, online (pdf): *Law Commission of Ontario* <https://www.lco-cdo.org/wp-content/uploads/2020/03/Defamation-Final-Report-Eng-FINAL-1.pdf>.

[599] *Ibid* at 14 ("7. Defamation Law and Privacy Law Have Distinct Objectives and Should Remain Separate").

[600] *Ibid* at 75-78.

framework that places legal obligations on platform intermediaries to uphold substantive legal protections for individuals targeted by abusive users on such platforms.[601]

However, a few caveats are in order regarding the application of defamation law generally in the context of TFGBV. Caution is necessary because defamation covers a wide range of expression, not all of which constitutes TFGBV. *Crouch v Snell*, for example, concerned online defamation but was a dispute between two businessmen and former co-founders, and involved no element of systemic oppression between them.[602] Moreover, defamation law has regularly been used by those in power to silence members of historically marginalized groups, including victims/survivors of sexual assault. This area of law may thus be a double-edged sword for historically marginalized groups, depending on the equities, power dynamics, and nature of the expression involved in a given case of alleged defamation.

First, any application of defamation law must take care not to conflate TFGBV-related defamation with other kinds of defamation that do not involve systemic oppression or historical inequity. For example, where the LCO report mentions activities that can constitute TFGBV, they are reduced to the notion of "online personal attack" and collapsed with activities such as travel reviews.[603] This is used as a broad umbrella term, but in the context of TFGBV, risks trivializing and erasing the systemic and misogynistic (and/or racist, ableist, and other structural discrimination-based) forces driving many such attacks, even if they occur on an individual basis.

Second, any application of defamation law must remain sensitive to how it has been exploited to silence victims/survivors of sexual assault and intimate partner violence and prevent future victims/survivors from speaking out.[604] Members of sociopolitically dominant groups also use defamation law to entrench types of systemic oppression that do not necessarily involve (but may intersect with) gender, such as discouraging individuals from speaking out against racism.[605] This systemic context is relevant to defamation law in the context of online platforms, as many members of historically marginalized groups would not have been able to speak out against abuses or speak truth to power and garner attention to an extent that they would be considered worth threatening with defamation claims in the first place, were it not for such platforms.[606]

---

[601] *Ibid* at 85-97, in which the LCO details its proposed platform regulation model for Ontario defamation law specifically.

[602] 2015 NSSC 340.

[603] Sue Gratton, "Defamation Law in the Internet Age" (March 2020) at 12, 55, online (pdf): *Law Commission of Ontario* <https://www.lco-cdo.org/wp-content/uploads/2020/03/Defamation-Final-Report-Eng-FINAL-1.pdf>.

[604] See e.g., "She accused a university prof of sexual assault. Now he's suing for defamation. Some fear the 'landmark' case could have a chilling effect" https://www.thestar.com/news/canada/2021/04/08/she-accused-a-university-prof-of-sexual-assault-now-hes-suing-for-defamation-some-fear-the-landmark-case-could-have-a-chilling-effect.html; Alicia Elliott, "How a Canadian Law Is Silencing Victims of Gender-Based Violence" (6 December 2018), online: *Flare* <https://www.flare.com/news/canadian-libel-law/>; "Women who speak out about their abuse online are frequently and increasingly threatened with legal proceedings, such as for defamation, which aims to prevent them from reporting their situation. Such behaviour may form part of a pattern of domestic violence and abuse." Dubravka Šimonović, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, UNHRC, 38th Sess, UN Doc A/HRC/38/47 (2018) at 31.

[605] See e.g., Katie Duke, "Calling a Racist a Racist: A Case for Reforming the Tort of Defamation" (2016) 37 Windsor Review of Legal and Social Issues 70.

[606] See e.g., Latoya A Lee, "Black Twitter: A Response to Bias in Mainstream Media" (2017) 26:6 Social Sciences; Nyx McLean, "Considering the Internet as Enabling Queer Publics/Counter Publics" (2014) Spheres; and Matthew D Luttig & Cathy J Cohen, "How social media helps young people — especially minorities and the poor — get politically engaged", *Washington*

Third, an intersectional feminist analysis of defamation law must examine the disparate role, impact, conceptualization, and imposed vulnerability or unearned resiliency of one's "reputation" and how that reputation is publicly perceived, harmed, bolstered, or protected, depending on one's gender, race, disability, sexual orientation, and class, for instance.[607] As an example, contrast the elevated concern for men's reputations and careers when sexual harassment or sexual assault claims against them are made public, with the rapidity with which women's careers are destroyed when the mere fact they have engaged in consensual sexual activity is publicized.[608]

Fourth, defamatory expression that involves attempts to weaponize the targeted person's sexuality against them should be interrogated for underlying misogynistic (as well as racist, transphobic, and/or homophobic) assumptions about female sexuality, 'reputation', and shame, while examining to what extent defamation law (and its underlying purpose of protecting 'reputation') implicitly upholds those assumptions.[609] An equality-advancing formulation of defamation law would avoid lending strength to outdated and insidious value judgements contributing to women's inequality, while at the same time still recognizing and providing redress for the substantive and material harms women experience when they are attacked on the basis of such misogynistic attitudes and beliefs, given the state of contemporary society. Alexa Dodge explains how such a conceptualization of law would operate, in the context of NCDII:

> [I]t is the breach of privacy, trust and bodily autonomy that is seen as problematic, and therefore the harm is not dependent on an understanding of sexuality or nudity as shameful. [...]
>
> If [...] we successfully change the existing beliefs that demonise sex (especially for women, sexual minorities, gender non-conforming people and other marginalised individuals), then the exposure of nude bodies and sexual activity could no longer be so easily weaponised as a tool capable of reputational ruin [...].[610]

Post (9 September 2016), online: <https://www.washingtonpost.com/news/monkey-cage/wp/2016/09/09/how-social-media-helps-young-people-especially-minorities-and-the-poor-get-politically-engaged/>.

[607] See e.g., Kimberlè Williams Crenshaw, "Beyond Racism and Misogyny: Black Feminism and 2 Live Crew" in Mari J Matsuda et al, eds, *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (New York & London: Routledge, 1993) 111 at 113, 119-20; and Safiya Noble, *Algorithms of Oppression* (New York: NYU Press, 2018).

[608] See e.g., Glenn Kauth, "Behind the headlines", *Canadian Lawyer Magazine* (4 January 2016), online: <https://www.canadianlawyermag.com/news/general/behind-the-headlines/270024>; EJ Dickson, "Meet the Paramedic Whose OnlyFans Was Outed by the 'New York Post'", Rolling Stone (17 December 2020), online: <https://www.rollingstone.com/culture/culture-news/onlyfans-medic-lauren-kwei-new-york-post-interview-1104943/>; "The Dalhousie students disciplined for sexist Facebook group are finding dentistry work, lawyer says", *National Post* (10 July 2015), online: <https://nationalpost.com/news/canada/the-dalhousie-students-disciplined-for-sexist-facebook-group-are-finding-dentistry-work-lawyer-says>; Zoe Whittall, "CanLit Has a Sexual-Harassment Problem", *Walrus* (9 March 2021), online: <https://thewalrus.ca/canlit-has-a-sexual-harassment-problem/>; Selam Jie Gano, "Remove Richard Stallman: Appendix A" (16 September 2019), online: *Medium* <https://selamjie.medium.com/remove-richard-stallman-appendix-a-a7e41e784f88>; and Melissa Jeltsen, "The Danger Of Valuing Men's Careers Over Women's Lives", Huffington Post (16 November 2018), online: <https://www.huffpost.com/entry/thousand-oaks-shooter-sexual-misconduct_n_5beb0be9e4b0caeec2beb019>.

[609] See generally Alexa Dodge, "'Try Not to be Embarrassed': A Sex Positive Analysis of Nonconsensual Pornography Case Law" (2021) Feminist Legal Studies.

[610] *Ibid* at 16 (inline citations omitted).

An effective platform liability model for TFGBV must remain much more victim/survivor-centred and sensitive, than defamation law traditionally has, to the right to equality and freedom from discrimination and the experiences of women, girls, and those at the intersections of multiple historically marginalized groups, when it comes to the role and impact of platformed TFGBV in perpetuating systemic oppression.

4.6. Systemic Approaches to Platform Liability for TFGBV

This section will discuss laws that do not expressly contemplate either platform liability or TFGBV, but could theoretically form the basis of platform liability for TFGBV, given the right circumstances. Such laws include statutory human rights law, commercial host liability, product liability, corporate criminal or tortious negligence, systemic negligence, and vicarious liability.

What differentiates these laws is that on their face, they address neither platform liability nor TFGBV specifically—let alone platform liability *for* TFGBV—and have not yet been used for that purpose in Canada at time of writing. However, these are laws which would speak to a platform company's role in perpetuating TFGBV and similar harms against other historically marginalized groups on a more institutional and systemic level. This is in contrast to the forced individualization—of an ultimately structural and societal issue—required by laws and legal remedies that focus solely on the direct perpetrator of the TFGBV in any given single case. Such individualization conceals the broader yet materially relevant systemic context and how digital platforms contribute to, create, or control that context. These laws are reviewed for their potential applicability to establishing platform liability for TFGBV, and may serve as starting points for future research, test litigation, or novel legal strategies as both TFGBV and digital platforms evolve.

First, future research and legal work should investigate the possibility of applying provincial and federal human rights statutes to digital platforms. To elaborate, if a platform is found to fit the definition of providing "goods, services, facilities or accommodation" as defined in the *Canadian Human Rights Act* or provincial equivalents, then it may be possible to find under a given set of circumstances that the platform company has denied equal access to a would-be user or group of users on the basis of a protected characteristic, such as gender or race.[611] Such an argument might draw on Sarah Jeong's observation that "[o]nline harassment makes products unusable. Harassment blows up phones with notifications, it floods inboxes, it drives users off platforms. Harassment is the app-killer."[612]

In *Ismail v British Columbia (Human Rights Tribunal)*, a restaurant patron sued the restaurant owner for violating British Columbia's *Human Rights Code*, due to the restaurant's stand-up comedian having made offensive discriminatory comments that constituted hateful speech and conduct. The BC Supreme Court stated the following in finding the restaurant owner liable:

> Ms. Pardy and her companions began the evening at Zesty on the patio. Their server then moved them inside, where the open mic night was ongoing. *It was not possible for them to enjoy the*

[611] In Quebec, for instance, platform companies would be subjected to the Quebec *Charter of Human Rights and Freedoms*, CQLR, c C-12, including its equality and non-discrimination provisions, in conjunction with the intermediary liability regime in section 22 of the *Act to establish a legal framework for information technologies*, CQLR, c C-1.1.

[612] Sarah Jeong, *The Internet of Garbage* (Vox Media, 2018) at 1377

> *normal restaurant services without being exposed to the comedy show.* I find that the comedy show *was part of the service that Zesty was providing* to the general public that night, along with food and beverages. […]
>
> The important point is that the women in *Badyal* [an analogous case] were discriminated against not in the provision of the normal pub service of food and drinks, but in the service of karaoke and dancing, which was an integral part of the pub's public service at that time. Likewise, Ms. Pardy and her friends were allegedly discriminated against by the emcee of a show, which was an integral part of Zesty's service to the public.
>
> I conclude that the Tribunal was correct in finding that the comments and actions of Mr. Earle occurred *during the provision of a service customarily available to the public*. […]
>
> In the end, this is not a case about the scope of expression in a comedy performance or an artistic performance. It is about verbal and physical abuse that amounts to adverse treatment based on sex and sexual orientation. […][613]

This case may be analogously applied to platform users who have no choice but to be exposed to TFGBV in the course of using the "normal services" of a particular platform, while the ability to engage in TFGBV itself *is also* an integral part of a platform's services for those users leveraging the platform to perpetrate it. For example, Corey Omer suggests, "An online intermediary could … probably be found civilly liable for the hateful or discriminatory speech of its users under provincial human rights statutes. Though there are no cases directly on point, the language in the provincial statutes tends to be very broad and can reasonably be read to capture online intermediaries".[614] In addition, Jane Bailey and Jacquelyn Burkell have examined the potential of applying human rights statutes to algorithmic bias,[615] which is relevant to the extent that many dominant social media platforms rely on algorithms as a core part of their content moderation and user engagement strategies, and have repeatedly demonstrated significant algorithmic bias in ways that engage the right to equality and freedom from discrimination for historically marginalized groups, including women.[616]

---

[613] *Ismail v British Columbia (Human Rights Tribunal)*, 2013 BCSC 1079 at paras 255-257, 339, and 341 (emphasis added).

[614] Corey Omer, "Intermediary Liability for Harmful Speech: Lessons from Abroad" (2014) 28:1 Harvard Journal of Law & Technology 289 at 308 (footnotes omitted).

[615] Jacquelyn Burkell & Jane Bailey, "Unlawful Distinctions?: Canadian Human Rights Law and Algorithmic Bias" (2016/2018) Canadian Yearbook of Human Rights 217.

[616] See e.g., Kevin Webb, "YouTube's algorithm is under fire for boosting a sexist conspiracy theory about black-hole researcher Katie Bouman", *Insider* (12 April 2019), online: <https://www.businessinsider.com/youtube-criticized-for-conspiracy-video-black-hole-katie-bouman-2019-4>; Salma El-Wardany, "Like Our Society, Instagram Is Biased Against Women Of Colour", *Refinery29* (10 December 2020), online: <https://www.refinery29.com/en-gb/2020/12/10150275/shadow-ban-instagram-censorship-women-of-colour>; Shirin Ghaffary, "The algorithms that detect hate speech online are biased against black people", *Recode* (15 August 2019), online: <https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter>; Karen Hao, "Facebook's ad-serving algorithm discriminates by gender and race", *MIT Technology Review* (5 April 2019), online: <https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/>; and Julia Carpenter, "Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you", *Washington Post* (6 July 2015), online: <https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/>.

Second, there may be a case to be made applying the concept of commercial host liability to digital platforms. Slane and Langlois write, regarding NCDII (in the context of advocating for limitations on CDA 230 but can apply more broadly as well):

> The reasoning is similar to other regulatory models that recognize that profit motive would otherwise encourage businesses to promote or deliberately turn a blind eye to illegal activity; examples include regulations on second-hand goods dealers and on establishments that serve alcohol to people who are likely to drive. Of course, with second-hand good dealers and bars, the business itself is legitimate, and it can carry on without supporting the illegal activity, if somewhat less lucratively—this is the sort of business that amateur porn hosts are generally part of. Put another way, safeguards against illegal activity carried out by users of a service can be imposed on businesses at a high risk of deliberately or tacitly condoning illegal activity of customers due to the profit motive; those businesses that cannot survive without this illegal activity should be shut down.[617]

Slane and Langlois appear to be applying the 'enabler' test as a threshold for liability in the above excerpt, where only businesses that "cannot survive" without the illegal activity would be captured. However, as they note, bars are legitimate establishments that are subject to commercial host liability due to the moral hazards and business incentives otherwise involved in their ventures. Therefore, while an 'enabler' provision may cover purpose-built TFGBV-dedicated platforms, platforms of general application, which are venues for a wide range of beneficial, legitimate, and legal expression, could be considered closer to bars being required to refrain from serving more to those already intoxicated by hate speech and violent extremism towards women and other systemically oppressed groups.

Third, certain functions (or dysfunctional aspects) of a given digital platform might be characterized as a "dangerous product" under product liability law. This was the legal strategy attempted in the case of *Herrick v Grindr*, discussed later in Section 5.1.3 ("*Matthew Herrick v Grindr LLC*"). The Citizen Lab has also pointed out that product liability, additionally in the form of class action proceedings, may apply to stalkerware apps.[618] However, stalkerware apps may be distinguished from digital platforms on the basis of being more easily discernible as a concrete "product" with a relatively clear isolated purpose.

Fourth, digital platforms might be held liable for corporate negligence, criminal negligence, or systemic negligence, for systemic harms done to historically marginalized groups, based either in tort law or the *Criminal Code* provisions set out above.[619] The following excerpt, for example, discusses applying negligence to platform-facilitated algorithmic harms to historically marginalized groups:

> According to Ruparelia, compared to intentional torts, "Negligence is better situated to respond to the reality that racism is not a series of discrete actions but rather 'an integrated system that

---

[617] Andrea Slane & Ganaele Langlois, "Debunking the Myth of 'Not My Bad': Sexual Images, Consent, and Online Host Responsibilities in Canada" (2018) 30:1 Canadian Journal of Women and the Law 42 at 62-63.

[618] Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (21 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment>. FSG at pages 77-81.

[619] See Cynthia Khoo, "Missing the Unintended Forest despite the Deliberately Planted Trees: Reasonable Foreseeability and Legal Recognition of Platform Algorithm-Facilitated Emergent Systemic Harm to Marginalized Communities" at 51-60 (Paper delivered at We Robot 2020, Ottawa, ON, 22 September 2020) [unpublished].

> elevates one group at the expense of another'." Negligence not only aligns with the focus on impact regardless of intent, as required in human rights law, but Ruparelia's description of negligence also fits situations that give rise to platform-facilitated emergent systemic harms to marginalized groups: such harms reflect the reality of adverse effects discrimination across online platforms, arising not from overt actions, but from a system of integrated components giving rise to conditions that disproportionately harm a given marginalized group in particular, if not exclusively.[620]

Fifth and last, future scholarship or litigation may find it worthwhile to consider additional laws of general application that may apply to digital platforms as corporate entities and employers, such as vicarious liability. This may be especially relevant where specific managers or senior executives have been known to have made conscious decisions against implementing mechanisms that would have reduced TFGBV and other forms of hate speech across their respective platforms.[621]

---

[620] *Ibid* at 54-55 (footnotes omitted).

[621] See e.g., Mark Bergen, "YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant", *Bloomberg* (2 April 2019), online: <www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warningsletting- toxic-videos-run-rampant>; and Jeff Horwitz & Deepa Seetharaman, "Facebook Executives Shut Down Efforts to Make the Site Less Divisive", *Wall Street Journal* (26 May 2020), online: <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>.

5. Platform Liability Models: Jurisdictional Scan

Part 5 of the report reviews platform liability models that have been implemented or proposed in six jurisdictions around the world, including their impacts and criticisms from relevant stakeholders. Some of the discussed legal frameworks do not focus specifically on technology-facilitated gender-based violence, abuse, and harassment (collectively, TFGBV). Rather, they address harmful user expression and behaviour generally speaking, illegal activity by users across the board, or in one case, focus on hate speech based only on "race, colour, religion, descent or national or ethnic origin" (which overlaps with, but neither wholly captures nor is wholly captured by TFGBV). However, examining what other jurisdictions have implemented or proposed, and to what effect, is valuable to informing potential legal reforms in Canada regarding platform liability for TFGBV.

Section 5.1 examines laws and proposals in the United States, specifically section 230 of the *Communications Decency Act* (CDA 230), proposed amendments to CDA 230, the *Allow States and Victims to Fight Online Sex Trafficking Act* (FOSTA), and a test case attempting to circumvent CDA 230, *Matthew Herrick v Grindr LLC*. Section 5.2 discusses the Netzwerkdurchsetzungsgesetz ('Network Enforcement Act', or NetzDG) in Germany. Section 5.3 discusses the United Kingdom's proposed Online Harms bill, based on the UK government's *Online Harms White Paper* and its initial and final responses to public consultation on the White Paper. Section 5.4 reviews enacted and proposed intermediary liability laws and instruments in the European Union, specifically: Articles 12-15 of the E-Commerce Directive, the Code of Conduct on Countering Illegal Hate Speech Online, the Communication and Recommendation on Tackling Illegal Content Online, and the proposed *Digital Services Act*. Section 5.5 discusses the *Enhancing Online Safety Act 2015* and forthcoming legislative reforms in Australia, as well as the *Sharing of Violent Abhorrent Material Act 2019*. Section 5.6 discusses the *Harmful Digital Communications Act 2015* and the Christchurch Call in New Zealand.

5.1. United States

The United States (US) might be considered to have had outsized influence over platform regulation issues around the world, in part due to many of the most dominant digital platforms in Western countries being based in the US, and in part due to the corresponding prominence of its intermediary liability laws, such as CDA 230 and section 512 of the *Digital Millennium Copyright Act* (DMCA), in addition to the US having actively pushed its approaches to platform liability onto other jurisdictions through trade agreements.[622] However, the United States has also begun departing, over recent years, from steadfast adherence to the broad liability shield that CDA 230 established for online platforms.

In addition, TFGBV is part of the "Biden Plan to End Violence Against Women", which was announced in November 2019 during President Joe Biden's candidacy. The plan includes establishing a National Task

[622] See Issue Spotlight No. 1 ("Copyright, Intermediary Liability, and Safeguarding Human Rights in Context").

Force on Online Harassment and Abuse,[623] which will study "rampant online sexual harassment, stalking, and threats, including revenge porn, deepfakes, and the connection between this harassment, mass shootings, extremism and violence against women".[624] The plan further states that the "Task Force will consider platform accountability, transparent reporting requirements for incidents of harassment and response, and best practices."[625] TFGBV experts have expressed encouragement at recognition of the issue, but maintain reservations regarding "how this plan will look in action, and how it will actually affect lives online and off".[626]

This section of the report will review CDA 230 and proposed reforms to the law; a specific statute that established an exemption from CDA 230 protection—the *Allow States and Victims to Fight Online Sex Trafficking Act* (known as FOSTA-SESTA)—and a test litigation case that attempted to circumvent CDA 230, *Matthew Herrick v Grindr LLC.* One specific proposed reform to CDA 230, by Danielle Citron and Benjamin Wittes, is rooted in Citron's extensive work focusing specifically on TFGBV.

5.1.1. Section 230 of the *Communications Decency Act*

The United States provides the highest standard and broadest scope of immunity to digital platforms for civil liability, due to section 230 of the CDA (CDA 230). CDA 230 shields online intermediaries from civil liability in two ways. First, and most famously, it exempts intermediaries—such as digital platforms—from being legally treated as if they themselves are the "publisher or speaker" of information provided by another person or entity using the intermediary's platform. The effect of this is to exempt online platforms from nearly all forms of civil liability for the words or actions of their users, and this exemption applies across all areas of law where *civil* liability is concerned.[627] Notably, CDA 230 does not apply to federal criminal liability and would not shield platforms where they would otherwise be liable under federal criminal laws, for hosting or facilitating certain kinds of illegal content or activities.[628] CDA 230 also does not apply to intellectual property cases, which are instead governed by a separate intermediary liability regime in the US *Digital Millennium Copyright Act* (DMCA).[629]

The second prong of CDA 230's protection for online intermediaries is particularly relevant to online platforms' ability to address TFGBV. Paragraph 230(c)(2) shields intermediaries from civil liability for *engaging in active moderation of user content*. This provision specifically enables online platforms to

---

[623] "The Biden Plan to End Violence Against Women", online: *Joe Biden for President: Official Campaign Website* <https://joebiden.com/vawa/#>.

[624] *Ibid*.

[625] *Ibid*.

[626] Samantha Cole, "Biden Has a Plan to Tackle Online Harassment. What Does It Actually Say?" (12 November 2020), online: *Vice* <https://www.vice.com/en/article/epdnjp/biden-plan-to-end-violence-against-women-online-harassment-what-does-it-say>.

[627] *Communications Decency Act*, 47 USC § 230 (1996) at section 230(e)(2). The sole exception is intellectual property law, which is governed by its own intermediary liability regime in section 512 of the U.S. *Digital Millennium Copyright Act*. DMCA, s 512. As of 2019, there is now an additional exception carved out by *SESTA/FOSTA* to impose platform liability in the case of content related to sex trafficking. However, this law has come under much criticism, including from sex workers, victims of sex trafficking, and human rights advocates, and will be discussed later in this report.

[628] *Communications Decency Act*, 47 USC § 230 (1996) at section 230(e)(1).

[629] *Digital Millennium Copyright Act*, 17 USC § 512 (1998) at section 512.

"restrict access to or availability of material that the provider […] considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected", without fear of legal repercussions resulting from their decisions made in the course of such moderation. Without this protection—but with near-certain immunity attached to a completely hands-off approach—online platforms would be disincentivized from engaging in even basic management of user content, for fear that such efforts would trigger the full suite of laws associated with being a traditional publisher rather than an Internet intermediary.[630]

It has proven virtually impossible to hold platform companies liable for user acts outside of copyright infringement. For example, US courts have ruled that even issuing an injunction to a platform company, in a case involving content posted by their user(s), would violate CDA 230 and is thus not legally possible.[631] This is especially the case with hate speech and speech-based abuse (that does not additionally violate privacy or other legal rights) due to US law and ingrained norms involving the First Amendment, which frequently shields even the actual speaker, before combining it with the strength of the CDA 230 shield for the platform.

Since the 2016 US presidential election, more and more actors across the entire political spectrum and throughout different sectors of society have increasingly called for reforms to CDA 230.[632] In addition to test litigation,[633] grassroots activism, and academic scholarship supporting various reforms, proposals to reform or restrict the effects of CDA 230 have included the introduction of several bills and an executive order by the 45th US president.[634] However, all of these initiatives have varied dramatically in rationale, motivation, advisability, legitimacy, and legal validity. Some reform proposals have been considered to be political posturing, or introduced in order to heighten fear of regulation and elicit more favourable treatment from platforms in their treatment of politically conservative content. Many of

---

[630] This is in fact what occurred in *Stratton Oakmont Inc v Prodigy Services Co,* 23 Media L Rep (BNA) (NY Sup Ct May 24, 1995). In this case, Prodigy hosted user bulletin boards which it managed through creating Content Guidelines, using software to automatically screen out offensive language, and delegating enforcement of the Guidelines to "Board Leaders". The plaintiff sued Prodigy for defamatory content posted by a user, and the court held that due to the above community management activities, Prodigy exercised "editorial control" over the bulletin boards, thus "render[ing] it a publisher with the same responsibilities as a newspaper." This outcome, combined with another court coming to the opposite conclusion on a similar set of facts in *Cubby, Inc v CompuServe Inc*, 776 F Supp 135 (SDNY 1991) (finding CompuServe *not* liable as a publisher for user content), is what eventually led to the enactment of CDA 230, including the moderation shield in paragraph 230(c)(2).

[631] See e.g., *Hassell v Bird*, Cal Rptr 3d 867 (S Ct Cal 2018), and the United States District Court for the Northern District of California holding *Google Inc v Equustek Solutions Inc*, 2017 SCC 34, unenforceable by way of *Communications Decency Act*, 47 USC § 230 (1996); and Vivek Krishnamurthy and Jessica Fjeld, "CDA 230 Goes North American? Examining the Impacts of the USMCA's Intermediary Liability Provisions in Canada and the United States" (2020) at 8, 19, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3645462>.

[632] For a tracker of all legislative reform proposals concerning CDA 230 (updated to March 2021 at time of writing), see Kiran Jeevanjee et al, "All the Ways Congress Wants to Change Section 230", *Slate* (23 March 2021) online: <https://slate.com/technology/2021/03/section-230-reform-legislative-tracker.html>.

[633] See e.g., *Herrick v. Grindr LLC*, 765 F App'x 586 (2d Cir 2019).

[634] Kiran Jeevanjee et al, "All the Ways Congress Wants to Change Section 230", *Slate* (23 March 2021) online:

<https://slate.com/technology/2021/03/section-230-reform-legislative-tracker.html>; Tony Romm and Elizabeth Dwoskin, "Trump signs order that could punish social media companies for how they police content, drawing criticism and doubts of legality", *Washington Post* (18 May 2020), online: <https://www.washingtonpost.com/technology/2020/05/28/trump-social-media-executive-order/>.

these proposals have been heavily criticized by platform liability experts and legal scholars,[635] including leading experts at the intersection of TFGBV and intermediary liability.[636]

While it is beyond the scope of this report to review each of these proposals, some examples are provided below to demonstrate the extent, variety, and nature of interest in amending or curtailing the scope of CDA 230 protection for digital platforms. Additionally, Section 5.1.4 below will delve into one specific proposal in more detail, put forward by Danielle Citron and Benjamin Wittes.

At the outset, many CDA 230 reform proposals appear to have been driven by some level of Republican partisanship, and have generally involved proposals to remove legal protection from platforms for certain perceived actions or omissions in their content moderation policies and decisions, such as engaging in alleged 'anti-conservative bias', failing to maintain political 'neutrality', or engaging in alleged political 'censorship'. For example, the above-mentioned US presidential executive order "directs the [US] Commerce Department to ask the F.C.C. to develop regulations addressing whether social media firms lose [CDA 230] immunity if they restrict access to posted material in bad faith".[637] This executive order was issued in direct retaliation against Twitter for fact-checking the then-president's tweets. Various legislative proposals have also required that a platform engage (or promise to engage) in 'good faith' and 'unbiased' moderation, on pain of losing CDA 230 protection.[638]

Other CDA 230 reform proposals have targeted specific issues or types of content, rather than a platform's overall approach to content moderation. For example, one senator's white paper advocated the removal of CDA 230 immunity "for failure to take down deep fake or other manipulated audio/video content".[639] Another proposal suggested stripping CDA 230 immunity for 30 days if a platform engaged in behavioural advertising.[640] During the lead-up to the 2020 US presidential election, Democratic presidential candidates advocated repealing CDA 230 altogether, or withdrawing its protection where platforms do not sufficiently moderate "speech that 'incite[s] or engage[s] in violence, intimidation, harassment, threats, or defamation'" against marginalized groups.[641] The Department of Justice recommended carve-outs of CDA 230's shield for "truly bad actors" such as platforms that "purposefully facilitate or solicit third-party content or activity that would violate federal criminal law" and for "claims that address particularly egregious content".[642]

---

[635] See e.g., Emma Llansó and Mana Azarmi, "CDT Leads Coalition in Opposition to The Online Content Policy Modernization Act, S.4632" (30 September 2020) online: *Center for Democracy and Technology* <https://cdt.org/insights/cdt-leads-coalition-in-opposition-to-the-online-content-policy-modernization-act-s-4632/>.

[636] See e.g., Danielle Citron, "(1) Long threat [sic] alert: yesterday, I laid low about the EO [Executive Order] on "online censorship" because I didn't want to be lured into, and be trolled by, absurd distortion of Section 230 and ultra vires nonsense. I have written about and worked on 230 reform for too long and too hard." (20 May 2020 at 10:52) online: *Twitter* <https://twitter.com/daniellecitron/status/1266381845735899141>.

[637] Charlie Savage, "Trump's Order Targeting Social Media Sites, Explained", *New York Times* (28 May 2020), online: <https://www.nytimes.com/2020/05/28/us/politics/trump-twitter-explained.html>.

[638] Zoe Bedell and John Major, "What's Next for Section 230? A Roundup of Proposals" (29 July 2020), online: *Lawfare* <https://www.lawfareblog.com/whats-next-section-230-roundup-proposals>.

[639] *Ibid.*

[640] *Ibid.*

[641] *Ibid.*

[642] *Ibid.*

Some legislative initiatives appear to have taken advantage of the ubiquitous fervour for CDA 230 reform in order to advance other major policy agendas. For example, a bill proposing the *Eliminating Abusive and Rampant Neglect of Interactive Technologies Act* (the "EARN IT Act") ostensibly targeted CSAM. However, it was widely criticized as actually being an attack on encryption and by extension the right to privacy. This was because the original text of the bill stipulated that platform companies would lose CDA 230 protection if they implemented end-to-end encryption on their platforms.[643]

Interest in reforming CDA 230 goes beyond a federal concern. In a bipartisan letter, 47 state attorneys general advocated for exempting state criminal law from CDA 230, as currently only federal criminal law is exempt, in addition to intellectual property law.[644]

5.1.2. *Allow States and Victims to Fight Online Sex Trafficking Act*

In 2017, two bills were introduced in the US House of Representatives and the US Senate, respectively, proposing to create the first major statutory exemption in twenty years to CDA 230's liability shield for online intermediaries. The Senate bill—the *Stop Enabling Sex Traffickers Act* (SESTA)—was ultimately incorporated into the House bill— the *Allow States and Victims to Fight Online Sex Trafficking Act* (FOSTA)—and enacted as FOSTA-SESTA in April 2018.[645] The legislation garnered significant prominence and controversy, in part due to its unprecedented limitation of broad civil immunity for platforms under CDA 230, in part due to its potential to set a significant precedent for further limitations on CDA 230, and in part due to strong vocal public support or opposition from victims and survivors of sex trafficking, sex workers, platform regulation and TFGBV experts, digital rights advocates, and major Internet companies, among other groups.[646]

Under FOSTA-SESTA, intermediary platform companies lose protection from civil liability if they "knowingly assist, support, or facilitate advertising activity that violates federal sex-trafficking law".[647] Additional provisions allow for state civil and criminal liability under certain circumstances.[648]

Although ostensibly proposed and passed with the objective of addressing platform-facilitated sexual exploitation, evidence has accumulated regarding the dangerous consequences of FOSTA-SESTA for sex workers. For example, a report from Hacking//Hustling on the law's impact stated:

---

[643] Lily Hay Newman, "The EARN IT Act Is a Sneak Attack on Encryption" (3 May 2020), online: *Wired* <https://www.wired.com/story/earn-it-act-sneak-attack-on-encryption/>.

[644] Matt Zimmerman, "State AGs Ask Congress to Gut Critical CDA 230 Online Speech Protections" (24 July 2013) online: *Electronic Frontier Foundation* <https://www.eff.org/deeplinks/2013/07/state-ags-threaten-gut-cda-230-speech-protections>.

[645] Danielle Citron and Quinta Jurecic, "FOSTA: The New Anti-Sex-Trafficking Legislation May Not End the Internet, But It's Not Good Law Either" (28 March 2018), online: *Lawfare* <https://www.lawfareblog.com/fosta-new-anti-sex-trafficking-legislation-may-not-end-internet-its-not-good-law-either>.

[646] *Ibid.*

[647] *Ibid.*

[648] Caitlyn Burnitis, "Facing the Future with FOSTA: Examining the Allow States and Victims to Fight Online Sex Trafficking Act of 2017" (2020) 10 University of Miami Race & Social Justice Law Review 139 at 151.

> The current legal changes of FOSTA-SESTA alter the structure of the web, disproportionately harming the physical safety and mental health of already vulnerable communities who have most benefited from using an internet based work model. [...]
>
> Comparing our initial data of online workers with that of WCIIA [Whose Corner is it Anyway], shows that those who are already being heavily policed on the streets do not feel the same immediate effects of FOSTA-SESTA. The street-based respondents from WCIIA already exist in a more heavily criminalized and policed economy. What FOSTA-SESTA did was push workers who had access to harm reduction working tools into less safe work environments, increasing their financial insecurity and exposure to violence.[649]

FOSTA-SESTA has not only had immediate and dire impacts for sex workers who relied on digital platforms to engage in harm reduction techniques.[650] It has also damaged historically marginalized communities by incentivizing Internet-wide prohibitions, bans, and automated takedowns in spaces created by and for historically marginalized users. As the following examples will demonstrate, expression impacted by FOSTA-SESTA includes 2SLGBTQQIA content; sex education; creative writing and art considered too risqué for more mainstream platforms; personal reflections in a 'safe space' relative to platforms like Facebook and Twitter; and consensual sexual content featuring under-represented communities, which subvert or reject sexist and heteronormative representations of sexuality perpetuated in popular media and mainstream pornography.

The most prominent example of such fallout was Tumblr's site-wide ban on 'adult content', which it implemented notwithstanding its historically unique "social, judgment-free culture, which many [users] cited as helping them understand their sexual orientation".[651] Carolyn Bronstein describes the deep sense of loss that accompanied the platform's announcement:

> The announcement sent shock waves through user groups for whom Tumblr had been a hub for curated, sex-positive and body-positive content, beloved for its anything-goes permissiveness and relaxed content moderation. [...] The outpouring of grief over the

---

[649] Danielle Blunt & Ariel Wolf, "Erased: The Impact of FOSTA-SESTA & the Removal of Backpage" (2020), at 42, online (pdf): *Hacking//Hustling* <https://hackinghustling.org/wp-content/uploads/2020/02/Erased_Updated.pdf>. See also Lura Chamberlain, "FOSTA: A Hostile Law with a Human Cost" (2019) 87:5 Fordham Law Review 2171.

[650] "Of those who utilized web-based harm reduction techniques, the most common tools used were sites dedicated to reviewing clients in an effort to flag those that with a history of violence, non-payment, or potential connections to law enforcement. Commonly known as "bad-date lists," sites such as these can fall within the vague parameters of what FOSTA-SESTA criminalizes. Another tool used by sex workers is a system of verification in which a new client gives the contact information of past providers to vouch for themselves. VerifyHim is just one example of the harm reduction tools that have been taken down after FOSTA-SESTA." Danielle Blunt & Ariel Wolf, "Erased: The Impact of FOSTA-SESTA & the Removal of Backpage" (2020) at 21, online (pdf): *Hacking//Hustling* <https://hackinghustling.org/wp-content/uploads/2020/02/Erased_Updated.pdf>.

[651] Paris Martineau, "Tumblr's Porn Ban Reveals Who Controls What We See Online" (12 April 2020), online: *Wired* <https://www.wired.com/story/tumblrs-porn-ban-reveals-controls-we-see-online/>.

> demise of Tumblr as a utopian environment and its reconfiguration as a sanitized corporate space was significant.[652]

Cookie Cyboid wrote the following about the impact on 2SLGBTQQIA communities in particular:

> A lot of queer communities connect online, and because our existence is seen as inherently sexual to some we can expect policies that limit sexual expression to hit queer people much harder. It's difficult to realize certain things about yourself as a queer person without the internet, and sex education for gay, lesbian, and trans people is severely lacking without the internet. I really fear for the younger generations of queer people growing up in a world where talking about sex online gets you banned.[653]

The Tumblr adult content ban disproportionately impacted the trans community as well:

> The emotional assault was visceral, similar to the feeling of seeing one's house torn down or being evicted. [...] The adult content ban erased lives and bodies, and a space for honest conversations about sexuality, decimating hardwon and validating digital communities. For many trans users, Tumblr was the primary safe space to talk about trans sexual practices and trans pleasure in a frank and open way.[654]

Helen Holmes further warns, "[A]s sex and sexuality is slowly bled from all corners of the 'respectable' internet, as evidenced by ongoing efforts to deprive sex workers of online platforms, art will continue to suffer from increased regulation as well."[655] Artists who do not voluntarily "adapt their works to fit ever-more-conservative major platform standards" have woken up to accounts deleted overnight, despite relying on them as primary sources of income and community, professional and otherwise.[656]

Risk-averse platform crackdowns also capture sex education and sexual health information. For example, comic artist Erika Moen faced takedowns and shadowbanning on platforms such as Twitter and Instagram for her work, which "provides queer-affirming sex education on a wide range of topics, from sexual how-tos to guidance on physical health".[657] As a result of this impact multiplied across all major Internet platforms, combined with creators self-censoring to the point of being hidden from their own audiences in attempts to avoid deletion, "stories and useful information about sexual health, work, and lifestyles are becoming increasingly hard to find".[658]

---

[652] Carolyn Bronstein, "Pornography, Trans Visibility, and the Demise of Tumblr" (2020) 7:2 Transgender Studies Quarterly 240 at 242.

[653] Cookie Cyboid, "Want To Know Why Tumblr Is Cracking Down On Sex? Look To FOSTA/SESTA" (25 December 2018), online: *Medium* <https://medium.com/the-establishment/want-to-know-why-tumblr-is-cracking-down-on-sex-look-to-fosta-sesta-15c4174944a6>. See also Marianne Eloise, "without nsfw content there is no tumblr" (6 December 2018), online: *Vice* <https://i-d.vice.com/en_uk/article/59vkba/without-nsfw-content-there-is-no-tumblr>.

[654] Carolyn Bronstein, "Pornography, Trans Visibility, and the Demise of Tumblr" (2020) 7:2 Transgender Studies Quarterly 240 at 241-42.

[655] Helen Holmes, "'First They Come for Sex Workers, Then They Come for Everyone,' Including Artists", *Observer* (27 January 2021), online: <https://observer.com/2021/01/first-they-come-for-sex-workers-then-they-come-for-everyone-including-artists/>.

[656] *Ibid*.

[657] *Ibid*.

[658] *Ibid*.

At the same time, abusive content such as TFGBV targeting precisely these same marginalized artists, creators, and users continues to flourish as a result of other (voluntary) content moderation policies and decisions from these same platform companies, as described in Part 3 ("Role of Digital Platforms in TFGBV"). Thus, members of historically marginalized groups are caught in an Internet-wide vise between their own expression being over-removed, while expression abusing and silencing them is under-removed. Laws that do not centre the lived experiences, insights, and expertise of those most impacted by TFGBV only exacerbate this double-bind.

5.1.3. *Matthew Herrick v Grindr LLC*

In *Matthew Herrick v Grindr LLC*, a user of the gay dating app Grindr sought to hold the platform liable for facilitating another user, his former partner, engaging in significant and sustained abuse towards him. For instance, the former partner impersonated the plaintiff on Grindr and continually sent men to his home and his workplace with the understanding that the plaintiff had invited them over for sex—approximately 1400 men within ten months.[659]

Upon sustained inaction from Grindr, despite multiple requests for assistance, the plaintiff attempted to sue the app company on grounds of product liability and negligence, as a deliberate attempt to circumvent CDA 230.[660] Counsel for the plaintiff explained their argument as follows:

> Grindr is a defectively designed and manufactured product insofar as it was easily exploited—presumably by spoofing apps available from Google and Apple—and didn't have the ability, according to the courtroom admissions of Grindr's own lawyers, to identify and exclude abusive users. For a company that served millions of people globally and used geolocating technology to direct those people into offline encounters, it was an arithmetic certainty that at least some of the time the product would be used by abusers, stalkers, predators and rapists. Failing to manufacture the product with safeguards for those inevitabilities [...] was negligent.[661]

The plaintiff's arguments focused on Grindr's own features and operations to attempt to avoid the application of CDA 230, which would shield the platform from liability arising from the words and actions of the former partner.

Grindr succeeded in having the case dismissed in its entirety by a district court, on the basis that the plaintiff's case would violate CDA 230. The US Court of Appeals for the Second Circuit affirmed the district court's decision. Regarding the product liability and negligence claims, the court held:[662]

---

[659] Carrie Goldberg, "*Herrick v. Grindr*: Why Section 230 of the Communications Decency Act Must be Fixed" (14 August 2019), online: *Lawfare* <https://www.lawfareblog.com/herrick-v-grindr-why-section-230-communications-decency-act-must-be-fixed>.

[660] *Ibid*.

[661] *Ibid.*

[662] *Herrick v. Grindr LLC*, 765 F App'x 586 (2d Cir 2019).

- a platform company's decisions regarding product design and safety features, which impact what users can do or say on a platform, themselves constitute choices about user content, and are thus protected by CDA 230;[663]
- Grindr was not liable for "failure to warn" because that claim was considered "inextricably linked to Grindr's alleged failure to edit, monitor, or remove the offensive content provided by [the plaintiff's] ex-boyfriend" and thus the claim is blocked by CDA 230;[664]
- the failure to warn claim also lacks causation because "any purported failure to warn Herrick when he first downloaded Grindr in 2011 is unrelated to his ex-boyfriend's subsequent use of the app";[665] and
- claims of negligence, intentional infliction of emotional distress, and negligent infliction of emotional distress all relied on "Grindr's allegedly inadequate response to Herrick's complaints [and are thus] barred because they seek to hold Grindr liable for its exercise of a publisher's traditional editorial functions".[666]

The Supreme Court of the United States declined to hear an appeal.[667] Although ultimately unsuccessful, *Herrick v Grindr* represents a high-profile and novel attempt in the United States to hold an online platform liable for user behaviours that would constitute TFGBV. Similar cases may be decided differently in other jurisdictions, given the lack of CDA 230 and associated impermeable jurisprudence, as well as variances in legislation, precedents, legal principles, and underlying values.

5.1.4. Citron and Wittes CDA 230 Reform Proposal

Danielle Citron and Benjamin Wittes have advanced a notable proposal for CDA 230 reform, building on a long line of Citron's work addressing intermediary liability and TFGBV. Their recommendations aim to narrow the sweeping scope of CDA 230 in the United States, in order to address TFGBV and other violative user acts on platforms, while preserving the traditional benefits of CDA 230, such as freedom of expression (and indeed promoting this value, where moderating hate speech is concerned). Citron and Wittes write that legal immunity for websites such as 'The Dirty' run exactly contrary to what US legislators intended to achieve when they enacted the CDA in 1996.[668] This has been in large part due to US courts' interpretation of CDA 230, rather than strictly the text of the provision itself:

> [T]he broad construction of CDA's immunity provision adopted by the courts has produced an immunity from liability far more sweeping than anything the law's words, context, and history support. Platforms have been protected from liability even though they republished content knowing it might violate the law, encouraged users to post

[663] *Ibid.*

[664] *Ibid.*

[665] *Ibid.*

[666] *Ibid.*

[667] Alexis Kramer, "Grindr Harassment Case Won't Get Supreme Court Review", *Bloomberg Law* (7 October 2019), online: <https://news.bloomberglaw.com/us-law-week/grindr-harassment-case-wont-get-supreme-court-review>.

[668] Danielle Keats Citron and Benjamin Wittes, "The Problem Isn't Just Backpage: Revising Section 230 Immunity" (2018) 2 Georgetown Law Technology Review 453 at 453-54 (footnotes omitted).

> illegal content, changed their design and policies to enable illegal activity, or sold dangerous products. As a result, hundreds of decisions have extended Section 230 immunity, with comparatively few denying or restricting it.[669]

Commenting on CDA 230's status as "a kind of sacred cow—an untouchable protection of near constitutional status",[670] Citron and Wittes emphasize, with particular relevance to TFGBV, that the "free expression calculus devised by the law's supporters often fails to consider the loss of voices in the wake of destructive harassment encouraged or tolerated by platforms. We suspect that the many benefits the immunity has enabled could have been secured at a slightly lesser price."[671] The authors assert that given the long line of jurisprudence entrenching the near insurmountable barrier that CDA 230 has posed to platform liability in even the most egregious of cases, legislative amendment is the most likely possible avenue of reform.[672] Such legislative reform is necessary to update a law passed when "it was impossible to foresee the threat to speech imposed by cyber mobs and individual harassers, whose abuse chills the speech of those unwilling to subject themselves to further damage. Then, the aggregative power of the Internet was not yet known. […] The potential for destruction [of victim's lives and well-being] is exponentially greater today than it was twenty years ago."[673]

Citron and Wittes propose a legal framework meant to "bring [CDA 230's] expressive and other costs into view along with its benefits so that courts can recalibrate the interpretative lens of the CDA's safe harbor."[674] This involves, first, making certain modifications to judicial interpretation and application of CDA 230, and second, a legislative proposal.

First, "courts should not apply Section 230's safe harbor unless the claims relate to the publication of user-generated content."[675] This serves as an important reminder to ensure that an implicated platform's owners and/or operators were not, in fact, involved in the creation or development of the harmful content, which would push them out of the category of being passive intermediaries and no longer entitled to safe harbour, as well as to ensure that the claims are about the content itself, and not the platform's own actions.[676] Citron and Wittes state, "Liability for aiding and abetting others' wrongful acts does not depend on the manner in which aid was provided. Designing a site to enable defamation

---

[669] *Ibid* at 460 (footnotes omitted).

[670] *Ibid* at 461.

[671] *Ibid* at 461.

[672] "It is not inevitable that society suffers these harmful consequences in exchange for a legal environment that fosters speech and innovation. This exchange is a choice—and it's a bad choice. Ideally, since Section 230 does not actually compel this exchange, the solution would be for courts to interpret Section 230 in a manner more consistent with its text, context, and history. This interpretation would go a long way to incentivize efforts to deter illegal material, which is what the CDA's drafters set out to do in the first place. However, this solution is probably a long-shot given the judiciary's current understanding of the law. If this assessment is correct, the only course is a potential statutory fix. We suggest a course correction for the courts and, if needed, a modest statutory change that would help reorient the current liability environment." *Ibid* at 467.

[673] *Ibid* at 463.

[674] *Ibid* at 465.

[675] *Ibid* at 467.

[676] For example, in one case, "The Ninth Circuit rejected the Section 230 defense because the defendant [platform] was not being sued for publishing third-party content. Instead, the lawsuit centered on defendant's failure to warn plaintiff about the rape scheme [despite the platform's owner knowing about it], not its failure to edit or remove content." *Ibid* at 468.

or sex trafficking could result in liability in the absence of a finding that a site was being sued for publishing or speaking."[677]

Citron and Wittes further point out that permitting websites such as 'The Dirty' to benefit from absolute protection under CDA 230 alone requires misguidedly interpreting "a provision enacted to encourage providers to take steps to *restrict* abusive material to shield them from liability for *encouraging* such material. This interpretation undermines the congressional goal of incentivizing self-regulation."[678] Thus, courts should interpret CDA 230 (and, one could extrapolate, similar laws in peer jurisdictions) to exclude "online service providers that knowingly traffic in, or solicit, illegal activity [thus] eliminating incentives for better behavior by those in the best position to minimize harm."[679]

Second, Citron and Wittes suggest two potential ways to amend the text of CDA 230, which would explicitly eliminate protection for "the worst actors" on the theory that "sites that encourage destructive online abuse or which are principally used for that purpose should not enjoy immunity from liability."[680] The first is to expand the pre-existing exemption for federal criminal law and intellectual property offences to gender-based violence and abuse, by amending the provision to read:

> Nothing in section 230 shall be construed to limit or expand the application of civil or criminal liability for any website or other content host that purposefully encourages cyber stalking, nonconsensual pornography, sex trafficking, child sexual exploitation, or that principally hosts such material.[681]

The second suggestion is to amend paragraph 230(c)(1), which defines the entities to which CDA 230 applies, by attaching conditions to obtaining CDA 230 protection. The amended text would state:

> No provider or user of an interactive computer service that *takes reasonable steps to prevent or address unlawful uses of its services once warned about such uses* shall be treated as the publisher or speaker of any information provided by another information content provider *in any action arising out of the publication of content provided by that information content provider.*[682]

The first proposal resembles a version of the 'enabler' provision in the Canadian *Copyright Act*,[683] rewritten to address instead TFGBV. The second amendment option would seem to bring CDA 230 in line with the current Canadian approach to platform liability in defamation law, the intermediary liability regime in Quebec, and Article 14 of the European Union E-Commerce Directive and associated jurisprudence. These regimes shield only passive hosts without knowledge of specific illegal content, if they have not acted to remove or disable access to that content. Some critics have raised concerns that

---

[677] *Ibid.*

[678] *Ibid* at 469 (emphasis in original, footnotes omitted).

[679] *Ibid* at 469. Citron and Wittes note the likely impact of this on platforms that do attempt to moderate harmful content in good faith: "Treating abusive website operators and Good Samaritans alike devalues the efforts of the latter and may result in less of the very kind of blocking that CDA in general, and Section 230 in particular, sought to promote."

[680] *Ibid* at 470-71.

[681] *Ibid* at 471.

[682] *Ibid* (emphasis in original).

[683] *Copyright Act*, RSC 1985, c C-42, ss 27(2.3) and 27(2.4).

"reasonable steps" may be too uncertain or vague to rely on as a consistent standard.[684] However, Citron and Wittes suggest that a flexible standard would allow the law to "take into account differences among online entities", since what is reasonable for one may not be considered so for another.[685]

5.2. Germany

Germany's Netzwerkdurchsetzungsgesetz (NetzDG) ("Network Enforcement Act") has garnered much attention for its comparatively more aggressive approach to platform liability for harmful or unlawful expression by users.[686] The law appears to have inspired the Canadian federal government, with Prime Minister Justin Trudeau including the following request in his 2019 mandate letter to the Minister of Canadian Heritage, Steven Guilbeault:

> Create new regulations for social media platforms, starting with a requirement that all platforms remove illegal content, including hate speech, within 24 hours or face significant penalties. This should include other online harms such as radicalization, incitement to violence, exploitation of children, or creation or distribution of terrorist propaganda.[687]

Germany passed NetzDG in June 2017.[688] The law requires social media platforms to comply with the following three obligations: remove "manifestly unlawful" content within 24 hours after being notified, or within seven days for more complex cases; provide users with a complaint mechanism for reporting such content for removal; and publish semi-annual transparency reports "detailing its content moderation practices" upon receiving over 100 complaints per year.[689] Non-compliance with any of these requirements is penalized with fines of up to 50 million euros,[690] but only "for 'systematic'

---

[684] See e.g., "No, Internet Companies Do Not Get A 'Free Pass' Thanks To CDA 230"(24 October 2019), online: *Techdirt* <https://www.techdirt.com/articles/20191020/15092343224/no-internet-companies-do-not-get-free-pass-thanks-to-cda-230.shtm>l and Zoe Bedell and John Major, "What's Next for Section 230? A Roundup of Proposals" (29 July 2020), online, *Lawfare*: <https://www.lawfareblog.com/whats-next-section-230-roundup-proposals>.

[685] Danielle Keats Citron & Benjamin Wittes, "The Problem Isn't Just Backpage: Revising Section 230 Immunity" (2018) 2 Georgetown Law Technology Review 453 at 471.

[686] In early 2020, France passed a similar law with even stricter obligations, which was subsequently struck down by a French court as unconstitutional. European Digital Rights, "French Avia law declared unconstitutional: what does this teach us at EU level?" (24 June 2020), online: *European Digital Rights* <https://edri.org/our-work/french-avia-law-declared-unconstitutional-what-does-this-teach-us-at-eu-level/>.

[687] Rt Hon Justin Trudeau, PC, MP, Prime Minister of Canada, "Minister of Canadian Heritage Mandate Letter" (13 December 2019), online: *Prime Minister of Canada* <https://pm.gc.ca/en/mandate-letters/2019/12/13/minister-canadian-heritage-mandate-letter>.The Report of the Standing Committee on Justice and Human Rights, however, on *Taking Action to End Online Hate*, does not include enacting such a law as part of its formal recommendations.

[688] Katherine Feenan and Kathleen Donovan, "Online Culture Shift: Safer for Women in Politics" (August 2019) at 12, online (pdf): *Public Policy Forum* <https://ppforum.ca/wp-content/uploads/2019/08/OnlineCultureShift-PPF-Aug2019-EN.pdf>.

[689] Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (15 April 2019) at 2, online (pdf): *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf>.

[690] *Ibid* at 3. See also Thomas Escrit, "Germany fines Facebook for under-reporting complaints", *Reuters* (2 July 2019), online: <https://www.reuters.com/article/us-facebook-germany-fine/germany-fines-facebook-for-under-reporting-complaints-idUSKCN1TX1IC>.

breaches of the law", as opposed to "an honest mistake in judgment, or overlook[ing] an item by error".[691] The law captures social networking platforms "with more than 2 million users located in Germany"[692] (out of a total population of just under 84 million[693]).

What constitutes "unlawful content" is based on pre-existing laws in Germany that have criminalized specific types of expression:

> NetzDG does not actually create new categories of illegal content. Its purpose is to enforce 22 statutes in the online space that already existed in the German criminal code and to hold large social media platforms responsible for their enforcement. The 22 statutes include categories such as "incitement to hatred," "dissemination of depictions of violence," "forming terrorist organizations," and "the use of symbols of unconstitutional organizations." NetzDG also applies to other categories, such as "distribution of child pornography," "insult," "defamation," "defamation of religions, religious and ideological associations in a manner that is capable of disturbing the public peace," "violation of intimate privacy by making photographs," "threatening to the commission of a felony" and "forgery of data intended to provide proof."[694]

In April 2020, Germany amended NetzDG by adding a requirement that social networking platforms "must now not only delete potentially criminal content but also report it to the Federal Criminal Police Office (BKA), [and moreover] some data will have to be forwarded to the authorities, even before they have established suspicion", including user data such as IP addresses, port numbers, or user passwords.[695] This requirement to pass user data to law enforcement, even before any verification of illegal content or action, elicited serious concerns regarding user privacy in addition to freedom of

---

[691] William Echikson & Olivia Knodt, "Germany's NetzDG: A key test for combatting online hate" (November 2018) at 4, online (pdf): *Archive of European Integration* <http://aei.pitt.edu/95110/1/RR_No2018-09_Germany's_NetzDG.pdf>.

[692] Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (15 April 2019) at 2, online (pdf): *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf>.

[693] "Germany Population (Live)" online: *Worldometer* <https://www.worldometers.info/world-population/germany-population/>.

[694] Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (15 April 2019) at 2, online (pdf): *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf>.

[695] Philipp Grüll, "German online hate speech reform criticised for allowing 'backdoor' data collection" (19 June 2020), online: *Euractiv* <https://www.euractiv.com/section/data-protection/news/german-online-hate-speech-reform-criticised-for-allowing-backdoor-data-collection/1480243/>; and Amélie Heldt, "Germany is amending its online speech act NetzDG... but not only that" (6 April 2020), online: *Internet Policy Review* <https://policyreview.info/articles/news/germany-amending-its-online-speech-act-netzdg-not-only/1464>.

expression.[696] A proposal to modify the law to 'freeze' the transfer of user data to law enforcement until at least after the content was reviewed and confirmed to meet grounds for deletion failed.[697]

NetzDG has been strongly criticized by civil liberties advocates as creating a high risk of over-removal of legitimate and legal speech, resulting in chilling effects and undue constraints on users' freedom of expression, among other issues such as the risks of privatized enforcement of speech laws.[698] Critics also feared that NetzDG would provide a model and legitimacy to repressive and authoritarian regimes that overtly censor political dissent, pointing to draft legislation introduced in Turkey in July 2020 as an example.[699] Rebecca Zipursky argues that NetzDG violates international human rights law, specifically Article 19 of the International Covenant on Civil and Political Rights (ICCPR), based on lack of proportionality, overbreadth, vague language, and delegation of legal determinations to private companies.[700] In fact, the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (at the time), David Kaye, issued a letter to the German government specifically to express concerns about NetzDG. He highlighted that the high fines, strict deadlines for content removal, and lack of judicial oversight may result in, respectively, lack of proportionality, undue interference with freedom of expression, and incompatibility with international human rights law.[701] At the same time, Zipursky notes that "much of Germany's new law is consistent with prior hate speech legislation deemed acceptable by the [UN] Human Rights Committee", and moreover that "there is value in preserving parts of it", provided NetzDG is amended to include more nuance.[702]

---

[696] See e.g., Jana Gooth, "Germany adopted a worrisome revision of #NetzDG today: All flagged content must now be forwarded by the platforms to what is the equivalent of the FBI. What the law entails and why it is bad - a thread" (18 June 2020 at 18:11), online, *Twitter* <https://twitter.com/janagooth/status/1273740119514787845>; and Janosch Delcker, "Germany's balancing act: Fighting online hate while protecting free speech", *Politico* (1 October 2020), online: <https://www.politico.eu/article/germany-hate-speech-internet-netzdg-controversial-legislation/>.

[697] Philipp Grüll, "German online hate speech reform criticised for allowing 'backdoor' data collection" (19 June 2020), online: *Euractiv* <https://www.euractiv.com/section/data-protection/news/german-online-hate-speech-reform-criticised-for-allowing-backdoor-data-collection/1480243/>.

[698] Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (15 April 2019) at 3, online (pdf): *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf>.

[699] See e.g. Svea Windwehr and Jillian C York, "Turkey's New Internet Law Is the Worst Version of Germany's NetzDG Yet"(30 July 2020), online: *Electronic Frontier Foundation* <https://www.eff.org/deeplinks/2020/07/turkeys-new-internet-law-worst-version-germanys-netzdg-yet> and "Turkey: Ruling party moves to tighten grip on social media giants", *Al Jazeera* (21 July 2020), online: <https://www.aljazeera.com/ajimpact/turkey-ruling-party-moves-tighten-grip-social-media-giants-200721141442734.html>." In addition, "Russia, Singapore, and the Philippines have all cited NetzDG in pending legislation that will limit speech online." Rebecca Zipursky, "Nuts About NETZ: The Network Enforcement Act and Freedom of Expression" (2019) 42:4 Fordham International Law Journal 1325 at 1361.

[700] Rebecca Zipursky, "Nuts About NETZ: The Network Enforcement Act and Freedom of Expression" (2019) 42:4 Fordham International Law Journal 1325 at 1354-64.

[701] Letter from the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, commenting on the draft law "Netzdurchführungsgesetz", presented by the Government on 14 March 2017 (1 June 2017) at 4, online (pdf): *UN Office of the High Commissioner of Human Rights* <https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf>.

[702] Rebecca Zipursky, "Nuts About NETZ: The Network Enforcement Act and Freedom of Expression" (2019) 42:4 Fordham International Law Journal 1325 at 1329, 1368.

Instances of mistaken removal have in fact occurred, such as satirical tweets intended to parody racist remarks made by a German far-right politician, Beatrix von Storch.[703] The politician's own posts on social media were also removed under NetzDG and their contents subsequently widely reported in the media due to that removal, which "seemed to confirm fears of the Streisand effect, or what one journalist dubbed the Storch effect."[704] Politicians from Germany's liberal Freie Demokratische Partei ("Free Democratic Party", or FPD) "stated that they refrain from posting on social media because of NetzDG".[705] Critics of the law have also documented instances of NetzDG being used to target legitimate expression (albeit unsuccessfully, in that the platforms declined to remove the content).[706]

A review of research regarding the empirical impacts of NetzDG since it has been implemented results in uncertainty regarding whether the law has been beneficial or effective on the whole. While the mandatory transparency reports provide some data for evaluation,[707] Tworek and Leersen note that focusing on takedown metrics alone "does not reveal whether NetzDG has achieved its purpose of combating hate speech and other online excesses".[708] This is due in part to inconsistent implementation[709] and non-standardization of data between companies—for instance, Facebook and Twitter counted complaints, while Google counted "content items" complained about, regardless of how many complaints involved the same item.[710] Moreover, knowing simply how much of a certain kind of content has been removed has limited significance without knowing how it "compares to the overall

---

[703] Linda Kinstler, "Germany's Attempt to Fix Facebook Is Backfiring", *The Atlantic* (18 May 2018) online: <https://www.theatlantic.com/international/archive/2018/05/germany-facebook-afd/560435/>.

[704] Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (15 April 2019) at 4, online (pdf): *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf>.

[705] *Ibid* at 4.

[706] Dean Sterling Jones, "Fake News Ban Targets Political Speech, Sexual Content" (11 February 2018), online: *Shooting the Messenger* <https://shootingthemessenger.blog/2018/02/11/fake-news-ban-targets-political-speech-sexual-content/>.

[707] See e.g., "Germany" (2021), online: *Twitter Transparency* <https://transparency.twitter.com/en/reports/countries/de.html>; "Removals under the Network Enforcement Law" (2021), online: *Google Transparency Report* <https://transparencyreport.google.com/netzdg/overview?>; "Where can I see Facebook's NetzDG Transparency Reports?" (2021), online: *Facebook Help Center* <https://www.facebook.com/help/1057152381103922>; and "Where can I see Instagram's NetzDG Transparency Reports?" (2021), online: *Instagram Help Center* <https://www.facebook.com/help/instagram/704881976636188>

[708] Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (15 April 2019) at 4, online (pdf): *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf>.

[709] "While Google and Twitter chose to include the NetzDG complaint in their flagging tool (visible in the first step described above), Facebook placed the access to its NetzDG complaint procedure separately from the content under its imprint and legal information. To be more specific, Facebook's complaint form according to NetzDG is not incorporated in their feedback function next to the contentious post." Amélie Heldt, "Reading between the lines and the numbers: an analysis of the first NetzDG reports" (2019) 8:2 Internet Policy Review at 11.

[710] *Ibid* at 6.

volume" of that content which remains on the platform.[711] William Echikson and Olivia Knodt, in a report evaluating NetzDG six months into enactment, also found difficulty assessing the law's impact.[712]

Tworek and Leersen point out that the vast majority of content removed since NetzDG has come into force was not, in fact, removed for violating one of the 22 German speech laws, but for violating a platform's own Community Standards.[713] It is only if reported content is found to warrant no action under platforms' own Community Standards that the content is then reviewed against the NetzDG-designated German laws of general application. As a result, "it may be that NetzDG's most important effect was to ensure swifter and more consistent removal of content within Germany under the companies' community guidelines".[714] However, Amélie Heldt writes that so long as platforms continue funneling all complaints through their own community standards first, "the effects on online speech remain more or less similar than before the coming into force of the NetzDG making it almost impossible to truly evaluate the impact of such regulation".[715]

Even if the primary impact of NetzDG has been only to galvanize enforcement of pre-existing internal community standards, such results would appear to confirm an important observation for non-US countries where major US digital platforms operate, with respect to the advisability of relying on self-regulation as opposed to imposing legal obligations:

> Germany's approach has seemed to illustrate that, currently, the only way countries outside the U.S. receive sustained attention from social media companies is if they are a massive market (like China or the European Union) or journalists uncover significant human rights violations or they threaten companies with significant fines. Real financial liability commanded platform companies' attention. Germany had tried a voluntary compliance system with the companies since 2015 but found it ineffective. The German government chose the path of law only after it deemed the companies insufficiently compliant. […] Since the introduction of the law, transparency reports indicate that compliance rates are far higher.[716]

Extrapolating from the collective insights of those who have studied the law and its impacts, it seems that while fears of a mass spike in wrongful takedowns appear not to have necessarily been borne out, neither has NetzDG necessarily achieved what it was intended to do, raising questions of efficacy and

---

[711] *Ibid* at 4.

[712] William Echikson & Olivia Knodt, "Germany's NetzDG: A key test for combatting online hate" (November 2018) at 6-7, online (pdf): *Archive of European Integration* <http://aei.pitt.edu/95110/1/RR_No2018-09_Germany's_NetzDG.pdf>. They also note that

[713] Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (15 April 2019) at 5, online (pdf): *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf>.

[714] *Ibid* at 6.

[715] Amélie Heldt, "Reading between the lines and the numbers: an analysis of the first NetzDG reports" (2019) 8:2 Internet Policy Review at14.

[716] Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (15 April 2019) at 9, online (pdf): *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf (emphasis added).

proportionality.[717] While greater enforcement of platforms' own community standards is a positive outcome, it seems possible that this outcome could be achieved with a law that does not raise as many legal and human rights concerns and risks as NetzDG has. What is clear and seemingly unanimously agreed upon, however, is that far greater and better data is needed to make drawing more meaningful conclusions possible.[718]

5.3. United Kingdom

In April 2019, the United Kingdom government published a proposed new regulatory framework that would impose specific responsibilities on digital platforms to address online harms from user content and user behaviours. The proposal, known as the *Online Harms White Paper* ("White Paper"), notably would impose a new statutory duty of care on platform companies, to be overseen and enforced by an independent regulator, and would cover both illegal content and harmful content that is not necessarily illegal.[719] The UK government subsequently confirmed that this agency would be Ofcom, the country's communications regulator.[720] The proposed regime appears to be largely based on ideas developed and work done by Lorna Woods and William Perrin at the Carnegie UK Trust.[721] The UK government held a public consultation on the White Paper from April to July 2019, and published an Initial Consultation Response ("Initial Response") in February 2020, which summarizes submissions and feedback received on the contents of the White Paper and provides updates on the UK government's position in some areas in response to overarching concerns raised.[722] In December 2020, the government published its Full Government Response to the Online Harms consultation ("Full Response"). Legislation implementing the government's proposed regime, in the form of an Online Harms Bill, is not expected until at least late 2021.[723]

---

[717] "In Germany, one year after implementation, the new law seems to be neither particularly effective at solving what it set out to do, nor as restrictive as many feared." Mozilla, "Inside Germany's crackdown on hate speech" (April 2019), online: *Mozilla Internet Health Report* <https://internethealthreport.org/2019/inside-germanys-crackdown-on-hate-speech/>.

[718] *Ibid*. See also: "The law's actual impacts on hate speech may be difficult to prove empirically, since this complex phenomenon is influenced by countless other factors as well, including political, cultural, demographic, and economic shifts. […] [I]t will require much more research — and greater access to data — to determine whether NetzDG is achieving its aim, and whether any benefits outweigh the harms to free speech." Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (15 April 2019) at 7, online (pdf): *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* <https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf>.

[719] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper (White Paper)* (London, April 2019) at 7, 31.

[720] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper: Full Government Response to the consultation (Full Government Response)* (London, December 2020) at 5 (para 2).

[721]Lorna Woods and William Perrin, "Online harm reduction – a statutory duty of care and regulator" (April 2019) at 2, 8-9, online (pdf): *Carnegie UK Trust* <https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf>.

[722] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper: Initial Consultation Response* (London: February 2020) at 3.

[723] "Online Harms bill: Warning over 'unacceptable' delay", *BBC* (29 June 2020), online: <https://www.bbc.com/news/technology-53222665>.

Key elements of the legal regime proposed in the White Paper, as modified by the Full Response, include:

- **Statutory Duty of Care:** A statutory duty of care would apply to companies that "(a) host user-generated content which can be accessed by users in the UK; and/or (b) facilitate public or private online interaction between service users, one or more of whom is in the UK", with search engines explicitly included.

- **Two-Tiered Approach:** The regulatory framework is intended to be "proportionate, risk-based and tightly defined in its scope", rejecting a one-size-fits-all approach "to reflect the diversity of online services and harms".[724] This includes a two-tiered approach which imposes a different set of legal obligations depending on whether a company is considered Category 1 or 2.[725] The new framework is further focused by explicitly excluding types of harms or illegal activity addressed through other laws, such as intellectual property, consumer protection, fraud, data protection, and activity more appropriately addressed through criminal law enforcement rather than a regulatory approach.[726]

- **Addressing Legal but Harmful Content:** All companies governed by the regulatory framework will have a legal obligation to address illegal content or behaviour, regardless of category. Where content is legal but harmful, only Category 1 companies will have an obligation to take action, based on such companies' extensive sharing functions and interaction at scale significantly increasing risk of harm, and to "address the current mismatch between companies' stated safety policies and many users' experiences online".[727] This is particularly significant in addressing TFGBV, given the high proportion of TFGBV that constitutes harmful but legal content.

- **Definition of Harmful Content:** While the initial White Paper included a non-exhaustive list specifying 23 types of harmful content,[728] the Full Response departed from this approach. Instead, the proposed legislation will "set out a general definition", where harmful content and activity falls under the new legal regime if "it gives rise to a reasonably foreseeable risk of a significant adverse physical or psychological impact on individuals".[729] Secondary legislation will set out certain "priority categories" of harmful content considered to pose particularly a

---

[724] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper: Full Government Response to the consultation (Full Government Response)* (London, December 2020) at 8 (para 17).

[725] "*Ibid* at 10 (para 28).

[726] "*Ibid* at 24-25 (para 2.4).

[727] "*Ibid* at 29 (paras 2.15-2.16).

[728] The list of harms was divided into "harms with a clear definition" (e.g., nonconsensual distribution of intimate images, harassment and stalking, hate crime, incitement of violence, terrorist content and activity); "harms with a less clear definition" (e.g., trolling, extremist content and activity, coercive behaviour, intimidation, disinformation), and "underage exposure to legal content" (e.g., children accessing pornography or other inappropriate material): UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper (White Paper)* (London, April 2019) at 31.

[729] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper: Full Government Response to the consultation (Full Government Response)* (London, December 2020) at 24 (para 2.2).

high risk of harm to users, including criminal offences (e.g., child sexual abuse material, hate crimes); content and activity harmful for children (e.g., pornography, violence); and content and activity that is legal for adults but related to self-harm (e.g., eating disorders, suicide).

- **Super-complaints:** Ofcom will accept "super-complaints" if there is "substantial evidence of a systemic issue affecting large numbers of people, or specific groups of people".[730]
- **Codes of Practice:** The Full Response also abandoned the White Paper's proposal of developing individual codes of practice to govern companies' actions for every type of harmful content included in the regulatory framework. This change occurred in response to consultation feedback that numerous codes could result in confusion, redundancy, and risk-averse over-removal of content.[731] Instead, the regulator will develop fewer codes, which will detail how companies can meet the new statutory duty of care and will focus on "systems, processes and governance".[732] The framework permits companies to implement alternative practices if they are demonstrably as or more effective than the codes.[733] The UK government published alongside the Full Response several interim voluntary codes of practice, intended to address terrorism and child sexual exploitation and abuse.[734]
- **Overarching Duty of Care:** An overarching duty of care still applies to companies in the absence of a specific code of practice for a particular kind of harm, resulting in residual persistent obligations such as "assessing and responding to the risk associated with emerging harms or technology".[735]
- **Proactive Monitoring:** The White Paper rejected including a general monitoring obligation in platforms' duty of care, based on concerns about disproportionate burden and user privacy.[736] It bears noting that such an obligation would run counter to Article 15 of the EU E-Commerce Directive, which prohibits Member States (albeit noting that the United Kingdom is no longer one) from imposing general monitoring obligations on intermediary providers such as online platforms.[737] The Full Response proposed that companies should "consider voluntarily using

---

[730] *Ibid* at 64.

[731] *Ibid* at 41.

[732] *Ibid* at 41 (para 2.48).

[733] *Ibid*.

[734] *Ibid* at 42 (paras 2.51-2.52). "The child sexual exploitation and abuse interim code of practice builds on the Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse, that were developed by the UK, US, Canadian, Australian and New Zealand governments, following consultation with tech companies and Non Governmental Organisations." *Ibid* at 42 (para 2.53).

[735] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper (White Paper)* (London, April 2019) at 43; UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper: Full Government Response to the consultation (Full Government Response)* (London, December 2020) at 42 (para 2.49).

[736] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper (White Paper)* (London, April 2019) at 68.

[737] Content filtering mechanisms for the purpose of preventing intellectual property infringement have also been ruled to violate the *Charter of Fundamental Rights of the European Union*. See *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)*, C-70/10 (2011) at para 51, and *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (Sabam) v Netlog NV*, C-360/10 (2012) at para 49.

automated technology to identify and remove terrorist content and activity from their public services".[738] The new legislation would empower Ofcom to issue and enforce a requirement to use "automated technology to identify and remove illegal terrorist content from their public channels, where this is the only effective, proportionate and necessary action available."[739] These voluntary and potentially required measures also apply to CSAM.[740] Obligations to use automated technology will be subjected to "robust safeguards" such as technological accuracy and standards of necessity and proportionality.[741]

- **Transparency Reports:** Category 1 companies will be required to publish transparency reports about how they are addressing online harms on their platforms. Transparency report obligations, including required information, will differ by type of company depending on service type, capacity, and audience, to ensure proportionality and usefulness.[742] The Secretary of State for the Department of Digital, Culture, Media and Sport will have the power to include additional companies outside of Category 1 where necessary.[743] Transparency obligations will also be informed by the *Government Report on Transparency Reporting in relation to Online Harms*,[744] which accompanied the Full Response.
- **Enforcement Mechanisms:** Enforcement mechanisms for breach of the new duty of care include significant financial penalties, up to the higher of 18 million euros (approximately $30 million CAD) or 10% of annual global turnover.[745] Ofcom will also have a 'last resort' power to "disrupt a company's business activities in the UK, including blocking access in the most serious circumstances."[746] In addition, criminal liability for senior management will apply beginning two years after the new framework comes into force, if they "fail to respond fully, accurately, and in a timely manner, to information requests" from the regulator.[747]

The White Paper proposals were critiqued by a wide range of stakeholders, including women's rights and gender equality organizations and digital rights advocates. Some of the key criticisms are presented below, while noting that the UK government may have subsequently addressed specific points in its Initial Response or Full Response.

---

[738] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper: Full Government Response to the consultation (Full Government Response)* (London, December 2020) at 46 (para 2.65).

[739] *Ibid* at 46 (para 2.66).

[740] *Ibid* at 44 (paras 2.57-2.59).

[741] *Ibid* at 44-47 (paras 2.60-2.63 and 2.66 and 2.70).

[742] *Ibid* at 67(paras 4.15-4.16)

[743] *Ibid* at 67-68 (paras 4.13-4.17).

[744] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *The Government Report on Transparency Reporting in relation to Online Harms* (Report) (London, December 2020).

[745] UK, Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper: Full Government Response to the consultation (Full Government Response)* (London, December 2020) at 73 (para 4.43).

[746] *Ibid* at 73 (para 4.43).

[747] *Ibid* at 75 (para 4.49).

Women's rights organizations, gender equality advocates, and experts on TFGBV, including technology-facilitated intimate partner violence, shared concerns that the proposed framework risked conflating harm and abuse towards women with harm and abuse towards children, and did not sufficiently consider the needs of survivors of intimate partner and sexual violence, including the need for trauma-informed support.[748] Organizations also sought express inclusion of misogynistic abuse as a harm within scope, clearer definitions of harms, more specific measures of success, and requirements for more granularly disaggregated data in transparency reports.[749] Another source of criticism was the restriction of online harms to companies such as social media platforms, thus excluding, for example, TFGBV deriving from smart-home devices, IoT-facilitated intimate partner violence, and intimate partner surveillance through stalkerware apps.[750]

In addition to the above critiques, Olga Jurasz and Kim Barker note that the "perceived gender neutrality of the White Paper proposals is a contradiction in terms—rather than being gender neutral in its content and proposed applicability to harms, it is in fact excluding the experiences of women and girls online. By doing so, it renders gender-based harms suffered by women and girls invisible within the proposals."[751] They assert that the White Paper's list of online harms may operate to the systemic detriment of women and girls, by "focusing exclusively on harms suffered by children and adolescents (to the exclusion of adults), harms suffered as a result of image-based abuse (to the exclusion of text-based abuse) or excluding certain harmful and abusive behaviours from the regulatory framework altogether (e.g., excluding gender as a protected characteristic from the hate crime framework in England & Wales and in Scotland)."[752]

Critiques from general digital rights advocates, who did not necessarily specialize in TFGBV, primarily stressed concerns about regulatory overreach combined with economic incentives impacting freedom of expression (through, e.g., upload filters, vague or overbroad provisions, or unclear definitions). Such groups also sought stronger procedural safeguards for user expression, such as due process and appeal mechanisms for wrongful removals (as opposed to wrongful leave-ups), and stressed economic

---

[748] See e.g., "Online Harms White Paper: Consultation - Fawcett Society Submission" (June 2019), online: *Fawcett Society* <https://www.fawcettsociety.org.uk/Handlers/Download.ashx?IDMF=71f448d2-fbc1-4b63-ab0c-2b23852836e7>; "Online harms and Caroline's Law – what's the direction for the law reform?" online: *The Open University* <https://www.open.ac.uk/research/news/online-harms-and-carolines-law%E2%80%93whats-direction-law-reform>; Aqsa Suleman, "Online Harms White Paper Should Represent Survivors of Abuse" (23 April 2019), online: *Against Violence & Abuse* <https://avaproject.org.uk/online-harms-white-papers-should-represent-survivors-of-abuse/>; "Written Submission to the Online Harms White Paper Consultation" (June 2019), online (pdf): *University College London's Gender and Internet of Things (IoT) Research Project* <https://www.ucl.ac.uk/steapp/sites/steapp/files/online_harms_white_paper_consultation_response_giot_june_2019_final.pdf>; Glitch UK and End Violence Against Women Coalition, "The Ripple Effect: COVID-19 and the Epidemic of Online Abuse" (September 2020), online (pdf): *End Violence Against Women* <https://www.endviolenceagainstwomen.org.uk/wp-content/uploads/Glitch-and-EVAW-The-Ripple-Effect-Online-abuse-during-COVID-19-Sept-2020.pdf>; and Athena Stevens and Rozina Ahmed, "Women's Equality Party responds to Online Harms white paper", online, *Women's Equality Party* <https://www.womensequality.org.uk/onlineharms>.

[749] *Ibid*.

[750] *Ibid*.

[751] "Online harms and Caroline's Law – what's the direction for the law reform?", online: *The Open University* <https://www.open.ac.uk/research/news/online-harms-and-carolines-law%E2%80%93whats-direction-law-reform>.

[752] *Ibid*.

concerns such as potential harm to market competition and impacts on small and medium-sized businesses.[753] General digital rights advocates also cautioned against incentivizing or compelling increased or unlawful surveillance of platform users, through proactive monitoring, and called for greater transparency as well. There was concern that the regulatory framework was not sufficiently grounded in human rights as a starting point; however, see Section 3.3.7 ("External Content Moderation Bodies") of this report for commentary regarding how advocacy for a 'human rights-based framework' has also been used to elevate freedom of expression and privacy over the right to equality.

Lastly, Blayne Haggart and Natasha Tusikov suggest that the White Paper's central deficiency is the fact that it "almost completely ignores the systemic conditions that have made commercial online platforms so problematic", i.e., their micro-targeting, algorithm-driven, engagement-maximizing business models.[754] They are joined by Privacy International in this, which criticizes the White Paper for insufficiently considering the role of data exploitation, such as targeted ads used to manipulate voters, exploit and harass people, discriminate, and contribute to an overall harmful online ecosystem.[755]

In response to these and other submissions, the UK government has updated or provided further detail regarding some of its original proposals. For example, the White Paper noted there would be different standards of obligations depending on whether content is illegal, or legal but harmful, but did not suggest some companies would be altogether exempt from obligations as part of this differentiation. The Full Response states that obligations related to legal but harmful content will apply only to Category 1 services, as noted above, in attempting to compel effective harm reduction without unduly infringing on users' other human rights—namely, privacy and freedom of expression—and to address the concerns of smaller platform companies.

The limitation of obligations regarding legal but harmful content is relevant to how effective the framework may be in addressing TFGBV. Much of TFGBV includes instances of expression that may not individually reach the level of a legal offence, but the cumulative impact of which can wreak substantive

---

[753] See Open Rights Group, "UK: Online Harms Strategy must "design in" fundamental rights" (19 April 2019), online: *European Digital Rights* <https://edri.org/our-work/uk-online-harms-strategy-must-design-in-fundamental-rights/>; Privacy International, "Privacy International's Response to the Open Consultation on the Online Harms White Paper" (1 July 2019), online (pdf): *Privacy International* <https://privacyinternational.org/sites/default/files/2019-07/Online%20Harms%20Response%20-%20Privacy%20International_0.pdf>; ARTICLE 19, "Response to the Consultations on the White Paper on Online Harms" (June 2019), online (pdf): *ARTICLE 19* <https://www.article19.org/wp-content/uploads/2019/07/White-Paper-Online-Harms-A19-response-1-July-19-FINAL.pdf>; Open Rights Group, "ORG policy responses to Online Harms White Paper" (May 2019) online (pdf): *Open Rights Group* <https://www.openrightsgroup.org/app/uploads/2020/03/ORG_Policy_Lines _Online_Harms_WP.pdf>; and Electronic Frontier Foundation and New America's Open Technology Institute, "Electronic Frontier Foundation and New America's Open Technology Institute Joint Comments" online (pdf): *Electronic Frontier Foundation* <https://www.eff.org/files/2019/07/03/uk_online_harms_white_paper_consultation_submission_electronic_frontier_foundation_and_new_americas_open_technology_institute.pdf>.

[754] Blayne Haggart and Natasha Tusikov, "What the U.K.'s Online Harms white paper teaches us about internet regulation" (17 April 2019), online: *Conversation* <https://theconversation.com/what-the-u-k-s-online-harms-white-paper-teaches-us-about-internet-regulation-115337>.

[755] Privacy International, "Privacy International's Response to the Open Consultation on the Online Harms White Paper" (1 July 2019) at 4-7, online (pdf): *Privacy International* <https://privacyinternational.org/sites/default/files/2019-07/Online%20Harms%20Response%20-%20Privacy%20International_0.pdf>. See also Open Rights Group, "ORG policy responses to Online Harms White Paper" (May 2019) at 1, online (pdf): *Open Rights Group* <https://www.openrightsgroup.org/app/uploads/2020/03/ORG_Policy_Lines_Online_Harms_WP.pdf>.

and legally significant systemic harm on the lives of women and girls, including their ability to participate in public life and benefit equally from the protection or benefit of the law—including their right to freedom of expression. Haggart and Tusikov point out that conventional criticisms based in concerns about freedom of expression, for example, "tend to ignore all the voices that are already stifled by the current de facto online rules" and provide "effectively an argument to continue stifling the speech of those currently affected by these behaviours".[756] The extent to which obligations to address legal but harmful content will meaningfully address TFGBV across the Internet would thus depend at least in part on which particular platforms are included in Category 1 or 2, and based on what criteria. For example, some platforms may have small user bases relative to Facebook or Google Search, but should be considered Category 1 services if they are 'purpose-built' platforms dedicated to hosting and distributing content that may not be considered illegal but which constitutes TFGBV.

5.4. European Union

In the European Union (EU), the starting point for intermediary liability is *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce)* ("E-Commerce Directive"), specifically Articles 12-15.[757] In addition to the E-Commerce Directive, the European Commission has established several non-binding instruments meant to sit on top of and further clarify platform companies' legal obligations as intermediaries: the Code of Conduct on Countering Illegal Hate Speech Online, and both a Communication and subsequent Recommendation on Tackling Illegal Content Online. In addition, the EU introduced a proposed *Digital Services Act* in December 2020, which preserves the intermediary liability provisions of the E-Commerce Directive while introducing significant reforms concerning platform liability. Each of these instruments will be discussed in turn.

5.4.1. E-Commerce Directive (2000/31/EC)

The E-Commerce Directive is binding on all countries ("member states") in the EU, and each country must achieve the substantive requirements of the Directive through domestic law. Articles 12 through 14 provide safe harbour from liability for user content to an online intermediary if it serves as a "mere conduit" (Article 12); provides only caching services (intermediate or temporary storage of information for later transmission) (Article 13); or hosts user content, provided the intermediary does not have "actual knowledge" of illegality and "acts expeditiously" if it does receive actual knowledge (Article

[756] Blayne Haggart and Natasha Tusikov, "What the U.K.'s Online Harms white paper teaches us about internet regulation", *Conversation* (17 April 2019), online: *Conversation* <https://theconversation.com/what-the-u-k-s-online-harms-white-paper-teaches-us-about-internet-regulation-115337>. Haggart and Tusikov further note that the White Paper is an improvement on the status quo as it seeks to explicitly regulate user content through a transparent and accountable public body, rather than continue to allow implicit regulation by opaque and unaccountable private companies.

[757] EC, *Commission Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market*, [2000] OJ, L 178/1.

14).[758] Article 15 prohibits placing a general monitoring obligation on Internet intermediaries, which has implications for proposed content moderation measures such as filtering tools.[759] The Court of Justice of the European Union (CJEU) has interpreted Articles 12-15 in cases that member states' national courts have referred to it.[760] Notably, the EU has also, as of April 2019, implemented controversial exceptions for copyright.[761]

Most digital platforms fall under Article 14 of the E-Commerce Directive, as they host user content and often play some moderating role, meaning they are thus neither "mere conduits" the way that Internet service providers are (Article 12), nor do they merely cache information as a technical pitstop (Article 13). In recognition of hosts' comparatively more involved function, Article 14 imposes obligations to obtain safe harbour that Articles 12-13 do not. Therefore, under EU law, digital platforms enjoy a more conditional safe harbour than is available under CDA 230 in the United States.

EU intermediary liability law also distinguishes between 'active' and 'passive' hosts. If a platform "played an active role of such a kind as to give it knowledge of, or control over, the data stored", then no exemption from liability applies.[762] In a trademark case, eBay was considered to have played an active role because its involvement included "optimising the presentation of" and promoting items that engaged in trademark infringement.[763] Only passive hosts—whose involvement with user content "is of a mere technical, automatic and passive nature" and who have "neither knowledge of nor control over the information which is transmitted or stored"—are eligible for safe harbour, assuming they meet the required conditions under Article 14.[764]

5.4.2. Code of Conduct on Countering Illegal Hate Speech Online

The EU Code of Conduct on Countering Illegal Hate Speech Online ("Code") is a non-binding voluntary agreement between the European Commission and several major technology companies, specifically, as of April 2021: Facebook, Twitter, YouTube, Microsoft, Instagram, Snapchat, Dailymotion,

---

[758] EC, *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')*, [2000] OJ, L 178/1.

[759] But see EC, *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.)*, [2019] OJ, L 130/92.

[760] See e.g., *Google France SARL v Louis Vuitton Malletier SA*, C-236/08 to C-238/08 (2010) ; *L'Oréal SA and Others v. eBay International AG and Others*, C-324/09 (2011); *Sotiris Papasavvas v O Fileleftheros Dimosia Etaireia Ltd and Others*, C-291/13 (2014); *Tobias Mc Fadden v Sony Music Entertainment Germany GmbH*, C-484/14 (2016); and *Eva Glawischnig-Piesczek v Facebook Ireland Limited*, C-18/18 (2019).

[761] EC, *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.)*, [2019] OJ, L 130/92, art. 17.

[762] *Google France SARL v Louis Vuitton Malletier SA*, C-236/08 to C-238/08 (2010).

[763] *L'Oréal SA and Others v eBay International AG and Others*, C-324/09 (2011) at para 123.

[764] EC, *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.)*, [2019] OJ, L 130/92 at Recital 42; Google France and Google Joined Cases C-236/08 to C-238/08, at para 113.

Jeuxvideo.com, and TikTok.[765] The Code, established in 2016, is meant to supplement pre-existing hate speech criminal laws in the European Union, and states:

> While the effective application of provisions criminalising hate speech is dependent on a robust system of enforcement of criminal law sanctions against the individual perpetrators of hate speech, this work must be complemented with actions geared at ensuring that illegal hate speech online is expeditiously acted upon by online intermediaries and social media platforms, upon receipt of a valid notification, in an appropriate time-frame. To be considered valid in this respect, a notification should not be insufficiently precise or inadequately substantiated.[766]

The Code defines "illegal hate speech" as "all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin".[767] This definition as well as the Code's legal basis is rooted in the *European Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law* ("Framework Decision").[768]

It is clear at the outset that the Code fails to address gender-based violence on digital platforms. The Code's exclusive focus on hate speech based on racism and xenophobia, being an outgrowth of the Framework Decision, means that it does not capture, either in principle or by way of signatory companies' commitments and practices, misogynistic or sexist hate speech—or hate speech based on other forms of gender-based oppression, such as sexual orientation or gender identity—that is not primarily viewed as racist or xenophobic hate speech. However, the Code is discussed here as an illustrative example of how governments have attempted to further online platform accountability for harmful content hosted on their site, which may potentially be extended to cover or otherwise inform the design of regulatory models for addressing TFGBV.

Under the Code, content that is flagged to companies as "illegal hate speech" must be assessed within 24 hours, first under their own community standards and rules, and then according to the definition in the Framework Decision. In addition to the 24-hour review and removal/disable-access deadline, companies under the Code agree to commitments such as establishing "clear and effective processes to review notifications [flags]", community guidelines that specify prohibitions on the "promotion of

---

[765] "To prevent and counter the spread of illegal hate speech online, in May 2016, the Commission agreed with Facebook, Microsoft, Twitter and YouTube a "Code of conduct on countering illegal hate speech online". In the course of 2018, Instagram, Snapchat and Dailymotion joined the Code of Conduct. Jeuxvideo.com joined in January 2019, and TikTok announced their participation in the Code in September 2020." "The EU Code of conduct on countering illegal hate speech online", online: *European Commission* <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en>.

[766] European Commission, "Code of conduct on countering illegal hate speech online", at 1, available at: <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en>.

[767] *Ibid* at 1.

[768] EC, *Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law*, [2008] OJ, L 328/55.

incitement to violence and hateful conduct", user education and awareness efforts, staff training on "current societal developments", and cooperation with civil society organizations and experts.[769]

Results of the Code have been monitored and published annually by the European Commission, relying on the assistance of reporting organizations and trusted flaggers[770] in different member states to test each company's adherence to the Code, by flagging content considered to be illegal hate speech over a period of 6-7 weeks each year. By June 2020, according to the fifth such annual evaluation, the Code had resulted in member companies reviewing 90.4% of submitted notifications (of illegal hate speech content) within 24 hours and removing 71% of such content[771]—a marked increase from the results of the first evaluation in 2016, in which 40% of flagged content was reviewed within 24 hours and 28.8% of content was removed.[772] All five evaluations found that the monitored platform companies tended to treat trusted flaggers differently than general users, with trusted flaggers' notifications resulting in higher content removal rates and more feedback and transparency from the company, compared to general users flagging content for hate speech.[773]

Significantly more content "calling for murder or violence of specific groups" was removed than content "using defamatory words or pictures to name certain groups".[774] The European Commission interprets this data to suggest that "the reviewers assess the content scrupulously and with full regard to protected speech."[775] Without knowing more regarding the kind of defamatory or image-based content that was flagged and left up, however, this may represent a failing in the context of TFGBV specifically, when it comes to the wide range of subtle ways in which women, girls, and intersecting marginalized identities are subjected to both individual and coordinated mass attacks online.

---

[769] European Commission, "Code of conduct on countering illegal hate speech online", at 1, available at: <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en>.

[770] A trusted flagger is "an individual or entity which is considered by a hosting service provider to have particular expertise and responsibilities for the purposes of tackling illegal content online". European Commission, "Commission Recommendation of 1.3.2018 on measures to effectively tackle illegal content online" (1 March 2018) at 10 (section 4(g)), online: *European Commission* <https://digital-strategy.ec.europa.eu/en/library/commission-recommendation-measures-effectively-tackle-illegal-content-online>. An example would be a civil rights NGO with which a social media platform has set up a dedicated backchannel for the purpose of flagging hate speech, which is separate from and processed more quickly than the main queue of content flagged by general users on the platform.

[771] Didier Reynders, "5th evaluation of the Code of Conduct" (June 2020), online (pdf): European Commission <https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf>.

[772] European Commission, "Results of the 1st monitoring exercise", available at: <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en#howitperforms>.

[773] "Only Facebook is informing users systematically (93.7% of notifications received feedback). Instagram gave feedback to 62.4% of the notifications, Twitter to 43.8% and YouTube only to 8.8%. Jeuxvideo.com sent feedback to 22.5% of the notifications. While Facebook is the only company informing consistently both trusted flaggers and general users, Twitter, YouTube and Instagram provide feedback more frequently when notifications come from trusted flaggers." Didier Reynders, "Countering illegal hate speech online: 5th evaluation of the Code of Conduct" (June 2020), online (pdf): *European Commission* <https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf>.

[774] *Ibid*.

[775] *Ibid*.

The Code has been criticized by free expression advocates,[776] with particular consternation regarding the Code's potential to promote and entrench "privatised enforcement"[777] by platform companies and "state interference by proxy" with legal speech.[778] According to Article 19, for instance, the Code's definition of "illegal hate speech" is overbroad and the Code's underlying authority, the Framework Decision, may not be compatible with international human rights standards on freedom of expression. Additionally, while the Code is non-binding, it results from governments pressuring private companies to suppress potentially lawful content, which may thus be seen as a form of censorship. The Code has also been criticized for lacking due process mechanisms, including consultation with and involvement of civil society organizations in its development (despite repeated references to them in the Code itself); placing companies rather than courts in the position of being *de facto* arbiters over speech; allowing the state to remove legal content through such companies that they may not be able to lawfully remove directly; and lack of an appeal mechanism to challenge wrongful or mistaken removals.[779]

5.4.3. Communication and Recommendation: *Tackling Illegal Content Online*

In 2018, the European Commission issued the *Recommendation on measures to effectively tackle illegal content online* ("Recommendation").[780] This is a non-binding legal instrument that aims to convert into specific guidelines and actionable best practices the goals and principles of an earlier—also non-binding—document from 2017, the European Commission's *Communication on Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms* ("Communication").[781] Both the Communication and Recommendation apply to *all* forms of illegal content as defined in European law, as opposed to being more narrowly focused as the Code is, or as are other platform liability regimes or content moderation frameworks specific to copyright infringement or terrorism, for example.

The Communication sets out principles and best practices for detecting and flagging illegal content, expeditious removal of content, reporting "evidence of criminal or other offences" to law enforcement authorities, promoting transparency and due process, safeguards against over-removal and abuse or

---

[776] See e.g., Jens-Henrik Jeppesen & Emma J Llansó, "Letter to European Commissioner on Code of Conduct for "Illegal" Hate Speech Online" (3 June 2016), online: *Center for Democracy & Technology* <https://cdt.org/insights/letter-to-european-commissioner-on-code-of-conduct-for-illegal-hate-speech-online/>; Evelyn Aswad, "The Role of U.S. Technology Companies as Enforcers of Europe's New Internet Hate Speech Ban" (2016) 1 Columbia Human Rights Review 1; and "European Commission's Code of Conduct for Countering Illegal Hate Speech Online and the Framework Discussion: Legal Analysis" (June 2016) online (pdf): *ARTICLE 19* <https://www.article19.org/data/files/medialibrary/38430/EU-Code-of-conduct-analysis-FINAL.pdf>.

[777] See e.g., Eugénie Coche, "Privatised enforcement and the right to freedom of expression in a world confronted with terrorism propaganda online" (5 November 2018) 7:4 Internet Policy Review 1.

[778] Aleksandra Kuczerawy, "The Code of Conduct on Online Hate Speech: an example of state interference by proxy?" (20 July 2016), online: *KU Leuven Centre for IT & IP Law* <https://www.law.kuleuven.be/citip/blog/the-code-of-conduct-on-online-hate-speech-an-example-of-state-interference-by-proxy>.

[779] See generally ARTICLE 19, "European Commission's Code of Conduct for Countering Illegal Hate Speech Online and the Framework Discussion: Legal Analysis" (June 2016) at 14-18, online (pdf): *ARTICLE 19* <https://www.article19.org/data/files/medialibrary/38430/EU-Code-of-conduct-analysis-FINAL.pdf>.

[780] EC, *Commission Recommendation of 1 March 2018 on measures to effectively tackle illegal content online, C(2018) 1177* [2018].

[781] EC, *Communication from Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms COM(2017) 555* (28 September 2017).

gaming of content moderation processes, and measures to prevent repeat infringers or the reappearance of previously removed content.[782]

The Recommendation builds on and is intended to "give effect to" the principles, best practices, and safeguards established in the Communication.[783] It applies to all hosting service providers who do business with EU residents, regardless of whether the company itself is based in the EU.[784] While all sizes and kinds of platforms are included, the Recommendation acknowledges that "account should be taken of the situation of hosting service providers which, because of their size or the scale on which they operate, have only limited resources and expertise".[785]

The Recommendation defines "illegal content" to mean "any information which is not in compliance with Union law or the law of a Member State concerned",[786] including hate speech, terrorist content, child sexual abuse material, and copyright infringement.[787] For all types of illegal content, the Recommendation sets out provisions with respect to: submitting and processing notices (i.e., flagging content); informing users if their content was flagged and removed barring extenuating circumstances, and providing a process to contest removals; facilitating out-of-court dispute resolution mechanisms; transparency (e.g., publishing clear, detailed explanations and annual transparency reports); proactive measures (e.g., automated detection); safeguards to ensure that content removal mechanisms and decisions are proportionate, accurate, and well-founded (including human oversight); protection against abuse of process (e.g., bad-faith notices or counter-notices); and cooperation between and among hosting services providers (i.e., the platform companies), EU member states, and trusted flaggers (e.g., information-sharing, fast-tracking requests, open communications channels).[788]

The Recommendation also includes a separate, additional set of more demanding standards and practices for "terrorist content"[789] specifically, including removing such content within one hour of flagging.[790] While evaluating the provisions on terrorist content and their potential equality and other

---

[782] *Ibid*.

[783] European Commission, "Commission Recommendation on measures to effectively tackle illegal content online - Frequently Asked Questions" (1 March 2018), online: *European Commission* <https://ec.europa.eu/commission/presscorner/detail/en/MEMO_18_1170>.

[784] EC, *Commission Recommendation of 1 March 2018 on measures to effectively tackle illegal content online, C(2018) 1177* [2018].

[785] *Ibid* at 6.

[786] *Ibid* at 10.

[787] European Commission, "Commission Recommendation on measures to effectively tackle illegal content online - Frequently Asked Questions" (1 March 2018), online: *European Commission* <https://ec.europa.eu/commission/presscorner/detail/en/MEMO_18_1170>.

[788] EC, *Commission Recommendation of 1 March 2018 on measures to effectively tackle illegal content online, C(2018) 1177*.

[789] Defined as "any information the dissemination of which amounts to offences specified in Directive (EU) 2017/541 or terrorist offences specified in the law of a Member State concerned, including the dissemination of relevant information produced by or attributable to terrorist groups or entities included in the relevant lists established by the Union or by the United Nations." *Ibid* at 10 (section 4(h).

[790] See generally *ibid* at 14-15.

human rights implications are beyond the scope of this report,[791] the European Commission's rationale behind singling out such content should be interrogated in the context of efforts to address TFGBV.

For instance, the Regulation justifies the one-hour removal window in part by stating that "terrorist content is typically most harmful in the first hour of its appearance online".[792] However, misogynistic expression and sexualized online abuse against women—such as the non-consensual distribution of intimate images (NCDII)—also proliferate rapidly and are able to achieve further reach and cause more devastating damage the longer it is available online.[793] NCDII poses additional substantive harms given users' abilities to download and forward such content into private chats, where they end up out of reach of the platforms or anyone else to retrieve or delete.

In terms of proportionality, imposing stronger content removal mechanisms for TFGBV would also seem not to carry the same equality, privacy, and freedom of expression risks associated with government counter-terrorism efforts in Canada and in other jurisdictions. These risks include harms such as racial profiling, Islamophobic discrimination, and the targeting of Indigenous rights activists, whose protest activities civil liberties groups have pointed out may be cast and wrongly targeted as threats to Canadian sovereignty and national security under Canada's *Anti-terrorism Act*.[794]

The Communication, the Recommendation, and the Code represent the European Commission's continued attempts to increase the responsibility of digital platforms for user content hosted on their

---

[791] See, however, literature concerning Counter-Terrorism Internet Referral Units (CTIRUs, or IRUs) in the United Kingdom and at Europol, formed specifically as counter-terrorism content moderation teams, but which some have suggested expanding to include other kinds of illegal content found on digital platforms: "Counter-Terrorism Internet Referral Unit" (last edited 23 March 2021) online: Open Rights Group Wiki <https://wiki.openrightsgroup.org/wiki/Counter-Terrorism_Internet_Referral_Unit>; Kilian Vieth, "Europol Policing the Web: Internet Content & Counter-Radicalization – An Interpretive Policy Analysis Approach" Master Thesis, FU Berlin, online (pdf): *Netzpolitik*

<https://cdn.netzpolitik.org/wp-upload/2017/08/MA_KilianVieth_EuropolPolicingtheWeb_finale.pdf>; Europol, "Europol's EU Internet Referral Unit Partners with Belgium, France and the Netherlands to Tackle Online Terrorist Content" (2 March 2018) online: *Europol* <https://www.europol.europa.eu/newsroom/news/europol%E2%80%99s-eu-internet-referral-unit-partners-belgium-france-and-netherlands-to-tackle-online-terrorist-content>; European Commission, "Communication from the Commission to The European Parliament, the European Council and the Council: Fifteenth Progress Report towards an effective and genuine Security Union" (13 June 2018) at 5-6, online: *European Commission*

<https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/european-agenda-security/20180613_com-2018-470-communication_en.pdf?utm_source=POLITICO.EU&utm_campaign=221667b5dd-EMAIL_CAMPAIGN_2018_06_13_12_58&utm_medium=email&utm_term=0_10959edeb5-221667b5dd-189773877>

[792] EC, *Commission Recommendation of 1 March 2018 on measures to effectively tackle illegal content online, C(2018) 1177* at 8 (Preamble, para 35).

[793] "[NCDII] is a growing issue due to the ease with which nude or sexual images can be created, shared, uploaded, and downloaded; the speed in which images can disappear after being downloaded; the difficulties associated with removing images once they are online; and the variety of imagery contained on multiple sites." Henry & Flynn, "Image-Based Sexual Abuse: Online Distribution Channels and Illicit Communities of Support" at 1933. For further discussion of expression-based TFGBV proliferating across digital platforms generally, see Section 3.2.2 ("Platformed TFGBV Is Networked, Socially Gamified, and Distributed").

[794] See e.g., Canadian Civil Liberties Association, "Submission to the Standing Committee on Public Safety and National Security regarding Bill C-59, An Act respecting national security matters" (January 18) at 14, online (pdf), *Canadian Civil Liberties Association* <https://ccla.org/cclanewsite/wp-content/uploads/2018/01/2018-01-17-Written-submissions-to-SECU-re-C-59.pdf>; and International Civil Liberties Monitoring Group, "Brief on Bill C-59, the National Security Act, 2017" (May 2019) at 37-38, online (pdf): *International Civil Liberties Monitoring Group* <https://iclmg.ca/wp-content/uploads/2019/05/C-59-brief-May-2019-update.pdf>.

respective websites, while stopping short of actual legislation or regulation. The hope is that platform companies will voluntarily adhere to such non-binding initiatives for fear of being faced with more stringent and binding laws and regulations, which may have significant legal or monetary consequences such as under NetzDG in Germany, or under the EU's General Data Protection Regulation (GDPR) in privacy law. Such efforts are an example of what Michael Karanicolas has described as "jawboning": "moral suasion, whereby platforms are pressured through threats of regulation to shift their broader approach to moderating content in order to bring it into line with categories that governments might seek to target."[795]

None of the three European Commission instruments described above modify the Article 14 safe harbour provision for online intermediaries under the E-Commerce Directive. However, commentators have noted that if companies opt to follow the practices set out in the Recommendation, which they may due to the 'jawboning' effect, then in that very process of taking a more active role in moderating what user content is or is not permitted on their respective platforms, they are that much "more likely to obtain the knowledge/control that would result in them losing the exemption from liability in Article 14".[796] The ability to obtain safe harbour under Article 14 is not only based on, "upon obtaining actual knowledge or awareness of illegal content, [acting] expeditiously to remove or disable access to it",[797] but also upon only being a passive host in the first instance, involved on a "mere technical, automatic and passive" level, according to Recital 42 of the E-Commerce Directive.[798] If an online platform is found not to be a "passive host" in the first place, but "provided assistance" such as "optimising the presentation of" the illegal content,[799] then it would not be eligible for the exemption.

However, legal experts have observed that the combination of Recital 42 and Articles 12-14 have created uncertainty and confusion requiring careful reading of how they have been interpreted in related EU jurisprudence to parse.[800] Moreover, according to Tarlach McGonagle, the "binary distinction between passive and active intermediaries [...] has long been under strain" and "no longer adequately

---

[795] Michael Karanicolas, "Squaring the Circle Between Freedom of Expression and Platform Law" (2020) 20 Pittsburgh Journal of Technology Law & Policy 177 at186. Karanicolas points out that jawboning may benefit the governments engaged in the practice, as much as it does the platform companies: "There are a number of reasons why jawboning is effective. First, platforms may view an independently managed policy-shift as cheaper and less unpredictable than having to comply with new, binding rules. Jawboning also neatly sidesteps constitutional challenges that might stand in the way of enacting new laws, as well as any political resistance if the new rules are unpopular or controversial. Governments who face resistance, either from the public or internally, to new legislation may be able to successfully bluff that the laws are just over the horizon." (footnotes omitted)

[796] Toby Headdon, "EU Commission issues guidance to online platforms for tackling illegal content online" (10 October 2017) online: *Lexology* <https://www.lexology.com/library/detail.aspx?g=b354f52d-b255-400c-ab03-a384b16fcea0>.

[797] *Ibid*.

[798] EC, *Commission Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market*, [2000] OJ, L 178/1 at Recital 42.

[799] EC, *Communication from Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms COM (2017) 555* (28 September 2017), at 11.

[800] See e.g., Maria Lillà Montagnani, "A New Liability Regime for Illegal Content in the Digital Single Market Strategy" in Giancarlo Frosio, ed, Oxford Handbook of Online Intermediary Liability (Oxford: Oxford University Press, 2020) 295 at 299 and Joris van Hoboken et al, "Hosting Intermediary Services and Illegal Content Online" (2018) at 33, online (pdf): *European Commission* <https://www.ivir.nl/publicaties/download/hosting_intermediary_services.pdf>.

reflects the complexity" of quasi-passive, quasi-active functions that intermediaries now widely provide, such as recommending and ranking.[801]

In its Communication on Illegal Content Online, the European Commission emphasized that "taking such voluntary, proactive measures [as set out in the Communication] *does not automatically lead to the online platform losing the benefit of the liability exemption* provided for in Article 14".[802] The distinction may come down to whether the platform is actively involved in the illegal content *before* the fact (active role, loss of Article 14 exemption), or whether they are only involved in finding and acting upon the illegal content *after* the fact (passive role with proactive content moderation, Article 14 exemption intact). This may be a highly fact-based and contextual question, the answer to which depends on the specific platform and category of content in question.[803]

Both the Communication and Recommendation have been heavily criticized by civil society groups such as European Digital Rights and the US-based Center for Democracy and Technology and Electronic Frontier Foundation. Critics have raised concerns including:

- the lack of process to verify after the fact that removed content was, in fact, illegal;
- overreliance on 'trusted flaggers';
- potential conflict with the prominently controversial EU Copyright Directive regarding the active/passive status of online hosts for the purposes of Article 14 of the E-Commerce Directive;
- potential *de facto* imposing of a general monitoring obligation, in contravention of Article 15 of the E-Commerce Directive (albeit, as a non-binding instrument, any such efforts on the part of a platform would be considered voluntary and thus not imposed);
- lack of data that such a system would achieve its intended aims;
- continued concerns with privatization of law enforcement with respect to online content;
- insufficient accountability mechanisms and processes that circumvent court-order mechanisms for content removal;
- incentivizing over-removal by emphasizing speed and automation;

---

[801] Tarlach McGonagle, "Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation" in Giancarlo Frosio, ed, Oxford Handbook of Online Intermediary Liability (Oxford: Oxford University Press, 2020) 467 at 476.

[802] *Ibid* at 10 (emphasis in original). The Communication further states, "[T]he mere fact that an online platform takes certain measures relating to the provision of its services in a general manner does not necessarily mean that it plays an active role in respect of the individual content items it stores and that the online platform cannot benefit from the liability exemption for that reason. In the view of the Commission, such measures can[,] and indeed should, also include proactive measures to detect and remove illegal content online, particularly where those measures are taken as part of the application of the terms of services of the online platform. This will be in line with the balance between the different interests at stake which the E-Commerce Directive seeks to achieve." (footnote omitted)

[803] See e.g., "This Recommendation acknowledges that due account should be taken of the particularities of tackling different types of illegal content online and the specific responses that might be required, including through dedicated legislative measures." European Commission, "Commission Recommendation of 1.3.2018 on measures to effectively tackle illegal content online" (1 March 2018) at 2, online: *European Commission* <https://digital-strategy.ec.europa.eu/en/library/commission-recommendation-measures-effectively-tackle-illegal-content-online>.

- lack of regard for the technological limits of automated content moderation and filtering; and
- potential violation of the right to privacy and chilling effects on freedom of expression through requirements to report user content to law enforcement authorities.[804]

5.4.4. *Digital Services Act* (Proposed)

The European Commission released a proposed suite of legislative reforms in December 2020, collectively known as the Digital Service Act package. The package encompasses both the *Digital Services Act* (DSA) and the *Digital Markets Act*.[805] Of the two, the DSA prominently addresses issues most relevant to this report, concerning platform accountability and liability for harmful and illegal content by platforms' users. Specifically, the DSA would establish clear rules and legal obligations to enforce the responsibilities of online platforms for risks to users and users' rights. The DSA is intended to update or replace key sections of the E-Commerce Directive and "if enacted, will represent the most significant piece of legislation in the digital market" since the E-Commerce Directive itself.[806] However, the baseline intermediary liability regime in Articles 12-15 of the Directive would be preserved in the DSA.[807] Each member state would be required to designate a Digital Services Coordinator (DSC), which would be the primary national authority responsible for administering the DSA in each country.[808]

To support the new legislation, the European Commission held a consultation from June to September 2020, requesting comments regarding how to keep users safe online, the current intermediary liability regime, competition and gatekeeping issues among digital platforms, challenges regarding individuals offering services through online platforms (e.g., issues concerning the gig economy), governance of the Single Market, and other issues such as online advertising.[809]

---

[804] See e.g., European Digital Rights, "Q&A on the Recommendation on measures to 'effectively tackle illegal content online'" (1 March 2018) online: *European Digital Rights*

<https://edri.org/our-work/qa-the-recommendation-on-measures-to-effectively-tackle-illegal-content-online/>; Emma Llansó and Laura Blanco, "EC Recommendation on Tackling Illegal Content Online Doubles Down on Push for Privatized Law Enforcement (1 March 2018) online: *Center for Democracy & Technology* <https://cdt.org/insights/ec-recommendation-on-tackling-illegal-content-online-doubles-down-on-push-for-privatized-law-enforcement/>; Graham Smith, "Towards a filtered internet: the European Commission's automated prior restraint machine" (25 October 2017), online: *Cyberleagle* <https://www.cyberleagle.com/2017/10/towards-filtered-internet-european.html >.

[805] European Commission, "The Digital Services Act package" (last updated 3 March 2021), online: *European Commission* <https://ec.europa.eu/digital-single-market/en/digital-services-act-package>. The *Digital Markets Act* aims to put in place *ex ante* competition rules aimed at large platform companies which are considered gatekeepers in the market, to promote new entrants and competitors in the EU's Digital Single Market.

[806] Leo Moore, John O'Connor and David Cullen, "European Commission launches consultation on Digital Services Act package" (29 June 2020), online: *Lexology* <https://www.lexology.com/library/detail.aspx?g=e389524c-6d25-4ae7-96eb-fa19f4ad8a37>.

[807] EC, *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*, Articles 3-5 and Article 7.

[808] *Ibid* at 15.

[809] "The Digital Services Act: ensuring a safe and accountable online environment" online: *European Commission* <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en>; European Commission, "Digital Services Act – deepening the internal market and clarifying responsibilities for digital services" online: *European Commission* <https://ec.europa.eu/info/law/better-

Key elements of the proposed DSA, relevant to addressing platform liability for TFGBV, include:

- **Tiered Approach:** Internet intermediaries are divided into four categories, each with their own set of obligations, which cumulatively include obligations of the tier(s) prior. The broadest category is Intermediary Services, which provide network infrastructure, such as Internet service providers and domain name registrars. Intermediary Services also encompass the second category, Hosting Services, such as cloud servers. Hosting Services, in turn, include two subsets of hosts that are each divided out into the third and fourth categories: Online Platforms, which align with the level of intermediary that is the focus of this report (e.g., social media platforms, as well as gig economy apps and online marketplaces), and Very Large Online Platforms (VLOPs).[810]
- **Notice and Action:** Hosting intermediaries—which include both online platforms and very large online platforms—must implement mechanisms that allow anyone to notify the intermediary about illegal content on the company's platform. The company must set up their system so that users or anyone else can submit a "sufficiently precise and adequately substantiated" notice that a "diligent" company could rely on to determine if the reported content is in fact illegal. Notices that include all of the required information are deemed to "give rise to actual knowledge or awareness" on the part of the intermediary, for the purpose of triggering an obligation to act under the Article 5 safe harbour provision (the equivalent of Article 14 under the E-Commerce Directive).[811]
- **Online Platform Obligations:** The DSA imposes a set of obligations on all online platforms, with the exception of "micro or small enterprises" as defined in a separate Recommendation.[812] Obligations that are particularly relevant to addressing TFGBV on digital platforms include transparency reporting, complaint and redress mechanisms, external dispute settlement, trusted flaggers, "measures against abusive notices and counter-notices", terms of service that include "due account of fundamental rights", and reporting criminal offences.[813]
- **Content Removal Complaint Process:** Article 17 of the DSA requires online platforms, including VLOPs, to establish processes that allow users to challenge decisions taken on the basis that their content was considered illegal or against the platform's terms and conditions. If a user is not satisfied with a platform's internal complaint process, Article 18 of the DSA entitles them to an out-of-court dispute settlement process before a certified independent

---

regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-Internal-Market-and-clarifying-responsibilities-for-digital-services>. Submissions to the consultation are available at the previously mentioned link.

[810] "The Digital Services Act: ensuring a safe and accountable online environment" (2020), online: European Commission <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en>.

[811] EC, *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC* at 51 (Art 14).

[812] "This Section shall not apply to online platforms that qualify as micro or small enterprises within the meaning of the Annex to Recommendation 2003/361/EC." *Ibid* at Art 16.

[813] European Commission, "The Digital Services Act: ensuring a safe and accountable online environment", online: *European Commission*

<https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en>.

body.[814] Notably, particularly with respect to TFGBV, this provision as drafted includes no requirement to enable users to challenge a platform's decision to *leave up* content or permit an abusive account to remain active.

- **Content Moderation and Recommender Transparency:** All intermediaries governed by the proposed DSA must publish, annually at minimum, "clear, easily comprehensible and detailed reports on any content moderation they engaged in during the relevant period."[815] The DSA details specific types of information that must be included in such reports.[816] Further, VLOPs that use recommender systems must "set out in their terms and conditions … the main parameters used in their recommender systems".[817]

- **VLOP Obligations:** VLOPs are defined as online platforms that provide their services to 45 million or more "average monthly active" users in the European Union; the number is intended to reflect 10 percent of the EU's population and will be adjusted as needed in the year the DSA is adopted.[818] VLOPs will have additional obligations on top of sharing the obligations the DSA places on the other three categories, as they "pose particular risks in the dissemination of illegal content and societal harms". Obligations unique to VLOPs include risk management, independent risk audits, public accountability, transparency in their recommendation systems and "user choice for access to information", data sharing obligations (with authorities and researchers), codes of conduct, and "crisis response cooperation".[819] These obligations are detailed in Section 4, Articles 26-33 of the proposed DSA.

- **Systemic Risk Assessment:** VLOPs are required to conduct annual assessments to identify, analyze, and assess "any significant systemic risks stemming from the functioning and use made of their services".[820] Potential systemic risks include:

    - dissemination of illegal content;
    - negative impacts on fundamental rights enshrined in the EU Charter, including prohibition of discrimination; and
    - "intentional manipulation" of the platform's service, including through "inauthentic use or automated exploitation", with "actual or foreseeable negative effect on the protection of public health, minors, civic discourse, or actual or foreseeable effects

---

[814] EC, *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC* at 53-54 (Art 18).

[815] *Ibid* at 50 (Art 13).

[816] *Ibid* at 50 (Art 13).

[817] *Ibid* at 61-62 (Art 29).

[818] *Ibid* at 59 (Art 25).

[819] European Commission, "The Digital Services Act: ensuring a safe and accountable online environment", online: *European Commission* <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en>.

[820] EC, *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC* at 59 (Art 26).

related to electoral processes and public security".[821] VLOPs' risk assessments must particularly consider their content moderation and recommendation systems, as well as the systemic impact of their advertisement selection and display systems.[822]

- **Systemic Risk Mitigation:** Upon having identified any systemic risks, the VLOP must implement "reasonable, proportionate, and effective mitigation measures, tailored to the specific systemic risks identified".[823] One key suggested measure is "adapting content moderation or recommender systems, their decision-making processes, the features or functioning of their services, or their terms and conditions".[824] Other suggested measures include working with trusted flaggers; limiting advertisement display; or cooperating with other online platforms to address the systemic risks.[825]

5.5. Australia

Australia is the sole jurisdiction this report examines that has established an independent regulator dedicated to addressing online expression constituting TFGBV, the eSafety Commissioner of Australia. Section 5.5.1 discusses the regulator's governing legislation, the *Enhancing Online Safety Act 2015*, and forthcoming legislative reforms under the title of the *Online Safety Act*. Section 5.5.2 examines another relevant law in Australia, the *Sharing of Violent Abhorrent Material Act 2019*.

5.5.1. *Enhancing Online Safety Act 2015* and Reforms (*Online Safety Act*)

The *Enhancing Online Safety Act 2015* ("EOS Act") establishes the office of the eSafety Commissioner of Australia, an independent regulator that has the power to investigate and compel social media services and online platforms to act in addressing a range of illegal and harmful content online, including TFGBV. Initially established as the "Children's eSafety Commissioner", subsequent amendments to its enabling legislation expanded the scope of the Commissioner's legislative mandate beyond only children. Currently, the eSafety Commissioner may:

- investigate and act on complaints regarding "serious cyberbullying material targeted at an Australian child", including relying on a two-tiered content removal framework for social media platforms (where Tier 1 social media platforms may opt in and platforms declared Tier 2 social media platforms—which includes large social media platforms, currently covering, in practice, only Facebook, Instagram, and YouTube—are subjected to binding legal obligations to remove flagged content within 48 hours, on pain of civil penalties);[826]

---

[821] *Ibid*.

[822] *Ibid*.

[823] *Ibid* at 60 (Art 27).

[824] *Ibid* at 60 (Art 27).

[825] *Ibid* at 60 (Art 27).

[826] eSafety Commissioner, 'Social Media Tier Scheme", online: *eSafety Commissioner* <https://www.esafety.gov.au/about-us/who-we-are/social-media-tier-scheme>.

- disclose information to a variety of adults and authority figures, including parents, guardians, teachers, school principals, and law enforcement authorities;
- issue and enforce removal orders and civil penalties to users as well as social media websites and other platforms regarding image-based abuse, such as NCDII (including the power to issue a formal warning, remedial direction, infringement notice, enforceable undertaking, and seek an injunction or civil penalty order in court);
- administer the Online Content Scheme under the *Broadcasting Services Act 1992 (Cth)*, which involves investigating complaints and taking action to address illegal and harmful content, including child sexual abuse material;
- issue notices to a content service or hosting service provider under the *Sharing Abhorrent Violent Materials Act* ("AVM notice"), to make them aware if they are hosting or providing access to such materials; and
- give direction to Internet service providers, such as requiring them to temporarily block certain websites after the Christchurch mosque shooting in New Zealand.[827]

In 2018, the Australian government commissioned an independent review of the EOS Act and the Online Content Scheme. The review concluded that the current regime has been too fragmented and insufficient to address harmful content online, primarily involving "retrofitting child protection safeguards into online services and products after harm emerges, or the damage is done".[828] The independent reviewer recommended significant overhaul of the online safety regime to create a more deliberate "fit for purpose" legislative framework, finding incremental changes to be insufficient.[829] The reviewer suggested that new legislation should include building online safety requirements directly into platform designs and require proactive and preventative measures on the part of platform companies, with respect to mitigating online harms.[830]

In response, the Australian government initiated a process to reform its existing legal regime through a new *Online Safety Act*. It held a public consultation from December 2019 to February 2020 to gather views regarding the existing legislation and proposed reforms.[831] A draft exposure bill was released in December 2020, allowing for a further period of public comment that ended in February 2021.[832] Shortly

---

[827] eSafety Commissioner, "Our legislative functions", online: *eSafety Commissioner* <https://www.esafety.gov.au/about-us/who-we-are/our-legislative-functions>.

[828] Lynelle Briggs AO, "Report of the Statutory Review of the Enhancing Online Safety Act 2015 and the Review of Schedules 5 and 7 to the *Broadcasting Services Act 1992* (Online Content Scheme)" (October 2018) at 2, online: *Australian Government* <https://www.communications.gov.au/publications/report-statutory-review-enhancing-online-safety-act-2015-and-review-schedules-5-and-7-broadcasting>.

[829] *Ibid* at 2.

[830] *Ibid*.

[831]Department of Infrastructure, Transport, Regional Development and Communications, Australian Government, "Consultation on Online Safety Reforms", online: *Australian Government* <https://www.communications.gov.au/have-your-say/consultation-new-online-safety-act>.

[832] The Hon Paul Fletcher, MP, Minister for Communications, Urban Infrastructure, Cities and the Arts, "New legislation to protect Australians against harmful online abuse" (23 December 2020), online: *Ministers for Infrastructure, Transport, Regional Development and Communications* <https://minister.infrastructure.gov.au/fletcher/media-release/new-legislation-protect-australians-against-harmful-online-abuse>.

thereafter, the government introduced a bill in the Australian House of Representatives that would implement the *Online Safety Act 2021*.[833] The Explanatory Memorandum accompanying the bill emphasizes that the new legislation would:

- retain and replicate a number of the provisions in the *Enhancing Online Safety Act 2015*, such as the non-consensual sharing of intimate images scheme;
- articulate a core set of basic online safety expectations to improve and promote online safety for Australians;
- create a new complaints-based, removal notice scheme for cyber-abuse being perpetrated against an Australian adult;
- broaden the regulatory scheme to capture harms occurring on services other than social media;
- reduce the timeframe for service providers to respond to a removal notice from the eSafety Commissioner from 48 to 24 hours;
- bring providers of app distribution services and search engine services clearly into the remit of the new online content scheme; and,
- establish a specific and targeted power for the eSafety Commissioner to request or require Internet service providers to disable access to material depicting, promoting, inciting or instructing in abhorrent violent conduct, for time-limited periods in crisis situations.[834]

In March 2021, the Australian House of Representatives passed the bill, which as of April 2021 is under consideration by the Senate.

The Senate Standing Committees on Environment and Communications received over 140 submissions concerning the bill, including from the Online Hate Prevention Institute, Digital Rights Watch, and Electronic Frontiers Australia.[835] There appears to have been broad support for the overall reforms and their objectives, including giving a more expansive framing, scope, and mandate to the eSafety Commissioner relative to its prior narrower focus on primarily children. However, a number of groups have raised key issues, concerns, and recommendations, including:

- the bill has been rushed through Parliament with no notable amendments or alterations from the exposure draft, despite the government having received over 370 submissions during the public comment period concerning that draft;[836]

---

[833] *Austl, A Bill for an Act relating to online safety for Australians, and for other purposes*, 2019-2020-2021, 46th Parl (first reading in Senate 17 March 2021).

[834] Parliament of the Commonwealth of Australia, House of Representatives, "Online Safety Bill 2021: Explanatory Memorandum", online: *Parliament of Australia* <https://parlinfo.aph.gov.au/parlInfo/search/display/display.w3p;query=Id%3A%22legislation%2Fems%2Fr6680_ems_3499aa77-c5e0-451e-9b1f-01339b8ad871%22>.

[835] Parliament of Australia, "Submissions Received by Committee", online: *Parliament of Australia* <https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Environment_and_Communications/OnlineSafety/Submissions>.

[836] Digital Rights Watch, "Submission to the Senate Standing Committees on Environment and Communications" (2 March 2021) at 1, online (pdf): *Parliament of Australia*

- the Commissioner should be given the authority to act on any content deemed harmful enough to have been made unlawful, which would include attacks on an identifiable marginalized group (not merely on an individual);[837]
- the online abuse provisions include an intention element that is not required by other legislation, including criminal law, and should be removed;[838]
- the provisions which address online abuse towards adults, defining "offensive or malicious" content, which is captured by the scheme, could be broadly over-interpreted and used to suppress and silence protected speech, including political expression;[839]
- the content removal scheme, as drafted, captures all sexual content, which is likely to cause significant harm to those working in the sex industry, including sex workers, similar to the US FOSTA-SESTA regime;[840]
- the scheme does not contain adequate appeals mechanisms for individuals and companies that receive removal notices, and should provide for meaningful and timely appeals;[841] and
- penalties included in the Bill should be made proportional to the person's ability to pay.[842]

5.5.2. *Sharing of Violent Abhorrent Material Act 2019*

In April 2019, Australia enacted the *Sharing of Abhorrent Violent Material Act 2019* (SAVMA) in an amendment to the *Criminal Code Act 1995*. SAVMA criminalizes the following failures to act:

- failure to refer details of a recording or livestream of conduct that has occurred or is occurring in Australia, and which constitutes abhorrent violent material, to the Australian Federal Police "within a reasonable time" after becoming aware of the material on their platform; and

---

<https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Environment_and_Communications/OnlineSafety/Submissions>.

[837] The Online Hate Prevention Institute, "Submission to the Senate Standing Committees on Environment and Communications" (1 March 2021) at 1-2, online (pdf): *Parliament of Australia* <https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Environment_and_Communications/OnlineSafety/Submissions>.

[838] *Ibid at* 3.

[839] Digital Rights Watch, "Submission to the Senate Standing Committees on Environment and Communications" (2 March 2021) at 2, online (pdf): *Parliament of Australia* <https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Environment_and_Communications/OnlineSafety/Submissions>.

[840] *Ibid* at 3.

[841] *Ibid* at 3, 8.

[842] Electronic Frontiers Australia, "Submission to the Senate Standing Committees on Environment and Communications" (2 March 2021), online (pdf): *Parliament of Australia* <https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Environment_and_Communications/OnlineSafety/Submissions>.

- failure to expeditiously remove or cease hosting on all of the platform company's services abhorrent violent material that can be accessed within Australia.[843]

"Abhorrent violent material" is defined to be audio, visual, or audio-visual material that records or streams 'abhorrent violent conduct' by one or more people, which a reasonable person would consider as offensive. This includes a terrorist act; murdering or attempts to murder another person; or torturing, raping, or kidnapping another person.[844]

The material must also have been produced by one or more people who engaged in the conduct; conspired to engage, aided, abetted, counselled, procured, or was in any way knowingly involved in the conduct; or attempted to engage in the conduct.[845] Whether or not the material has been altered, or whether the conduct itself was engaged in within Australia (as opposed to the online content featuring the conduct being accessible in Australia), is irrelevant for determining liability under SAVMA, as is the geographical location of the content service provider.[846]

SAVMA additionally empowers the eSafety Commissioner of Australia to issue a written notice informing a content service provider or hosting service provider of specific abhorrent violent material on their platform, provided they have reasonable grounds.[847]

The obligation to report abhorrent violent material to the police applies to all Internet service providers, content service providers, and hosting service providers. Failure to meet this obligation results in a fine of 800 "penalty units", or $177,600 AUD at time of writing.[848]

The obligation to expeditiously remove or cease hosting abhorrent violent material in Australia applies to content service providers, where their service can be used to access content online. The penalty for breaching this obligation, for an individual, is a criminal conviction and 3 years' imprisonment or a fine of up to 10,000 penalty units ($2.22 million AUD), or both. For a corporation, the penalty is a criminal conviction and the greater of 50,000 penalty units ($11.1 million AUD) or 10 percent of the company's annual turnover during the 12 months prior to the offence.[849]

SAVMA has received intense criticism from platform regulation experts such as Daphne Keller and Evelyn Douek, on the following grounds:

---

[843] *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth), 2019/38 at sections 474.33 and 474.34.

[844] *Ibid* at section 474.32(1). Each of these actions is further defined in the Act.

[845] *Ibid* at section 474.31(1).

[846] *Ibid* at sections 474.31(2) and (3), 474.34(6).

[847] *Ibid* at sections 474.35-74.36.

[848] "Under most Commonwealth laws, financial penalties are expressed in terms of 'penalty units' instead of dollar figures. For example, a maximum penalty may be expressed as '10 penalty units' instead of $2100. On 1 July 2020, the value of a penalty unit will increase from $210 to $222." Department of Agriculture, Water and Environment, Australian Government, "Notice 89-2020 - Increase to Commonwealth penalty unit value" (19 June 2020), online: *Australian Government* <https://www.agriculture.gov.au/import/industry-advice/2020/89-2020>.

[849] *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth), 2019/38, ss 474.33(1) and 474.34(9), 474.34(10).

- it charges platform companies with a single objective (removing abhorrent violent material) without providing guidelines or safeguards and overwhelmingly incentivizes over-removal through significant penalties;
- it ignores the difficulties of journalistic decision-making with respect to publishing depictions of violence and would likely result in all news reports being removed across "all but the bravest or most risk-tolerant companies" out of an abundance of caution;[850]
- it may allow for overly broad interpretations of "abhorrent violent material" and "abhorrent violent conduct", leading to removal of legally protected content and unintended suppression of expression by journalists, human rights defenders, and activists;[851]
- it unrealistically requests, in effect, perfect enforcement of content standards by platforms;[852]
- it injects further opacity around speech regulation on digital platforms;[853] and
- it moves in a troubling direction of the spectre of holding Internet service providers liable for user content and behaviours.[854]

SAVMA also garnered a joint letter of concern from the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, raising similar concerns.[855]

5.6. New Zealand

New Zealand has one law in force relevant to platform liability for TFGBV, the *Harmful Digital Communications Act 2015*, which Section 5.6.1 examines. Section 5.6.2 discusses the non-binding Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online ("Christchurch Call").

---

[850] Daphne Keller, "Australia Shows the World How Not to Regulate Platforms, News, and Public Information" (11 April 2019), online: *Center for Internet and Society* <https://cyberlaw.stanford.edu/blog/2019/04/australia-shows-world-how-not-regulate-platforms-news-and-public-information>.

[851] Global Network Initiative, "The Global Network Initiative Expresses Concern About the Freedom of Expression and Privacy Implications of Australia's 'Sharing of Violent Abhorrent Material' Bill" (5 April 2019), online: *Global Network Initiative* <https://globalnetworkinitiative.org/gni-concerns-foe-privacy-australia-bill/>.

[852] Evelyn Douek, "Australia's 'Abhorrent Violent Material' Law: Shouting 'Nerd Harder' and Drowning Out Speech" (2020) 94 Australia Law Journal 41 at 12-16 (SSRN).

[853] *Ibid* at 17-19.

[854] *Ibid* at 19-21.

[855] Letter from the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, Fionnuala Ní Aoláin, commenting on the Criminal Code Amendment (Sharing of Abhorrent Violent Material) Law 2019 (4 April 2019), online (pdf): <https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24533>.

5.6.1. *Harmful Digital Communications Act 2015*

The *Harmful Digital Communications Act 2015* (HDCA) makes it a criminal offence in New Zealand if a person posts a digital communication that (a) was intended to cause harm; (b) would cause harm to "an ordinary reasonable person in the position" of the targeted person; and (c) did in fact cause harm.[856] "Harm" is defined as "serious emotional distress".[857] The HDCA provides a list of factors to assess in determining whether or not the digital communications would cause harm, including extremity of language, age and characteristics of the targeted individual, anonymity and repeat communications, reach of circulation, truth or falsity, and context.[858]

The HDCA is rooted in ten "communication principles" that must inform the harm assessment.[859] If someone receives a digital communication that violates any of the ten principles, they may submit a complaint to a designated independent agency under the Act,[860] currently the non-profit organization Netsafe.[861] The organization has no enforcement power and only provides a dispute resolution process. If Netsafe does not resolve the complaint to the targeted person's satisfaction, only then can the person apply for a court-ordered takedown of the violating post, and the person must show that they have already exhausted the Netsafe process.[862]

An intermediary liability regime under the HDCA provides a conditional safe harbour to "online content host[s]", defined as: "the person who has control over the part of the electronic retrieval system, such as a website or an online application, on which the communication is posted and accessible by the user".[863] This regime effectively amounts to a notice-and-action framework (specifically, notice-notice-takedown)[864] where the host is protected from both civil and criminal liability for a user's post so long as it complies with a set of obligations upon receiving a complaint about content violating the HDCA.[865]

Specifically, the online content host must do the following to retain their safe harbour:

---

[856] *Harmful Digital Communications Act 2015* (NZ), 2015/63 at section 22(1).

[857] *Ibid* at section 4.

[858] *Ibid* at section 22(2).

[859] A person violates one or more of the ten communications principles if their digital communications involves any of the following: 1) discloses sensitive personal information about someone; 2) is threatening, intimidating, or menacing; 3) is "grossly offensive to a reasonable person" in the targeted person's position; 4) is "indecent or obscene"; 4) harasses an individual; 5) makes a false allegation; 7) publishes something in breach of confidence; 8) incites or encourages someone to message another person to cause the recipient harm; 9) incites or encourages someone to die by suicide; or 10) denigrates an individual based on "colour, race, ethnic or national origins, religion, gender, sexual orientation, or disability". *Ibid* at section 6.

[860] *Ibid* at section 6.

[861] Netsafe, "Our Service", online: *Netsafe* <https://www.netsafe.org.nz/aboutnetsafe/our-service/>.

[862] Netsafe, "What is the HDCA?" (1 April 2021), online: *Netsafe* <https://www.netsafe.org.nz/what-is-the-hdca/

[863] *Harmful Digital Communications Act 2015* (NZ), 2015/63 at section 4.

[864] World Intermediary Liability Map, "*Harmful Digital Communications Act*, 2015", online: *World Intermediary Liability Map* <https://wilmap.law.stanford.edu/entries/harmful-digital-communications-act-2015>.

[865] *Harmful Digital Communications Act 2015* (NZ), 2015/63 at section 23.

- within 48 hours of receiving a valid notice of complaint, forward the notice to the author of the post (omitting identifying personal information of the complainant unless there is consent to reveal the information), and inform the author they may submit a counter-notice within 48 hours of receiving the notice;
- if the host cannot contact the author despite reasonable attempts, it must "take down or disable the content as soon as practicable", but "no later than 48 hours" after the complaint;
- if the author consents to removal in a valid counter-notice, the host must "take down or disable the specific content as soon as practicable";
- if the author refuses removal in a valid counter-notice, the host must leave the content up and notify the complainant "as soon as practicable", including identifying personal information if the author consents; and
- if the author submits no valid counter-notice, the host must "take down or disable the specific content as soon as practicable but no later than 48 hours after notifying the author".[866]

Put otherwise, under this regime, content that is the subject of a complaint can only remain on the platform if the author submits a valid counter-notice refusing removal within 48 hours of being notified, upon which the complainant is informed and the platform's role ends. Presumably at this point, the complainant is left to pursue a Netsafe complaint and then a court takedown order if needed, in addition to any other relevant legal action available. Netsafe is also empowered to submit notices of complaint to online content hosts on behalf of impacted individuals.[867]

The online content host further cannot benefit from the safe harbour unless it provides an "easily accessible mechanism" enabling users to submit complaints through the notice-and-action process described above;[868] or if the content was posted "on behalf, or at the direction, of" the platform.[869]

The HDCA empowers a court to issue any of the following orders to an online content host: a) take down or disable public access to posted or sent material; b) release to the court the identity of an author behind anonymous or pseudonymous communication; c) publish a correction in a manner specified by the court; or d) give a right of reply to the person targeted by the communication in question.[870] In considering whether or not to issue an order, the court must take into account factors such as: contents and caused or likely level of harm; purpose and intention; occasion, context, and subject matter; how far the content has reached; age and vulnerability of the targeted person; truth or falsity of the contents; public interest; defendant's conduct; and technical and operational practicalities of fulfilling the order if issued.[871] The HDCA also explicitly requires the court to "act consistently with the rights and freedoms contained in the New Zealand Bill of Rights Act 1990."[872]

---

[866] *Ibid* at section 24.

[867] *Ibid* at section 25(1).

[868] *Ibid* at section 25(2).

[869] *Ibid* at section 25(3).

[870] *Ibid* at section 19(2).

[871] *Ibid* at section 19(5).

[872] *Ibid* at section 19(6).

5.6.2. Christchurch Call

The Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online ("Christchurch Call") is an international, multi-stakeholder agreement established in response to a white supremacist terrorist attack on two mosques in Christchurch, New Zealand. The shooter killed 51 people and injured 50, while livestreaming the entire attack online. Both the recorded video of the livestream as well as a 74-page white supremacist manifesto that the killer posted in conjunction with the attack, meant to be disseminated and incite further mass violence, quickly went and remained viral despite all efforts to stem the tide by social media platforms. Canada is a founding supporter of the Christchurch Call, and platform company signatories include Facebook, Google, Twitter, YouTube, Microsoft, LINE, Amazon, and Daily Motion. Platform companies which have joined the Christchurch Call have agreed to the following key commitments:

- transparent and specific measures to prevent the uploading and dissemination of terrorist and violent extremist content, including immediate and permanent removal, relying on technical measures and notice-and-takedown procedures while remaining consistent with human rights;
- greater transparency in and enforcement of community standards and terms of service, as well as regular transparent public reporting;
- immediate and effective mitigation of the specific risk that terrorist and violent content is livestreamed, including real-time review; and
- review of algorithms and similar content curation processes that may amplify terrorist and violent extremist content or otherwise drive users towards it.[873]

Critics of the Christchurch Call have pointed to several issues, including in a joint statement by digital rights organizations.[874] Concerns include the lack of initial public consultation in drafting the text of the Call; failure to clearly define "online service provider" or "terrorist and violent extremist content"; failure to acknowledge the human rights challenges and limitations of upload filters and automated content moderation; and outsourcing of speech regulation to private companies.[875]

---

[873] Christchurch Call to Eliminate Terrorist & Violent Extremist Content Online, "About", online: *Christchurch Call to Eliminate Terrorist & Violent Extremist Content Online* <https://www.christchurchcall.com/call.html>.

[874] "Civil Society Positions on Christchurch Call Pledge", online: *Google Docs* <https://drive.google.com/file/d/1RfXLUnx662mmOJv3Z2c0NXEpsAXS8HGN/view>.

[875] Javier Pallero, "Access Now on the Christchurch Call: rights, wrongs, and what's next" (15 May 2019), online: *Access Now* <https://www.accessnow.org/access-now-on-the-christchurch-call-rights-wrongs-and-whats-next/>;

Priyal Pandey, "One Year Since the Christchurch Call to Action: A Review" (August 2020) 389 ORF Issue Brief 1, online (pdf): <https://www.orfonline.org/wp-content/uploads/2020/08/ORF_IssueBrief_389_Christchurch.pdf>; and Liz Woolery, "Three Lessons in Content Moderation from New Zealand and Other High-Profile Tragedies" (27 March 2019) online: *Center for Democracy & Technology* <https://cdt.org/insights/three-lessons-in-content-moderation-from-new-zealand-and-other-high-profile-tragedies/>.

6. Constitutional and Critical Analysis of Platform Liability for TFGBV

In assessing the legal, regulatory, policy, and technical approaches that governments and platform companies have applied to addressing technology-facilitated gender-based violence, abuse, and harassment (TFGBV) and similar forms of abusive content online, several critical issues emerge that render any proposed platform liability legislation a multifaceted and complex exercise. In addition to potential challenges implicating constitutionality,[876] extensive research, scholarship, and analysis abounds demonstrating the difficulties of platform liability for user content related to the particular nature of digital platforms themselves. Part 6 of the report discusses some of these issues.

Section 6.1 discusses key considerations in assessing the constitutionality of laws that may constitute limitations on freedom of expression, in the context of platformed TFGBV. The focus is on factors that militate towards constitutionality, such as the role of the right to equality and freedom from discrimination, and the nature and scope of the legislation. Section 6.2 examines potential unintended consequences of platform regulation that is not carefully done, such as wrongful removal of legitimate or beneficial expression; counterproductive misalignment between legal obligations and the specific platforms that the obligations apply to; and complications that arise with potentially embedding in law privatized regulation of public discourse. Section 6.3 briefly highlights additional challenges involved in achieving meaningful reform to address TFGBV.

6.1. Equality and Freedom of Expression in Canadian Constitutional Law

While platform liability for TFGBV may be a developing legal issue in Canada, the scourge of misogynistic, racist, and other forms of hate-based and discriminatory speech has long pre-dated the Internet. Laws to curb such speech in public spaces have routinely been met with constitutional challenges, based on claims that even narrow restrictions targeting the most extreme forms of hate speech amount to an unjustifiable violation of the right to freedom of expression. Similarly, proposals to regulate or place liability on digital platforms for abusive users invariably meet opposition and criticism rooted in concern for potential impacts on freedom of expression for Internet users in general.

Canadian constitutional and human rights law has repeatedly recognized the necessity and justifiability of placing limitations on freedom of expression, in order to uphold equality rights and protect

[876] Regarding platform regulation recommendations that appear in the June 2019 report by the House of Commons Standing Committee on Justice and Human Rights (JUST), "Taking Action to End Online Hate", for instance, Lex Gill notes that if enacted, some of the "proposed measures [such as content monitoring and removal obligations] could provoke complex questions regarding jurisdiction and enforcement" and may be "vulnerable to constitutional challenge as a violation of section 2(b) or section 8 privacy rights under the [*Canadian Charter of Rights and Freedoms*]". Lex Gill, "The Legal Aspects of Hate Speech in Canada" (June 2020) at 18, online (pdf): *Public Policy Forum* <https://ppforum.ca/wp-content/uploads/2020/07/1.DemX_LegalAspects-EN.pdf>.

historically marginalized and vulnerable groups.[877] It bears emphasizing that the right to equality and freedom from discrimination are as fundamental and as protected by the *Canadian Charter of Rights and Freedoms* as is freedom of expression. Canadian constitutional law is clear that these *Charter*-guaranteed rights are "part of a matrix, rather than a hierarchy in which some are more equal than others."[878] Moreover, as will be discussed below, restricting speech-based abuse directly promotes and protects the equality *and* freedom of expression rights of those who are targeted on the basis of a historically marginalized group identifier, or those who are otherwise silenced or engage in self-censorship due to belonging to one or more historically marginalized groups.[879] Multiple decisions from the Supreme Court of Canada have assessed and affirmed the constitutionality of laws prohibiting hate speech,[880] including hate speech published and distributed online.[881]

The platform liability context can be distinguished from circumstances that gave rise to much of the leading hate speech jurisprudence, due to the platform's intermediary role, which is typically at least one step removed from the actual speaker or publisher who is the defendant in most cases in this area. The all-important layer of users whose expression is facilitated by online platforms must not be ignored, and precedents cannot necessarily be applied directly from speaker (or publisher) to platform. Prioritizing users' expression is especially important where users also include members of historically marginalized groups accessing a level of public participation, mass mobilization, and political expression that would be—and demonstrably was—prohibitively difficult in the absence of the Internet, because traditional channels of public communication have been, and continue to be, bastions of systemic discrimination and inequality.[882]

At the same time, the underlying reasoning and legal principles supporting the constitutionality of hate speech prohibitions in Canadian law have only become more relevant than ever, in the context of TFGBV, combined with the cultural, political, and technosociological environments collectively forged

---

[877] *Canada (Human Rights Commission) v Taylor*, [1990] 3 SCR 892; *R v Andrews*, [1990] 3 SCR 870; *R v Keegstra*, [1990] 3 SCR 697; *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11.

[878] Jane Bailey, "Twenty Years Later *Taylor* Still Has It Right: How the *Canadian Human Rights Act*'s Hate Speech Provision Continues to Contribute to Equality" (2010) 50 Supreme Court Law Review 349 at para 47 (QL).

[879] See Section 6.1.3 ("TFGBV Is Low-Value Expression Far from the Core of Section 2(b)") and Section 6.1.4.2 ("Systemic Inequality and the Limitations of 'Counterspeech'").

[880] *Canada (Human Rights Commission) v Taylor*, [1990] 3 SCR 892; *R v Andrews* [1990] 3 SCR 870; *R v Keegstra,* [1990] 3 SCR 697; *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11.

[881] *Lemire v Canada (Human Rights Commission)*, 2014 FCA 18.

[882] See e.g., Paul Orlowski, "Indigenous Representation in the Media" in Kirsten Kozolanka & Paul Orlowski, eds, *Media Literacy for Citizenship: A Canadian Perspective* (Canadian Scholars, 2018) 164; Maite Taboada & Fatemeh Torabi Asr, "Tracking the gender gap in Canadian media" (3 February 2019), online: *Conversation* <https://theconversation.com/tracking-the-gender-gap-in-canadian-media-110082>; Manisha Krishnan, "In The Midst of a Race Reckoning, Global News Laid Off Some of Its Most Vocal Internal Critics" (26 August 2020), online: *Vice* <https://www.vice.com/en/article/jgx4ek/in-the-midst-of-a-race-reckoning-global-news-laid-off-some-of-its-most-vocal-internal-critics>; Vicky Mochama, "Canadian media continue to uphold whiteness at work: Mochama", *Toronto Star* (21 June 2017), online: <https://www.thestar.com/opinion/commentary/2017/06/21/canadian-media-continue-to-uphold-whiteness-at-work-mochama.html>; Jane Lytvynenko, "Vancouver Talk Radio Host Fired After Trainwreck Interview On Race" (13 October 2016), online: *Canadaland* <https://www.canadaland.com/cknw-host-fired/>; and Lindsay Richardson, "'We need more': Report finds Indigenous women given less opportunity in Canadian film and TV", *APTN* (23 May 2019), online: <https://www.aptnnews.ca/national-news/we-need-more-report-finds-indigenous-women-given-less-opportunity-in-canadian-film-and-tv/>.

by digital platforms and their users.[883] Law and context combine to justify legal reforms that would impose some degree of legal obligation or indirect liability on digital platforms for TFGBV by a user. The most effective legal reforms would account for the distinct role of digital platforms in the Internet ecosystem, differentiated from the direct perpetrator of TFGBV, while simultaneously recognizing the facilitative—and sometimes more active—role of digital platforms in the devastating and widespread perpetuation of TFGBV.

This part of the report (Section 6.1) will discuss each of the following points in turn, to demonstrate how established constitutional principles in Canadian equality and freedom of expression law apply with similar if not greater force in the context of digital platforms and TFGBV. Section 6.1.1 introduces the right to freedom of expression as established in Canadian constitutional law and the test for proportionality that a law must meet for its limitation of a right to be justifiable. Section 6.1.2 discusses the weight that the right to equality and freedom from discrimination holds in conducting the proportionality analysis. Section 6.1.3 demonstrates how TFGBV constitutes 'low-value' expression that does not advance or operates against values underlying freedom of expression, and warrants a lesser degree of constitutional protection.

Section 6.1.4 highlights the importance of context, which in the case of platformed TFGBV, means the relevant constitutional analysis must take into account critical contextual factors such as the particular nature of *platformed* TFGBV and platform governance, systemic inequality, and the role of the state in addressing private abuse of historically marginalized groups. Section 6.1.5 explains how legislation that is solely dedicated TFGBV is more likely to withstand constitutional scrutiny, whereas constitutionality would be jeopardized if TFGBV were 'bundled' with other issues separate from addressing private abuse perpetuating systemic oppression of historically marginalized groups. This is because the former would advance two elements of proportionality: ensuring an intelligible standard in the legislation's objective and scope, and drafting a law that is more remedial and salutary in nature, rather than punitive.

6.1.1. Right to Freedom of Expression and Constitutional Proportionality

Laws prohibiting types of TFGBV that do not already fall under other pre-existing grounds of illegal conduct—such as defamation, sextortion, criminal harassment, violation of privacy, or NCDII—may face constitutional challenge on the grounds that they infringe the right to freedom of expression. Freedom of expression is guaranteed under section 2(b) of the *Canadian Charter of Rights and Freedoms*, which protects "any activity that conveys or attempts to convey meaning […] Indeed, hate propaganda, defamatory libel, and publishing false news have all been found to fall within the ambit of s. 2(b)."[884] The scope of freedom of expression is deliberately kept broad in Canada. Where section 2(b) protection applies to certain forms of expression, limits on or prohibitions of that expression are constitutional if they are shown to be justified under section 1 of the *Charter*.

TFGBV encompasses a wide spectrum of harmful expression, from 'casual' sexist remarks, subtle slights, and day-to-day discrimination, up to intimidating or violative comments and behaviours,

[883] See generally Jane Bailey, "Twenty Years Later *Taylor* Still Has It Right: How the Canadian Human Rights Act's Hate Speech Provision Continues to Contribute to Equality" (2010) 50 Supreme Court Law Review 349.

[884] *Crouch v Snell*, 2015 NSSC 340 at para 102, citing *Irwin Toy Ltd v Quebec (Attorney General)*, [1989] 1 SCR 927; *R v Keegstra*, [1990] 3 SCR 697; *R v Lucas*, [1998] 1 SCR 439; *R v Zundel*, [1992] 2 SCR 731.

misogynistic threats, and expression or actions taken online that amount to violence.[885] Some of these categories of harmful expressive conduct or speech are already illegal on the basis of other areas of law, as mentioned above. Of the overwhelming amounts of TFGBV that remain, some instances are appropriately constitutionally protected, some are currently constitutionally protected but perhaps less appropriately so, and some do not receive any constitutional protection at all. This spectrum of constitutionality requires analyzing whether or not the expression attracts section 2(b) protection in the first instance, and whether or not the expression remains impervious to legal restriction after a section 1 analysis.

It is beyond the scope of this report to parse every kind of TFGBV and assess where on that particular constitutional spectrum such instances of TFGBV would lie—particularly as the section 1 analysis would require a specific law to evaluate alongside the expression it restricts. However, such an exercise, to the extent it encourages clarity and stops short of counterproductive over-categorization, may be the responsibility of legislators addressing platform liability for TFGBV through legal reform. The purpose of the following sections in this part of the report is to highlight key principles that speak to the constitutional validity of laws that restrict TFGBV, including through legal regimes applied to digital platforms to target TFGBV by their users.

6.1.1.1. Threats of Violence Are Not Protected Expression

At the outset, the Supreme Court of Canada (SCC) has determined that threats of violence are not protected under section 2(b). This is particularly significant for TFGBV, as a formidable proportion of TFGBV includes both threats of violence and violent expression against women, girls, and members of intersecting historically marginalized groups. Chief Justice McLachlin (as she then was) stated:

> This Court's jurisprudence supports the proposition that the exclusion of violence from the s. 2(b) guarantee of free expression extends to threats of violence […] It makes little sense to exclude acts of violence from the ambit of s. 2(b), but to confer protection on threats of violence. Neither are worthy of protection. Threats of violence, like violence, undermine the rule of law. As I wrote in dissent in *R. v. Keegstra* […], threats of violence take away free choice and undermine freedom of action. They undermine the very values and social conditions that are necessary for the continued existence of freedom of expression […].[886]

Short of violence, threats of violence, or violent expression, much expression constituting TFGBV may fall within the scope of section 2(b). The analysis would thus most likely turn to assessing the constitutionality of specific legislation targeting TFGBV, based on a proportionality test discussed next.

6.1.1.2. Section 1 Proportionality Analysis and TFGBV

The test developed in *R v Oakes* (*Oakes*) is used to determine justifiability under section 1 of the *Charter*.[887] A law that otherwise infringes a *Charter* right is constitutional if it targets a "pressing and

---

[885] Suzanne Dunn, "Is it Actually Violence? Framing Technology-Facilitated Abuse as Violence" in Jane Bailey, Asher Flynn & Nicola Henry, eds, *Emerald International Handbook of Technology-Facilitated Violence and Abuse* (UK: Emerald Publishing Ltd, 2021).

[886] *R v Khawaja*, 2012 SCC 69 at para 70 (citations omitted).

[887] *R v Oakes*, [1986] 1 SCR 103.

substantial objective", and is proportional.[888] Proportionality is determined through assessing if the infringement of the *Charter* right is rationally connected to the identified objective, if the means chosen to further the objective interfere as little as reasonably possible with the right, and if the benefit of the infringing measure outweighs its negative effects.[889] It is worth highlighting two important points regarding proportionality in the context of platform liability for TFGBV.

First, 'minimal impairment' does not mean that the law can intrude on a particular right only in the most minimal way possible or imaginable, but that such intrusion is "within a range of reasonably supportable alternatives".[890] The SCC in *Canada (Human Rights Commission) v Taylor* (*Taylor*) stated, in the context of section 13(1) of the *Canadian Human Rights Act* (*CHRA*) (prohibiting hate speech disseminated via telephone), "[T]he question is not so much whether the objective of s. 13(1) can be accomplished in a less restrictive way as it is whether the sacrifice required in order to combat successfully discriminatory effects is so severe as to make the impact of s. 13(1) upon the freedom of expression unacceptable."[891] The SCC in *Saskatchewan (Human Rights Commission) v Whatcott* (*Whatcott*), which upheld the constitutionality of a hate speech provision in Saskatchewan's provincial human rights statute, similarly reiterated that "while it may 'be possible to imagine a solution that impairs the right at stake less than the solution Parliament has adopted' there is often 'no certainty as to which will be the most effective'".[892] A regime that seeks to impose liability on digital platforms for TFGBV (whether by their users or by the platform in a less-than-intermediary role) would thus not have to be the least intrusive approach conceivable, but it must be reasonable in any trade-offs involved.

Related to the question of whether a law is minimally impairing *for its purpose* is the question of effectiveness in achieving that purpose. In *Lemire v Canada (Human Rights Commission)*, which upheld the constitutionality of section 13(1) of the *CHRA* in the context of hate speech disseminated online, the Federal Court of Appeal (FCA) stated: "Section 1 does not entitle or require courts to search out an optimal remedy for a complex social problem — a task for which they are not equipped. This is a matter for the legislature. The role of the courts is to ensure that the statutory remedy selected is within the range of what is reasonable."[893] As Jane Bailey writes, "[T]he government need not prove that an impugned measure will in fact achieve its stated goal, only that 'it is reasonable to suppose that the limit may further the goal'."[894]

Similarly relevant is assessing the likelihood of harm of the targeted expression, about which the SCC in *Whatcott* stated:

---

[888] The SCC in *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 86 (WL), recognized that proportionality may be the focus of most contention in the context of laws targeting expression-based abuse, given broad consensus concerning the pressing and substantive nature of addressing hate speech and its associated harms to vulnerable and historically marginalized groups.

[889] *Ontario (Attorney General) v G*, 2020 SCC 38 at para 71.

[890] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 101.

[891] [1990] 3 SCR 892, 1990 CarswellNat 1030 at para 68 (WL).

[892] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 101 (emphasis added).

[893] *Lemire v Canada (Human Rights Commission)*, 2014 FCA 18 at para 105.

[894] Jane Bailey, "Twenty Years Later *Taylor* Still Has It Right: How the Canadian Human Rights Act's Hate Speech Provision Continues to Contribute to Equality" (2010) 50 Supreme Court Law Review 349 at para 18 (QL), citing *Alberta v Hutterian Brethren of Wilson Colony*, 2009 SCC 37.

> This Court has addressed such criticism [relying on "likelihood or risk of harm" versus requiring evidence of a "clear causal link"] in a number of situations involving the applicability of s. 1 and has adopted a "reasonable apprehension of harm" approach. This approach recognizes that a precise causal link for certain societal harms ought not to be required. A court is entitled to use common sense and experience in recognizing that certain activities, hate speech among them, inflict societal harms. […] As was clear from *Taylor*, and reaffirmed through the evidence submitted by interveners in this appeal, the discriminatory effects of hate speech are part of the everyday knowledge and experience of Canadians. I am of the opinion that the Saskatchewan legislature is entitled to a reasonable apprehension of societal harm as a result of hate speech.[895]

The SCC's guidance regarding effectiveness, risk of harm, and causality applies with particular force to laws addressing TFGBV, given its myriad forms both subtle and overt, the long-term nature of its repercussions, and how TFGBV, by definition, evolves alongside and adapts to technological advances.

Second, Canadian hate speech laws in multiple legal contexts at both the provincial and federal levels have been upheld as constitutional, after thorough analyses of their proportionality relative to what was at stake. As an overview and preview of some of the issues discussed in the subsections below, Bailey summarizes some of the key factors collectively involved in those determinations:

> The Court concluded that these provisions [section 13(1) of the *CHRA* in *Taylor* and section 319(2) of the *Criminal Code* in *Keegstra*] restricted non-violent attempts to convey meaning and thus violated subsection 2(b) of the *Charter*, but were nevertheless justifiable in that:
>
> > (i) Hate propaganda as defined in the provisions lay far from the core values of the search for the truth, democratic participation, and self-fulfillment underlying freedom of expression, making their restriction more easily justifiable;
> >
> > (ii) The Code and *CHRA* provisions served pressing and substantial objectives underscored by other *Charter* values such as equality and multiculturalism, as well as Canada's international human rights obligations, respectively being aimed at: limiting the risk of harm that hate propaganda poses to target group members and to racial, ethnic, and religious harmony in Canada and promoting equality of opportunity unhindered by discriminatory practices based upon membership in, among others, a particular racial, religious, or ethnic group;
> >
> > (iii) Prohibiting the dissemination of hate propaganda as defined in the provisions was rationally connected with their objectives in that censure of the expression restricted fostered the protection of target group members and promoted equality, diversity, and multiculturalism in Canadian society; and
> >
> > (iv) The provisions were tailored to restrict public, rather than private dissemination of the expression in issue and, as such, their likely salutary effects

[895] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at paras 132, 143.

> on equality, multiculturalism, and protection of target group members outweighed their deleterious impact on expression.[896]

The FCA in *Lemire* later confirmed that "the application of section 13 [of the *CHRA*] to the Internet has not, in my opinion, changed the minimal impairment analysis under section 1. The medium may be different but the essential message of *Taylor* and *Whatcott* remains the same."[897] In fact, Evans JA suggested that the Internet makes that essential message all the more important to uphold:

> Communications through the Internet take a variety of highly effective forms, including material that incorporates text, graphics, and video. Indeed, a statutory prohibition of the communication of hate speech without including such a widely used and powerful means of communication as the Internet would be an exercise bordering on futility. To conclude that the application of section 13 to Internet communications is not a minimal impairment of section 2(b) rights would seriously jeopardize Parliament's ability to pursue the legitimate objective of curbing hate speech in order to prevent discrimination against members of targeted groups.
>
> Justice Rothstein recognized the power of this relatively new form of communication in *Whatcott* when he said [...]: "In terms of the effects of disseminating hateful messages, there is today the added impact of the Internet." It is true that the hate messages in *Whatcott* were disseminated by "low tech" means: the distribution of flyers and the insertion of personal advertisements in newspapers. However, the section of the Saskatchewan Code impugned in *Whatcott* defines very broadly the prohibited means of communicating hate messages, and may well include Internet or other computer mediated communications. Nothing in the Court's reasons suggests that this feature of the section threatened its constitutional validity.[898]

While most of the relevant jurisprudence on hate speech concerns hate and discrimination based on race, ethnicity, and religion, the underlying reasoning and principles apply to hate and discrimination based on sex, gender identity, or sexual orientation. In 2019, an Ontario court decided *R v Sears*,[899] "the first case in Canada in which women were specified as a targeted group in a conviction under the hate propaganda provisions of the *Criminal Code*".[900] The defendants, the editor and publisher of a community publication titled *Your Ward News* (YWN), had regularly written and published "misogynistic messages […] evocative of the types of expression and devices used to expose groups to hatred."[901] The publication simultaneously included abundant anti-Semitic hate speech, for which the defendants were also convicted. Before concluding that "both men were fully aware of the unrelenting promotion

---

[896] Jane Bailey, "Private Regulation and Public Policy: Toward Effective Restriction of Internet Hate Propaganda" (2003) 49 McGill Law Journal 59 at 69-71 (footnotes omitted).

[897] *Lemire v Canada (Human Rights Commission)*, 2014 FCA 18 at para 60.

[898] *Ibid* at paras 61-62.

[899] *R v Sears*, 2019 ONCJ 104.

[900] Canadian Race Relations Foundation, "Hate Crime in Canada" (last modified 2 March 2020), online: *Canadian Race Relations Foundation* <https://www.crrf-fcrr.ca/en/news-a-events/articles/item/26823-hate-crime-in-canada>.

[901] *R v Sears*, 2019 ONCJ 104 at para 10 (Appendix A).

of hate in YWN and intended that hatred to be delivered to others", the judge further described the contents of YWN which related to women:

> Any position communicated that essentially denies that an entire half of the world's population are human beings is so outrageously reprehensible that the word "hate" is starkly inadequate. One cannot refer to people as chattel, advocating violence against them, demeaning them as inferior to men, without promoting hatred towards them.[902]

In the context of laws addressing TFGBV, the proportionality analysis should take into account the fact that public messages, images, videos, and posts conveying meanings similar to the misogynistic contents of YWN—unstintingly quoted and laid out in the *Sears* decision—are routinely disseminated, celebrated, and amplified across a variety of digital platforms by critical masses of users (in conjunction with some platforms' algorithms), acting individually or in coordination or socialization with each other as described in Section 3.2.2 ("Platformed TFGBV Is Networked, Socially Gamified, and Distributed").

6.1.2. Right to Equality Must Inform Proportionality Analysis

The SCC has repeatedly emphasized that the right to equality and freedom from discrimination, as protected by section 15 of the *Charter*, must inform the proportionality analysis of laws restricting hate-based expression, and weigh in favour of the constitutional validity of such laws. The right to equality is also enshrined in international human rights law and international human rights treaties that Canada has ratified, including: the *International Covenant on Civil and Political Rights* (ICCPR), the *Convention on the Elimination of All Forms of Discrimination Against Women* (CEDAW), the *Convention on the Rights of Persons with Disabilities* (CRPWD), and the *International Convention on the Elimination of All Forms of Racial Discrimination* (ICERD).[903] The SCC stated in *Taylor*:

> In seeking to prevent the harms caused by hate propaganda, the objective behind s. 13(1) [of the CHRA] is obviously one of pressing and substantial importance sufficient to warrant some limitation upon the freedom of expression. It is worth stressing, however, the heightened importance attached to this objective by reason of international human rights instruments to which Canada is a party and ss. 15 and 27 of the *Charter*. […]
>
> That the values of equality and multiculturalism are enshrined in ss. 15 and 27 of the Charter further magnify the weightiness of Parliament's objective in enacting s. 13(1). These *Charter* provisions indicate that the guiding principles in undertaking the s. 1 inquiry include respect and concern for the dignity and equality of the individual and a recognition that one's concept of self may in large part be a function of membership in a particular cultural group. As the harm flowing from hate propaganda works in opposition to these linchpin *Charter* principles, the importance of taking steps to limit its pernicious effects becomes manifest.[904]

---

[902] *Ibid* at para 29.

[903] Government of Canada, "International Human Rights Treaties to which Canada is a Party" (last modified 30 July 2019), online: *Department of Justice* <https://www.justice.gc.ca/eng/abt-apd/icg-gci/ihrl-didp/tcp.html>.

[904] *Canada (Human Rights Commission) v Taylor*, [1990] 3 SCR 892, 1990 CarswellNat 1030 at paras 42 and 45 (WL).

The SCC in *Keegstra* further reinforced the weight of equality in upholding the constitutionality of criminal liability for hate speech,[905] citing a submission by LEAF:

> [T]he intervener L.E.A.F. made the following comment in support of the view that the public and wilful promotion of group hatred is properly understood as a practice of inequality:
>
> > "Government sponsored hatred on group grounds would violate section 15 of the *Charter*. Parliament promotes equality and moves against inequality when it prohibits the wilful public promotion of group hatred on these grounds. It follows that government action against group hate, because it promotes social equality as guaranteed by the Charter, deserves special constitutional consideration under section 15."
>
> I agree with this statement. In light of the *Charter* commitment to equality, and the reflection of this commitment in the framework of s. 1, the objective of the impugned legislation is enhanced insofar as it seeks to ensure the equality of all individuals in Canadian society. The message of the expressive activity covered by s. 319(2) [of the *Criminal Code*] is that members of identifiable groups are not to be given equal standing in society, and are not human beings equally deserving of concern, respect and consideration. The harms caused by this message run directly counter to the values central to a free and democratic society, and in restricting the promotion of hatred Parliament is therefore seeking to bolster the notion of mutual respect necessary in a nation which venerates the equality of all persons.[906]

Further, the right to equality and freedom from discrimination must reflect substantive equality, which is the notion that to achieve true equality, people in different positions may have to be treated differently. This is in opposition to formal equality, which is the idea of treating everyone the same way across the board, regardless of their respective starting points or social locations. When formal equality results in discriminatory impacts, such as a particular law applying to everyone 'equally' but disproportionately harming a specific historically marginalized group due to systemic factors, that is known as 'adverse impact discrimination'. According to the SCC in *Fraser v Canada (Attorney General)*:

> There is no doubt […] that adverse impact discrimination "violate[s] the norm of substantive equality" which underpins this Court's equality jurisprudence [...]. At the heart of substantive equality is the recognition that identical or facially neutral treatment may "frequently produce serious inequality" [...]. This is precisely what happens when "neutral" laws ignore the "true characteristics of [a] group which act as headwinds to the enjoyment of society's benefits" [...].[907]

The distinction between substantive and formal equality is relevant to analyzing platform liability for TFGBV because most content moderation policies and decisions are applied with formal equality to platforms' users. In fact, a substantive equality approach to online content moderation was recommended by the 2014-2020 UN Special Rapporteur on the promotion and protection of the right

---

[905] *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at paras 78-84 (WL).

[906] *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 80 (WL).

[907] *Fraser v Canada (Attorney General)*, 2020 SCC 28 at paras 47-48 (in-text citations omitted).

to freedom of opinion and expression, David Kaye, in his 2018 thematic report on content regulation to the UN Human Rights Council:

> Meaningful guarantees of non-discrimination require companies to transcend formalistic approaches that treat all protected characteristics as equally vulnerable to abuse, harassment and other forms of censorship. Indeed, such approaches would appear inconsistent with their own emphasis that context matters. Instead, when companies develop or modify policies or products, they should actively seek and take into account the concerns of communities historically at risk of censorship and discrimination.[908]

The failure to apply substantive equality to distinguish between users' social locations and power dynamics is what results in regressive situations such as Facebook having a set of rules that formally protected white men, but not Black children, from hate speech.[909] Substantive equality would also be relevant to determining whether someone is being subjected to a coordinated harassment campaign, for example, or is being subjected to an outpouring of legitimate criticism for a valid reason.

Where TFGBV is concerned, section 28 of the *Charter* is additionally relevant. Section 28 specifically states that the rights and freedoms in the *Charter* are guaranteed "equally to male and female persons", protecting gender equality in the guarantee of the other rights and freedoms within, as acknowledged in *Keegstra*.[910] Thus, for example, women's freedom of expression online must be as equally guaranteed and protected as men's. The constitutional guarantees in sections 2(b), 15, and 28 for women—individually and in concert—are not reflected in the misogynistic vitriol, gender-based targeted harassment, and threats of sexual violence and death that are the current price of online engagement for many women and members of intersecting marginalized groups.

Over-emphasizing the concept of freedom of expression in its philosophical ideal as a basis to oppose hate speech restrictions and other laws targeting similar expression for fear of "unacceptably chilling expression and unduly compromising [commitment to the freedom] fails to accord due regard for the fundamental democratic right to equality by rendering freedom of expression as the ground upon which equality may trespass only insofar as it does so minimally".[911] Yet, the seeming elevation of freedom of expression above the right to equality has been a common refrain in discourse concerning platform liability and content moderation in particular, perhaps in part due to many of the most prominent platforms being headquartered in the United States, and in part due to the historical ties between digital rights advocacy and cyberlibertarianism. As Mary Anne Franks, Kate Klonick, and others have noted, "the gravitational pull of the First Amendment" and its "grip […] on the public imagination" in the United States has exerted constant pressure to extend the doctrine's legal and social boundaries,

---

[908] David Kaye, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (2018 thematic report on content regulation), 2018, A/HRC/38/35, at para 48.

[909] Julia Angwin and Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children", *ProPublica* (28 June 2017), online: <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>.

[910] *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 82 (WL).

[911] Jane Bailey, "Twenty Years Later *Taylor* Still Has It Right: How the Canadian Human Rights Act's Hate Speech Provision Continues to Contribute to Equality" (2010) 50 Supreme Court Law Review 349 at para 46 (QL).

in addition to disproportionately influencing the major US-based platform companies' own internal policies and approaches to content moderation.[912]

The paramountcy that freedom of expression enjoys in the United States runs contrary to Canadian constitutional law explicitly stating, "A hierarchical approach to rights, which places some over others, must be avoided, both when interpreting the *Charter* and when developing the common law."[913] In fact, Canada's commitment to equality and freedom from discrimination—at least to the extent reflected in hate speech jurisprudence—is a major and deliberate point of divergence from First Amendment jurisprudence in the United States, as the SCC in *Keegstra* made clear:

> Where s. 1 operates to accentuate a uniquely Canadian vision of a free and democratic society, however, we must not hesitate to depart from the path taken in the United States. Far from requiring a less solicitous protection of *Charter* rights and freedoms, such independence of vision protects these rights and freedoms in a different way. […] [T]he international commitment to eradicate hate propaganda and, most importantly, the special role given equality and multiculturalism in the Canadian Constitution necessitate a departure from the view, reasonably prevalent in America at present, that the suppression of hate propaganda is incompatible with the guarantee of free expression [citations omitted].[914]

The above-mentioned Canadian vision "reflects a more comprehensive conception of both private and public forces affecting individual liberty than that adopted in the US. […] The thinner conception of liberty as freedom from government restriction underlying the US approach fails to take sufficient account of the de-liberating impact of hate propaganda on target group members".[915] The greater weight that Canadian constitutional law accords to the right to equality, combined with its recognition that private (non-government) violations of the right to equality and freedom from discrimination may warrant as robust a response as state violations,[916] has potential implications both for how Canadian law should approach other relevant legislation in the platform liability context, such as section 230 of the U.S. *Communications Decency Act*, as well as the constitutional analysis of the chosen approach.

---

[912] Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (2 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment>. See also Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harvard Law Review 1598.

[913] *Dagenais v Canadian Broadcasting Corp*, [1994] 3 SCR 835 at 877. See also *Singh v Minister of Employment and Immigration*, [1985] 1 SCR 177, 1985 CarswellNat 152 at para 115 (WL) ("[I]t is important to bear in mind that the rights and freedoms set out in the *Charter* are fundamental to the political structure of Canada and are guaranteed by the *Charter* as part of the supreme law of our nation. I think that in determining whether a particular limitation is a reasonable limit prescribed by law which can be 'demonstrably justified in a free and democratic society' it is important to remember that the courts are conducting this inquiry in light of a commitment to uphold the rights and freedoms set out in other sections of the *Charter*".)

[914] *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 60 (WL).

[915] Jane Bailey, "Private Regulation and Public Policy: Toward Effective Restriction of Internet Hate Propaganda" (2003) 49 McGill Law Journal 59 at 75, 102.

[916] *Ibid* at 75 ("The Canadian approach recognizes both public and private sources of oppression on individual liberty and expressly, through section 1 of the *Charter*, acknowledges the potential role for government in ameliorating the negative impacts of private sources of oppression on individuals".)

Although *Taylor* was decided over 30 years ago, Bailey's words about the decision ring even truer today, despite the passage of a decade since they were written:

> Affirmation of the *Taylor* majority's equality-based vision is perhaps more important today than when it was first expressed almost 20 [now 30] years ago. Social, technological and economic conditions exacerbate hate propaganda's risk to targeted vulnerable communities and their members. In the circumstances, judicial or legislative retreat into a detached, marketplace-of-ideas philosophy would be simply unacceptable. Equality demands and deserves more.[917]

As several leading SCC decisions and many scholars and researchers have demonstrated, the law must acknowledge how social location impacts one's ability to exercise freedom of expression, while giving full effect to the right to equality and freedom from discrimination as guaranteed by the *Charter*. Lawmakers and policy analysts must take power imbalances, gender-based and other forms of systemic discrimination, and historical and ongoing oppression of marginalized groups into account. Ignoring the material reality of these impacts on women, girls, and intersecting marginalized identities results in a bankrupt conceptualization and hollowing of both the right to equality and the right to freedom of expression, and how they are best upheld in the context of digital platforms and TFGBV.

6.1.3. TFGBV Is Low-Value Expression Far from the Core of Section 2(b)

Even on grounds of upholding freedom of expression alone, there is an argument to grant a lower level of constitutional protection to much of the expression that constitutes TFGBV, especially in its most extreme forms. TFGBV operates to the detriment of freedom of expression itself—specifically, the freedom of expression of those targeted and impacted by TFGBV. The SCC has established that while freedom of expression protects a broad range of expression, "not all expression will be treated equally in determining an appropriate balancing of competing values […] [D]ifferent types of expression will be relatively closer to or further from the core values behind the freedom, depending on the nature of the expression. This will, in turn, affect its value relative to other *Charter* rights".[918]

The type of hate speech at the centre of cases such as *Taylor*, *Whatcott*, *Keegstra*, *Andrews*, *Lemire*, and *Sears* constitutes low-value expression which "contributes little to", or runs counter to,[919] all three rationales at the heart of the right to freedom of expression: "the quest for truth, the promotion of individual self-development or the protection and fostering of a vibrant democracy where the participation of all individuals is accepted and encouraged".[920] The reasons provided by the SCC and the FCA in all of these decisions remain equally cogent in the context of TFGBV and hate propaganda based on sex or gender identity.

---

[917] Jane Bailey, "Twenty Years Later *Taylor* Still Has It Right: How the Canadian Human Rights Act's Hate Speech Provision Continues to Contribute to Equality" (2010) 50 Supreme Court Law Review 349 at para 84 (QL).

[918] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 112. See also *Canada (Human Rights Commission) v Taylor*, [1990] 3 SCR 892, 1990 CarswellNat 1030 at para 36 (WL).

[919] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 114.

[920] *Ibid* at para 113.

With respect to the question for truth, Dickson CJ (as he then was) wrote:

> [T]he argument from truth does not provide convincing support for the protection of hate propaganda. […] Indeed, expression can be used to the detriment of our search for truth; the state should not be the sole arbiter of truth, but neither should we overplay the view that rationality will overcome all falsehoods in the unregulated marketplace of ideas. There is very little chance that statements intended to promote hatred against an identifiable group are true, or that their vision of society will lead to a better world. To portray such statements as crucial to truth and the betterment of the political and social milieu is therefore misguided.[921]

Regarding individual self-fulfillment, autonomy, and human flourishing, *Keegstra* established that to the extent such fulfillment is achieved through the ability to exercise freedom of expression, hate speech "represents a most extreme opposition to the idea that members of identifiable groups should enjoy this aspect of the s. 2(b) benefit".[922] The Court in *Whatcott* expanded on how hate speech perniciously subdues both individual flourishing and participation in democracy:

> [H]ate propaganda opposes the targeted group's ability to find self-fulfillment by articulating their thoughts and ideas. It impacts on that group's ability to respond to the substantive ideas under debate, thereby placing a serious barrier to their full participation in our democracy. Indeed, a particularly insidious aspect of hate speech is that it acts to cut off any path of reply by the group under attack. It does this not only by attempting to marginalize the group so that their reply will be ignored: it also forces the group to argue for their basic humanity or social standing, as a precondition to participating in the deliberative aspects of our democracy.[923]

In addition to the examples cited throughout this report and elsewhere, empirical research has demonstrated that TFGBV silences women online and curtails their ability to exercise freedom of expression.[924] For example, the *Guardian* found that among their journalists who had been subjected to abuse—disproportionately female and racialized journalists—"53% had stopped reading comments, 33% said they stayed away from public debate, and 14% had seriously considered leaving journalism […] 20% had refused assignments as a result of abuse."[925] The impacts of abuse are not limited to misogynistic and sexist speech and conduct online; racist abuse has led to those targeted having "had to quit jobs, forgo education, leave their homes, avoid certain public places, curtail their own exercise

---

[921] *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 92 (WL).

[922] *Ibid* at 93 (WL).

[923] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 75.

[924] See Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (2 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment> ("That racist and sexist speech can produce chilling effects has been backed up by empirical studies showing that the targets of bigoted speech may experience not only psychological effects — lack of confidence, social anxiety, fear — but also physiological effects, such as increased heart rate and stress. This in turn can lead to targets censoring themselves as a means of avoiding these negative effects").

[925] Becky Gardiner, "'It's a terrible way to go to work:' what 70 million readers' comments on the Guardian revealed about hostility to women and minorities online" (2018) 18:4 Feminist Media Studies 592 at 601.

of speech rights, and otherwise modify their behavior and demeanor."[926] Freedom of expression is not advanced when journalists—moreover the very journalists contributing to media diversity and its associated freedom of expression and democratic implications for under-represented groups in the public sphere—are driven off of the Internet and away from careers and professions built on freedom of expression and unflinching use of that freedom.

In other words, hate speech "can achieve the self-fulfillment of the publisher [or commenter], but often *at the expense of that of the victim*. These are important considerations in [...] assessing the constitutionality" of laws prohibiting such speech.[927] As Mari Matsuda astutely articulated, "Tolerance of hate speech is not tolerance borne by the community at large. Rather, it is a psychic tax imposed on those least able to pay."[928] More often than not, there is a physical, financial, professional, and political tax as well. Speaking similarly to the notion of disproportionately borne burdens in the context of TFGBV, Citron writes, "Defeating online aggressions that deny victims their ability to engage with others as citizens outweighs the negligible contribution that [TFGBV] makes to cultural interaction and expression. […] We should be less troubled about limiting the expressive autonomy of [perpetrators of TFGBV] who use their voices to extinguish victims' expression."[929]

TFGBV that shares characteristics with the hate speech seen in the jurisprudence either does not further, or significantly stifles, participation in democracy. Particularly pertinent to the context of platform liability for TFGBV is the SCC's recognition that characterizing hate speech as "political" expression is no shield to otherwise justified limits, and may be all the more reason to restrict such expression:

> I recognize that hate propaganda is expression of a type which would generally be categorized as "political," thus putatively placing it at the very heart of the principle extolling freedom of expression as vital to the democratic process. Nonetheless, expression can work to undermine our commitment to democracy where employed to propagate ideas anathemic to democratic values. Hate propaganda works in just such a way, arguing as it does for a society in which the democratic process is subverted and individuals are denied respect and dignity simply because of racial or religious characteristics. This brand of expressive activity is thus wholly inimical to the democratic aspirations of the free expression guarantee. […]
>
> I am very reluctant to attach anything but the highest importance to expression relevant to political matters. But given the unparalleled vigour with which hate propaganda repudiates and undermines democratic values, and in particular its condemnation of the view that all citizens need be treated with equal respect and dignity so as to make participation in the political process meaningful, I am unable to

---

[926] Mari J Matsuda, "Public Response to Racist Speech: Considering the Victim's Story" (1989) 87 Michigan Law Review 2320 at 2337.

[927] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 114 [emphasis added].

[928] Mari J Matsuda, "Public Response to Racist Speech: Considering the Victim's Story" in Mari J Matsuda et al, *Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment* (New York: Routledge, 1993) 17 at 18.

[929] Danielle Keats Citron, "Restricting Speech to Protect It" in Susan J Brison & Katharine Gelber, eds, *Free Speech in the Digital Age* (New York: Oxford University Press, 2019) 122 at 130-31.

> see the protection of such expression as integral to the democratic ideal so central to the s. 2(b) rationale.[930]

Whatcott and *Lemire* make this same point in upholding the constitutionality of hate speech provisions in human rights legislation (the latter with respect to online expression specifically):

> [I]if one understands an effect of hate speech as curtailing the ability of the affected group to participate in the debate, relaxing the standard in the context of political debate is arguably more rather than less damaging to freedom of expression. As argued by some interveners, history demonstrates that some of the most damaging hate rhetoric can be characterized as "moral", "political" or "public policy" discourse.
>
> Finding that certain expression falls within political speech does not close off an enquiry into whether the expression constitutes hate speech. Hate speech may often arise as a part of a larger public discourse but, as discussed in *Keegstra* and *Taylor*, it is speech of a restrictive and exclusionary kind. Political expression contributes to our democracy by encouraging the exchange of opposing views. Hate speech is antithetical to this objective in that it shuts down dialogue by making it difficult or impossible for members of the vulnerable group to respond, thereby stifling discourse. Speech that has the effect of shutting down public debate cannot dodge prohibition on the basis that it promotes debate.[931]

The SCC's recognition that hate speech's political nature can undermine, rather than promote, freedom of expression is significant. When platform companies have faced criticism for allowing prominent purveyors of speech that constitutes TFGBV or falls under criminal or human rights hate speech provisions to remain on the platform, the companies have regularly cited the political relevance of such expression, and the notion of freedom of expression in the abstract, to defend their decisions. Large social media companies such as Facebook also implicitly use the political nature of hate speech as a disingenuous shield in systemic and institutional decisions to avoid implementing content moderation mechanisms that would advance the core values of freedom of expression on their platforms (truth, individual self-fulfillment and self-autonomy, and democratic participation) because it would disproportionately impact right wing-leaning content in the process (as a result of such content disproportionately violating content moderation policies in the messages conveyed).[932]

In fact, empirical research has demonstrated that legally restricting online abuse "can, when done carefully and well, enhance and diversify speech rather than chill it."[933] Specifically, the results of a survey that Jon Penney conducted found that laws criminalizing online harassment and intimidation

---

[930] *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at paras 95, 97 (WL).

[931] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at paras 116-17. See also *Lemire v Canada (Human Rights Commission)*, 2014 FCA 18 at para 66 ("Although the expression of political views is at the core of the protection provided by section 2(b), hate speech does not get a pass simply because its subject matter could be regarded as political or of public interest.").

[932] See Section 3.4 ("Critiques of Platform Approaches to Speech-Based TFGBV").

[933] Danielle Keats Citron & Mary Anne Franks, "The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform" (2020) 2020 University of Chicago Legal Forum 45 at 68.

> had a statistically significant salutary impact on women's willingness to share personal content online. This gender effect likely evidences that if women are aware of a law that penalizes or criminalizes online harassment and bullying, they feel less likely to be attacked or harassed and are thus more secure and willing to share, speak, and engage online. In other words, *these statutes may actually lead to more speech, expression, and sharing online among adult women online, not less*.[934]

Penney and Citron further summarize the results and its implications in a separate co-authored paper:

> In short, there was little evidence to support claims that the law would have substantial or significant chilling effects for online activities. […] [Women] were more likely to spend time online, more likely to share personally created or authored content online, and more likely to contribute to social network sites online. […]
>
> The more victims speak out, the more victims who have retreated from online engagement might return. A law that facilitates victim speech and engagement can help empower victims and, in the long term, prevent, mitigate, and reverse the negative impacts of online abuse and chilled speech. Public discourse and broader democratic deliberation would be enriched, with a wider array of voices, contributions, and perspectives, especially those from women and members of marginalized groups, who are most often targeted by online abuse.[935]

Based on this research, Penney highlights that as much as past Internet law and policy debates have primarily been concerned with the 'chilling effects' of government regulation, more attention must be paid to the 'chilling effects' of online abuse in the absence of such regulation.[936] He argues that "any new Canadian regulatory framework or scheme must take online abuse seriously" and that any intermediary safe harbour or legal immunity "must include express mandates, provisions, exceptions, and incentives to address" TFGBV and similar online abuse.[937]

Such findings, combined with the recognition of women's and other historically marginalized groups' pre-existing inability to equally exercise freedom of expression due to online abuse and harassment, make it increasingly clear that "the argument that rules governing such behaviour would stifle legitimate speech is effectively an argument to continue stifling the speech of those currently affected by these behaviours. […] The choice here isn't between free speech and censorship; it's between who

---

[934] Jonathon Penney, "Can Cyber Harassment Laws Encourage Online Speech?" (15 August 2017), online: *Berkman Klein Center for Internet & Society at Harvard University* <https://medium.com/berkman-klein-center/can-cyber-harassment-laws-encourage-online-speech-4e1ae884bfb> (emphasis added). See also Jonathon W Penney, "Internet Surveillance, Regulation, and Chilling Effects Online: A Comparative Case Study" (2017) 6:2 Internet Policy Review 1.

[935] Danielle Keats Citron & Jonathon W Penney, "When Law Frees Us to Speak" (2019) 87 Fordham Law Review 2317 at 2330, 2332-33.

[936] Jonathon W Penney, "Online Abuse, Chilling Effects, and Human Rights" in Elizabeth Dubois & Florian Martin-Bariteau, eds, *Citizenship in a Connected Canada: A Research and Policy Agenda* (Ottawa: University of Ottawa Press, 2020) 207 at 210-211.

[937] *Ibid* at 217.

will and won't be heard."[938] Indeed, the choice is not even between who will and will not be heard, because those whose expression would be chilled by laws restricting online abuse can still be heard, since they remain as free to engage in non-abusive expression as they have been able to all along. Thus, the choice is only whether or not women, and other marginalized and vulnerable groups and individuals who are silenced by TFGBV, will be heard or not.

Protecting freedom of expression cannot be an honest exercise without considering *whose* freedom of expression is being protected, and at whose expense. Currently, some groups of people enjoy, with impunity, an extreme freedom that they abuse in order to annul others' basic freedom. Permitting this state of affairs to continue does not advance the cause of free expression, but rather, protects the freedom to abuse marginalized groups at the expense of their fundamental freedom to express anything at all. A coherent and principled approach to platform regulation and platform liability should put human rights and substantive equality at its centre. This includes both Canadian human rights and equality law as well as international human rights law.

6.1.4. Critical Context: Platforms, Systemic Inequality, and Private Abuse

The SCC has stressed the importance of context in determining whether or not a limit on a particular freedom is proportionate under section 1 of the *Charter*. Applying this contextual principle to platform liability for TFGBV is particularly crucial given its sociotechnological dynamics, requiring nuanced understanding of both gender-based violence, abuse, and harassment as a pre-existing systemic problem combined with understanding how given technological affordances and platform environments interact with the problem. According to *Keegstra*, the analysis must not "lose sight of the factual circumstances" and "the proper judicial perspective under section 1 must be derived from an awareness of the synergetic relation between two elements: the values underlying the *Charter* and the circumstances of the particular case".[939] In stating this, *Keegstra* drew on the "contextual approach" that Wilson J set out in *Edmonton Journal v Alberta*:

> [A] particular right or freedom may have a different value depending on the context. It may be, for example, that freedom of expression has greater value in a political context than it does in the context of disclosure of the details of a matrimonial dispute. The contextual approach attempts to bring into sharp relief the aspect of the right or freedom which is truly at stake in the case as well as the relevant aspects of any values in competition with it. It seems to be more sensitive to the reality of the dilemma posed by the particular facts and therefore more conducive to finding a fair and just compromise between the two competing values under s. 1.[940]

Taylor likewise emphasized that "in balancing interests within s. 1 one cannot ignore the setting in which the s. 2(b) freedom is raised. It is not enough to simply balance or reconcile those interests

---

[938] Blayne Haggart & Natasha Tusikov, "What the U.K.'s Online Harms White Paper Teaches Us about Internet Regulation" (17 April 2019), online: *Conversation* <https://theconversation.com/what-the-u-k-s-online-harms-white-paper-teaches-us-about-internet-regulation-115337>.

[939] *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 50 (WL).

[940] *Edmonton Journal v Alberta (Attorney General)*, [1989] 2 SCR 1326 at 1355-56, cited in *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 50 (WL).

promoted by a government objective with abstract panegyrics to the value of open expression."[941] Instead, the SCC said a contextual approach requires appreciating whether or to what extent the restriction in fact "debilitates or compromises" principles underlying the *Charter* right in question.[942]

Where platform liability for TFGBV is concerned, there are at least three critical contextual factors that legal reform must take into account: the sociotechnological context of digital platforms and how they undermine the concept of the 'marketplace of ideas'; the pre-existing systemic inequality impacting marginalized individuals' ability to engage in 'counterspeech'; and the reality of private abuse as an ongoing threat to marginalized and vulnerable groups. Each of these factors will be discussed below.

6.1.4.1. Platform Dynamics and a Dysfunctional 'Marketplace of Ideas'

The first contextual factor is the technological, sociocultural, and political context of digital platforms themselves. Such context includes: understanding how individuals use and exploit platforms' technological affordances to engage in TFGBV; how historically marginalized groups also rely on platforms for beneficial ends and in exercising their own *Charter* rights; how platforms' business models incentivize content moderation policies and decisions; and how the contemporary role and power of digital platforms has impacted public discourse and the information environment for the average individual, for those whose online expression and behaviours constitute TFGBV, and for women and others targeted by TFGBV. Many of these issues were discussed throughout Part 3 ("Role of Digital Platforms in TFGBV").

As an example of how this factual context may impact analyzing a platform liability law for TFGBV, the platform-facilitated normalization of gender-based hate speech may, over time and if left unchecked, raise the bar for what expression meets the threshold of sufficient extremity to fall under either a criminal or human rights hate speech provision. In 2007, a report commissioned by the Department of Justice included the following findings, summarized by Jane Bailey:

> Many groups [targeted by hate crimes] perceive that what their members experience as being racially charged events will not meet legal thresholds that focus on clear hateful motivation against an identifiable group. As the representative of one South Asian non-governmental organization noted, this approach leaves out "more culturally pervasive" forms of racism. *These data suggest a perception that the more broadly vilified a racial or religious group is, the less protection its members can expect from the law*. Such views are particularly disturbing in light of growing anti-Muslim and anti-Semitic sentiment in both Europe and Canada.[943]

While the report concerns targeting groups and individuals based on race and religion, similar dynamics would apply to TFGBV (which, additionally, intersects with race and religion). As explained in Section 3.2.3 ("Platformed TFGBV Normalizes and Escalates Violence against Women"), TFGBV proliferating across digital platforms contributes to making gender-based violence, abuse, and harassment—

---

[941] *Canada (Human Rights Commission) v Taylor*, [1990] 3 SCR 892, 1990 CarswellNat 1030 at para 48 (WL).

[942] *Ibid* at para 48 (WL) ("Rather, a contextual approach to s. 1 demands an appreciation of the extent to which a restriction of the activity at issue on the facts of the particular case debilitates or compromises the principles underlying the broad guarantee of freedom of expression").

[943] Jane Bailey, "Twenty Years Later *Taylor* Still Has It Right: How the Canadian Human Rights Act's Hate Speech Provision Continues to Contribute to Equality" (2010) 50 Supreme Court Law Review 349 at para 60 (QL) (emphasis added).

including sexist and misogynistic rhetoric—more "culturally pervasive" and thus normalized. This can desensitize or subtly influence even those who may not initially share or condone such beliefs, attitudes, and behaviours, making what would have in the past been considered an extreme statement or sentiment seem less so over time.[944]

In addition, modern-day parasocial relationships[945] between right-wing 'influencers' on social media and their audiences bear a striking resemblance to the social dynamics that *Taylor* describes regarding the impugned telephone campaign in the case. In *Taylor*, the SCC stated:

> Dr. Ravault [an expert witness] was also able to demonstrate how the authors of hate messages are able through subtle manipulation and juxtaposition of material to give a veneer of credibility to the content of the messages. The combination of the telephonic medium and the material is, we believe, particularly insidious, because, while a public means of communication is used, it is one which gives the listener the impression of direct, personal, almost private, contact by the speaker, provides no realistic means of questioning the information or views presented and is subject to no counter-argument within that particular communications context.[946]

Rebecca Lewis, who has extensively researched far-right extremism on YouTube, observes likewise:

> For years, YouTube has described this in democratizing terms. Indeed, people in their bedrooms can broadcast directly to their fans, creating a sense of intimacy and authenticity not present in older forms of media. In practice, however, that means a range of anti-feminist, Islamophobic, and even white supremacist content creators share far-right propaganda in the form of incredibly intimate personal stories, spoken to their audiences as if they are speaking to close friends. […]
>
> Parasocial relationships can seem particularly strong when a creator streams for hours on end, and when a viewer, such as Caleb, is lonely or confused. And I argue in my research that it is these relationships — the trust-building, personal storytelling, and seeming authenticity — that convincingly sells audiences on far-right ideas.[947]

The parasocial and quasi-personalized element of online media consumption, combined with the ubiquitous information and sociocultural environment surrounding users on digital platforms, directly undermines concepts such as the 'marketplace of ideas', which is often referenced in opposition to laws restricting hate-based or discriminatory expression. The 'marketplace of ideas' refers to a "classic image of competing ideas and robust debate" that "assumes that a process of robust debate, if uninhibited by governmental interference, will lead to the discovery of truth, or at least the best

---

[944] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 74 ("As the majority becomes desensitized by the effects of hate speech, the concern is that some members of society will demonstrate their rejection of the vulnerable group through conduct. Hate speech lays the groundwork for later, broad attacks on vulnerable groups").

[945] Parasocial relations are "one-sided relationships in which fans feel as though they genuinely know and are close to the celebrities whose content they view." See Becca Lewis, "All of YouTube, Not Just the Algorithm, is a Far-Right Propaganda Machine" (8 January 2020), online: *FFWD* <https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430>.

[946] *Canada (Human Rights Commission) v Taylor*, [1990] 3 SCR 892, 1990 CarswellNat 1030 at para 82 (WL).

[947] Becca Lewis, "All of YouTube, Not Just the Algorithm, is a Far-Right Propaganda Machine" (8 January 2020), online: *FFWD* <https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430>.

perspectives or solutions for societal problems".[948] However, the SCC has rejected the primacy of what Bailey deems a "detached and philosophical story of democracy",[949] in determining the constitutionality of laws to protect historically marginalized and vulnerable groups from hate speech.

In upholding the constitutionality of criminal liability for hate speech, the SCC in *Keegstra* cited the following from the Cohen Committee report, which remains all too relevant in the contemporary context of frictionless sharing and amplification, networked abuse, mass trolling, conspiracy-fuelled voter bases, and viral disinformation that outruns and outlasts fact-checking and corrections:

> The Cohen Committee noted that individuals can be persuaded to believe "almost anything" [...] if information or ideas are communicated using the right technique and in the proper circumstances [...]:
>
> > "... we are less confident in the 20th century that the critical faculties of individuals will be brought to bear on the speech and writing which is directed at them. In the 18th and 19th centuries, there was a widespread belief that man was a rational creature [...]
> >
> > We cannot share this faith today in such a simple form. [...] [I]t is too often true, in the short run, that emotion displaces reason and individuals perversely reject the demonstrations of truth put before them and forsake the good they know. [...] We act irresponsibly if we ignore the way in which emotion can drive reason from the field."
>
> [...] Moreover, the alteration of views held by the recipients of hate propaganda may occur subtly, and is not always attendant upon conscious acceptance of the communicated ideas. Even if the message of hate propaganda is outwardly rejected, there is evidence that its premise of racial or religious inferiority may persist in a recipient's mind as an idea that holds some truth, an incipient effect not to be entirely discounted.[950]

The SCC in *Whatcott* also rejected the "marketplace of ideas" argument as insufficient to protect true freedom of expression and its underlying values, let alone the right to equality or the dignity and safety of historically and systemically persecuted groups in society:

> I do not say that the marketplace of ideas may not be a reasonable alternative, and where a legislature is so minded, it will not enact hate speech legislation. However, in *Keegstra*, Dickson C.J. set out a compelling rationale for why Parliament's preference to regulate hate speech through legislation rather than to trust it to the hands of the marketplace was also reasonable. He noted that "the state should not be the sole arbiter of truth, but neither should we overplay the view that rationality will overcome all falsehoods in the unregulated marketplace of ideas" [...]. In his view, paradoxically, hate speech undermines the principles upon which freedom of expression is based and

---

[948] Stanley Ingber, "The Marketplace Of Ideas: A Legitimizing Myth" (1984) 1984:1 Duke Law Journal 1 at 3.

[949] Jane Bailey, "Twenty Years Later *Taylor* Still Has It Right: How the Canadian Human Rights Act's Hate Speech Provision Continues to Contribute to Equality" (2010) 50 Supreme Court Law Review 349 at para 43 (QL).

[950] *R v Keegstra*, [1990] 3 SCR 697, 1990 CarswellAlta 192 at para 66 (WL) (in-text citations omitted).

> "contributes little to the . . . quest for truth, the promotion of individual self-development or the protection and fostering of a vibrant democracy where the participation of all individuals is accepted and encouraged" […]. That is because a common effect of hate speech is to discourage the contributions of the minority. While hate speech may achieve the self-fulfillment of the publisher, it does so by reducing the participation and self-fulfillment of individuals within the vulnerable group. These drawbacks suggest that this alternative is not without its concerns.[951]

Researchers, historians, and equality advocates across law, human rights, science and technology studies, and media and communications studies have also demonstrated how the 'marketplace of ideas', connected to the notion of the 'public sphere' as popularized by the philosopher Jürgen Habermas,[952] was no such utopia even in its time:

> Habermas's influential account of the bourgeois public sphere celebrated the coffeehouses, newspapers, and other forums where 'members of the public' could rationally debate socio-political issues and build consensus. His account of early modern publics depicts spaces where anyone can enter, bracketing their own social status to engage with others as equal peers in deliberation. Fraser notes that these were highly exclusionary spaces, particularly for women, persons of color, and members of the working classes.[953]

That exclusionary nature of public space continues today, not least where it comprises online platforms cultivating hostile environments to members of marginalized groups. Citron writes that "[a]n online discourse which systematically under-represents people—particularly women and people of color—cannot effectively process our various attitudes and convert them into truly democratic decisions."[954]

In addition, the context, business models, design logic, and sociocultural and political forces operating across digital platform environments have only served to further distance the reality of the online public sphere and 'marketplace' from the intended ideal. Specifically, the public forum that many of the largest digital platforms constitute "seems to encourage the distorted 'democratic' idea that all

---

[951] *Saskatchewan (Human Rights Commission) v Whatcott*, 2013 SCC 11 at para 104.

[952] Pieter Boeder, "Habermas' heritage: The future of the public sphere in the network society" (2005) 10:9 First Monday, online: *First Monday* <https://firstmonday.org/article/view/1280/1200>.

[953] R Stuart Geiger, "Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space" (2016) 19 Information, Communication, and Society 787 at 793. Geiger goes on to discuss Fraser's critique of this "public sphere", which almost exactly describes a similar state of affairs today: "Fraser critiques the hegemonic way in which certain public venues for socio-political debate and discussion came to be known as 'the public,' while other kinds of venues and activities that were populated by women, minorities, and working class individuals were excluded from this concept of the public. *Subordinated groups had to enter hostile spaces in order to have their discourse be considered part of the public*, and Fraser extensively reviews the literature about how *dominant groups engage in various practices to silence, intimidate, and chill participation by non-dominant groups*. Fraser's feminist critique recasts the bourgeois public sphere as but one of many public spheres – albeit one that was exclusively assumed to represent the population as a whole." (emphasis added).

[954] Danielle Keats Citron, "Cyber Civil Rights" (2009) 89 Boston University Law Review 61 at 105 ("If expressing opinions online subjects someone to the risk of assault, even if the damage is only temporary, the result will change the kinds of people who participate in online discourse. If we believe the Internet is, and should remain, a wild west with incivility and brutality as the norm, then those who are impervious to such conduct will remain online while the vulnerable may not.").

opinions are equally worthy of respect – regardless of whether they have any factual grounding".[955] Franks notes, regarding the virality of disinformation, "In an age of instantaneous transmission, there often is literally *no time* to correct falsehoods before they go 'viral.' By the time corrections are made — if they are made at all — it is often too late to correct first impressions or undo harm. [...] Repeated exposure to false information, even in a corrective context, increases the likelihood that [it] will be remembered as true."[956] Moreover, trusting the 'marketplace of ideas', or the online public sphere, to autonomously sort out truth or validity from falsehoods or malicious intent assumes that individual Internet users have the wherewithal, knowledge, astuteness, and time to engage in such evaluations on a daily basis; "there is so much information that distinguishing between fact and fiction requires more time and energy than most are willing (or able) to invest."[957]

In fact, the current functioning and dynamics of discourse across online platforms is particularly conducive to behaviours and strategies that involve manipulating and gaming platform features for the purposes of distorting public opinion or silencing marginalized voices. Zeynep Tufekci describes the phenomenon as follows:

> The most effective forms of censorship today involve meddling with trust and attention, not muzzling speech itself. As a result, they don't look much like the old forms of censorship at all. They look like viral or coordinated harassment campaigns, which harness the dynamics of viral outrage to impose an unbearable and disproportionate cost on the act of speaking out. They look like epidemics of disinformation, meant to undercut the credibility of valid information sources. They look like bot-fueled campaigns of trolling and distraction, or piecemeal leaks of hacked materials, meant to swamp the attention of traditional media.[958]

Even the assumption of one single, shared marketplace or public sphere no longer holds, due to audience fragmentation and algorithmically created filter bubbles. According to Tufekci, online speech is "no longer public in any traditional sense".[959] Despite the appearance that social media is "where masses of people experience things together simultaneously […] in reality, posts are targeted and delivered privately, screen by screen by screen. Today's phantom public sphere has been fragmented and submerged into billions of individual capillaries."[960]

Further, Jane Bailey emphasizes that the explicitly profit-driven nature and incentives of digital platforms leave little to be desired in the way of remedying TFGBV, particularly where suppliers of

---

[955] Richard Moon, "The Demise of Freedom of Expression" (19 October 2016), online: *Centre for Free Expression* <https://cfe.ryerson.ca/blog/2016/10/demise-freedom-expression>.

[956] Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (2 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment> (emphasis in original).

[957] Dax D'Orazio, "Freedom of Expression, Misinformation, and Anti-Vaxxers: The Right Thing to Do Is Not Obvious" (25 March 2020), online: *Centre for Free Expression* <https://cfe.ryerson.ca/blog/2020/03/freedom-expression-misinformation-and-anti-vaxxers-right-thing-do-not-obvious>; see also Shaheen Shariff & Karen Eltis, "Addressing Online Sexual Violence: An Opportunity for Partnerships between Law and Education" (2017) 27 Education & Law Journal 99 at 109.

[958] Zeynep Tufekci, "It's the (Democracy-Poisoning) Golden Age of Free Speech" (16 January 2010), online: *Wired* <https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/>.

[959] *Ibid*.

[960] *Ibid*.

online abuse, misogyny, and related hateful or harmful expression are all too happy to meet and stoke ever-present demand.[961] Real-world markets fail to live up to an already flawed metaphor as a suitable model of organization for democratic exchange of ideas and truth-seeking. As Bailey writes, "the economic rationality behind supplying discriminatory content combines with relatively widespread idealization of the Internet as a wide open, anarchic marketplace of ideas, to work against effective restriction of Internet hate propaganda solely through private regulation".[962]

Given all of the above, "the core claim of the marketplace of ideas that permitting all speech, including hate speech, is ultimately valuable because it leads to the discovery and acceptance of truth may well be fundamentally misguided".[963] Those who oppose laws to address TFGBV on the grounds of harming freedom of expression, or for fear of disrupting the 'marketplace', thus continue to rely on a set of premises that may have never been true for anyone other than a "high-status white man",[964] and are even more tenuous today in the context of online platforms and TFGBV.

6.1.4.2. Systemic Inequality and the Limitations of 'Counterspeech'

The second critical contextual factor in assessing proportionality of platform liability for TFGBV requires recognizing that due to systemic inequality, the status quo already amounts to an ongoing state of violation of the right to freedom of expression where women, girls, and others subjected to online violence, abuse and harassment are concerned. This state of violation continues as long as members of such groups are systemically dissuaded from being able to fully exercise that freedom on terms equal to that of individuals from dominant sociopolitical classes, due to the latter perpetrating online abuse and misogyny. As Nicolas Suzor et al. write, "The [alleged] conflict between protecting freedom of speech online and preventing abuse is a false dichotomy that rests on a refusal to account for power among individual users: systemic discrimination and abuse have serious negative impacts on the agency and participation of people who experience them."[965]

Systemic inequality is relevant to interrogating another common argument against enacting laws to prevent or mitigate TFGBV: the idea of 'more speech' or 'counterspeech'. The Dangerous Speech Project defines counterspeech as "any direct response to hateful or harmful speech which seeks to undermine it".[966] While responses may include initiatives such as bystander interventions or fact-checking by

---

[961] Jane Bailey, "Private Regulation and Public Policy: Toward Effective Restriction of Internet Hate Propaganda" (2003) 49 McGill Law Journal 59 at 95-96.

[962] *Ibid* at 96.

[963] Stefan Theil, "Nadine Strossen, *Hate: Why We Should Resist It with Free Speech, Not Censorship*", Book Review (2019) 69 University of Toronto Law Journal 404 at 406.

[964] Kevin Munger, "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment" (2017) 39 Political Behavior 629 at 631.

[965] Nicolas Suzor, Molly Dragiewicz, Bridget Harris, Rosalie Gillett, Jean Burgess & Tess Van Geelen, "Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online" (2019) 11:1 Policy and Internet 84 at 89; see also Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (2 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment> ("The conventional account of the relationship between free speech and autonomy often omits the fact that the exercise of some people's speech risks chilling other people's speech").

[966] "Counterspeech" (last visited 14 April 2021), online: *Dangerous Speech Project* <https://dangerousspeech.org/counterspeech/>.

unaffected parties, the following discussion will only focus on the deficiencies of 'counterspeech' as a remedy for impacted historically marginalized groups and individuals specifically. Systemic inequality and power differentials between those who engage in hateful or systemically harmful speech or conduct, and those who are the targets of such acts or expression, combined with the current context of digital platforms as present-day 'governors' and central arenas of public discourse, militate against the effectiveness of counterspeech as a tenable solution to TFGBV.

First, the chilling effects of online abuse on women and other marginalized groups do not disappear because their speech occurs in circumstances that allow it to be categorized in the abstract as 'counterspeech'. Many scholars, including gender equality and racial justice advocates, have pointed out for decades that the recommendation of 'more speech' or 'counterspeech' "frequently is useless because it may provoke only further abuse or because the insulter is in a position of authority over the victim".[967] As Ruth Coustick-Deal asserts, counterspeech is "available only to those who already have privilege, usually white men. They don't feel the same kind of fear, or live with the constant threat of sexual violence directed at them. It is easy to advocate counter speech when you can engage in it freely and without repercussions."[968]

Second, "[w]hether women can fight speech with more speech depends on whether, and to what extent, women can speak".[969] Systemic gender inequality rooted in historical and ongoing oppression continues to raise and maintain formal and informal barriers to women's speech, even in the absence of TFGBV and platformed misogyny. Coustick-Deal notes that "[s]tructural bias that favours white male voices in entertainment, publishing, and the film industry means that for every racist film or book on one end of the seesaw, a number of obstacles prevent any counter[speaker] from just hopping on at the other side. Balance is a myth."[970] Becky Gardiner similarly observes that "the hostility to women and people of colour below the line [in the comments sections of news articles] mirrors a historical institutional hostility to women and people of colour 'above the line'—the discriminatory hiring and commissioning practices over many decades that have left them struggling to get published at all."[971] When it comes to those who speak and those who are spoken about, the direction of power between them matters: "powerful people can generally do more, say more, and have their speech count for more than can the powerless. If you are powerful, there are more things you can do with your words."[972]

---

[967] Richard Delgado, "Words That Wound: A Tort Action for Racial Insults, Epithets, and Name Calling" in Mari J Matsuda et al, *Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment* (New York: Routledge, 1993) 89 at 95.

[968] Ruth Coustick-Deal, "What's wrong with counter speech?" (6 February 2017), online (blog): *Ruth Coustick-Deal* <https://medium.com/@ruthcoustickdeal/https-medium-com-whats-wrong-with-counter-speech-f5e972b13e5e>.

[969] Rae Langton, "Speech Acts and Unspeakable Acts" (1993) 22 Philosophy and Public Affairs 293 at 314.

[970] Ruth Coustick-Deal, "What's wrong with counter speech?" (6 February 2017), online (blog): *Ruth Coustick-Deal* <https://medium.com/@ruthcoustickdeal/https-medium-com-whats-wrong-with-counter-speech-f5e972b13e5e>. Coustick-Deal later writes, "There is a sad irony that we are calling for a strategy to counter hate speech, which is usually against oppressed groups — such as people of colour, [I]ndigenous people, women, and queer people — but which is actually *least available* to those groups" (emphasis in original).

[971] Becky Gardiner, "'It's a terrible way to go to work:' what 70 million readers' comments on the Guardian revealed about hostility to women and minorities online" (2018) 18:4 Feminist Media Studies 592 at 605.

[972] Rae Langton, "Speech Acts and Unspeakable Acts" (1993) 22 Philosophy and Public Affairs 293 at 299; see also Charles R Lawrence III et al, "Introduction" in Mari J Matsuda et al, *Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment* (New York: Routledge, 1993) 1 at 10 ("Matsuda draws a distinction between dissent—or the right to criticize the powerful institutions that govern our lives—and hate speech, which is directed against the least powerful segments of

'Counterspeech' would not be necessary if women's, and Black, Indigenous, and other racialized people's expression was elevated and protected as simply 'speech' equal to that of their oppressors in the first place.

Third, power differentials and systemic oppression render counterspeech ineffective where misogynistic, sexist, racist, or otherwise discriminatory or unconsciously biased readers or listeners disregard or dismiss speech precisely because its speaker is from one or more such marginalized demographics. One prominent study on countering racist harassment on Twitter used bots with different profile characteristics to rebuke real users tweeting racist slurs, with all bot accounts sending the same message. The study found that "messages sent by white men caus[ed] the largest reduction in offensive behavior among a subject pool of white men," and moreover, only white men with a "high number of Twitter followers" reduced harassment, and only temporarily (the effect lasted a month).[973] In contrast, "there was actually an *increase in racist harassment* among the subjects who received a message sent by a black bot with few followers".[974]

Kevin Munger's study provides empirical support to the reasoned criticism above of the 'counterspeech' approach: 'counterspeech' messaging was only effective when the speaker was perceived to be "a high-status white man", and if engaged in by the very victims who are the targets of such harassment, the method would have only resulted in more of the precise abuse that 'counterspeech' is purportedly a key solution to reducing. Further, there is the possibility that the gamification of online abuse, including the context-collapsing repurposing of women's or racialized persons' expression as mere targets for entertainment, results in what Rae Langton describes as a form of speech *disablement*—when one's speech no longer counts as anyone having spoken at all:

> If you are powerful, you sometimes have the ability to silence the speech of the powerless. One way might be to stop the powerless from speaking at all. Gag them, threaten them, condemn them to solitary confinement. But there is another, less dramatic but equally effective, way. Let them speak. Let them say whatever they like to whomever they like, but stop that speech from counting as an action. More precisely, stop it from counting as the action it was intended to be. […] Some speech acts are unspeakable for women in some contexts: although the appropriate words can be uttered, those utterances fail to count as the actions they were intended to be.[975]

In the case of Anita Sarkeesian and other outspoken feminist gamers, for example, their online expression was stripped of any possible effect as 'counterspeech' meant to illuminate search for truth, and repurposed into mere fodder for further abuse and profit by the very harassers against whom they were 'counterspeaking'.[976]

---

our community. This distinction, Matsuda argues, is a principled one, given the historical contexts of subordination that she uses as a starting point for developing legal theory".)

[973] Kevin Munger, "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment" (2017) 39 Political Behavior 629 at 631.

[974] *Ibid* (emphasis added).

[975] Rae Langton, "Speech Acts and Unspeakable Acts" (1993) 22 Philosophy and Public Affairs 293 at 299.

[976] Alice E Marwick & Robyn Caplan, "Drinking male tears: language, the manosphere, and networked harassment" (2018) 18 Feminist Media Studies 543 at 544 ("In a statement on her blog, Sarkeesian wrote: 'Carl is a man who literally profits from harassing me and other women: he makes over $5,000 a month on Patreon for creating YouTube videos that mock, insult and

Fourth, counterspeech may cause further damage, in cases of what Mary Anne Franks terms 'unanswerable speech', which comprises many forms of platformed TFGBV. "There is no 'counterspeech' to the nonconsensual publication of a person's nude image, the dissemination of a home address, or the disclosure of undocumented status. No 'process of education' can undo their damage."[977] It would likely cause re-traumatization and further psychological harm and injury to dignity, to expect a targeted woman or girl to enter into a 'reasoned debate' with 'more speech' about the release of her nude photos without consent, or with a man sending her rape threats or death threats.

Fifth and lastly, reliance on 'counterspeech' as a primary method of battling online abuse inappropriately places the burden on those already targeted and victimized by such abuse, while potentially exposing them to even further escalated harassment in the process.[978] As Coustick-Deal writes, "When we ask for counter speech, rather than a removal of content in response to hate speech, we are placing a huge burden upon an oppressed group to spend their time and energy on speaking back to their oppressors— facing harassment, threats, and more oppression when they do so. Rather than holding oppressors accountable, we once again place the burden on the oppressed to carry out further labour just to defend their existence."[979]

6.1.4.3. Private Abuse as an Ongoing Threat to Historically Marginalized Groups

The third essential contextual factor in effectively addressing TFGBV on digital platforms is recognizing that for historically marginalized groups and individuals, such as women and girls, violence and abuse of power by private, non-state actors can be as great and relentless a threat as violence or abuse of power by the state—if not more so—depending on the circumstances. Often, women and girls are subjected to threats from both private and public actors. This is particularly the case for those whose identities place them at the intersection of multiple historically marginalized groups. Acknowledging this reality is crucial to correctly assessing the proportionality of state action that restricts the ability of private individuals from dominant social groups to perpetrate TFGBV against individuals who are systemically marginalized and more vulnerable to both private and public abuse.

---

discredit myself and other women online, and he's not alone. He is one of several YouTubers who profit from the cottage industry of online harassment and antifeminism; together, these people have millions of followers who are regularly encouraged by the videos and tweets of these individuals to harass me and other women who make videos daring to assert the basic humanity of women, people of color, trans folks, and members of other marginalized groups (Anita Sarkeesian 2017)'").

[977] Mary Anne Franks, "The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?" (2 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment>; see also Danielle Keats Citron, "Restricting Speech to Protect It" in Susan J Brison & Katharine Gelber, eds, *Free Speech in the Digital Age* (New York: Oxford University Press, 2019) 122 at 130 ("Posts with a woman's nude photo, home address, and supposed interest in sex are not facts or ideas to be debated in the service of truth. When dealing with falsehoods impugning someone's character, the victim does not have an affirmative case she is trying to convey—she is only seeking to dispel the harm from posters' attacks. Even if victims could respond, their replies may never be seen. The truth may be unable to emerge from a battle of posts. Images of a private person's naked body have little value to the general public and can destroy that person's career. They ensure that victims are undateable, unemployable, and unable to partake in online activities. Furthermore, as Professor Daniel Solove aptly notes, 'Truth isn't the only value at stake'." (in-text citations omitted)).

[978] Ruth Coustick-Deal, "What's wrong with counter speech?" (6 February 2017), online (blog): *Ruth Coustick-Deal* <https://medium.com/@ruthcoustickdeal/https-medium-com-whats-wrong-with-counter-speech-f5e972b13e5e>.

[979] *Ibid*.

The SCC has warned: "In interpreting and applying the *Charter* I believe that the courts must be cautious to ensure that it does not simply become an instrument of better situated individuals to roll back legislation which has as its object the improvement of the condition of less advantaged persons."[980] Legal reforms to address TFGBV must not fall casualty to the privileged lens of "better situated individuals". Bailey elaborates on the impacts of discriminatory private ordering for marginalized social groups, in the absence of public regulation:

> To say that one's ability to engage and participate in a marketplace of ideas depends on being "free from" government intrusion needs to be recognized as a privileged understanding. For many socially vulnerable groups, historic exclusion from participation did not emanate from government action, but rather from discriminatory forms of private social ordering. From this vantage point, state inaction reinforces discriminatory private ordering.
>
> State action to limit discrimination may be experienced not as an unwelcome intrusion on freedom, but as a freedom enhancer for those whose ability to participate has been impeded by privately imposed forms of bigotry and control. In fact, for those whose democratic participation has been limited by non-state actors' discriminatory conduct, equality may appear as the matrix from which the substantive ability to experience other rights and freedoms flows. For many, without intervention to level the playing field in the marketplace of ideas, the right to freedom of expression may well ring hollow.
>
> From this vantage point, courts need to be cautious about situations in which majoritarian viewpoints are mischaracterized as political dissent that must be assiduously protected from state intrusion. Approaching discrete *Charter* rights as equally legitimate components of a matrix that must be read together creates an environment more amenable to this principled and democratically vital analysis.[981]

Canadian courts have also recognized that "there is a distinction to be drawn between legislation that 'acts as the "singular antagonist of the individual"' (e.g., criminal justice legislation) and legislation that mediates between different groups (e.g., social legislation)",[982] finding that "a lower standard of s. 1 justification may be appropriate" in the case of the latter.[983] This is significant given Mari Matsuda's observation that the "kinds of injuries historically left to private individuals to absorb and resist through private means are no accident. […] [A]bsence of law is itself another story with a message, perhaps unintended, about the relative value of different human lives."[984]

---

[980] *R v Edwards Books and Art Ltd*, [1986] 2 SCR 713 at 779.

[981] Jane Bailey, "Twenty Years Later *Taylor* Still Has It Right: How the Canadian Human Rights Act's Hate Speech Provision Continues to Contribute to Equality" (2010) 50 Supreme Court Law Review 349 at paras 48-49 (QL).

[982] *Crouch v Snell*, 2015 NSSC 340 at para 121, citing *RJR-MacDonald Inc. v Canada (Attorney General)*, [1995] 3 SCR 199, *Harper v Canada (Attorney General)*, 2004 SCC 33.

[983] *RJR-MacDonald Inc. v Canada (Attorney General)*, [1995] 3 SCR 199, 1995 CarswellQue 119 at para 68 (WL).

[984] Mari J Matsuda, "Public Response to Racist Speech: Considering the Victim's Story" in Mari J Matsuda et al, *Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment* (New York: Routledge, 1993) 17 at 18.

6.1.5. Considerations in Legislative Drafting

Two further aspects of legislation are key to assessing its constitutionality: clarity and precision of drafting (or whether there is an 'intelligible standard') and whether the legislation is remedial or punitive in nature.

6.1.5.1. Intelligible Standard

The subject matter of legislation must be sufficiently clearly defined so as to provide an 'intelligible standard' to adhere to in making decisions or determining whether or not particular circumstances are captured by the law. Where legislation restricting certain types of expression is concerned, both clarity and specificity are critical in ensuring the law is not struck down for vagueness, overbreadth, or arbitrariness. For example, *Taylor* noted that section 13(1) of the *CHRA* was constitutional in part because the "terms of the section, in particular the phrase 'hatred or contempt', are sufficiently precise and narrow to limit its impact to those expressive activities which are repugnant to Parliament's objective of promoting equality and tolerance in society."[985] This conclusion, as well as the SCC's decision in *Keegstra* upholding as constitutional the *Criminal Code*'s hate speech provision, further delineated the extreme level of sentiment required to constitute 'hatred' or 'contempt'.

In contrast, the Supreme Court of Nova Scotia struck down the provincial *Cyber-safety Act* as an unjustifiable violation of freedom of expression. A central element of the law's failing was the sheer breadth of expression captured in the words of the legislation, without any built-in accounting for context (leaving it open to, for instance, potential exploitation by the powerful and the privileged being told uncomfortable truths):

> I must consider all the types of expression captured by the *Act*. The *Act* restricts "any electronic communication through the use of technology […] that is intended or ought reasonably be expected to cause fear, intimidation, humiliation, distress or other damage or harm to another person's health, emotional well-being, self-esteem or reputation, and includes assisting or encouraging such communication in any way". It is not difficult to come up with examples of expressive activity that falls within this definition, and at the same time promotes one of the core freedom of expression values. […]
>
> I find that the *Act* provides no intelligible standard according to which Justices of the Peace and the judiciary must do their work. It does not provide sufficiently clear standards to avoid arbitrary and discriminatory applications. The Legislature has given a plenary discretion to do whatever seems best in a wide set of circumstances. There is no "limit prescribed by law" and the impugned provisions of the *Act* cannot be justified under s. 1.[986]

Given the wide range of expression that constitutes TFGBV, combined with the intermediary role that platforms play in facilitating all manner of user expression and the legal obligations that would be imposed upon platform companies to govern user expression on their platforms, attentiveness to clarity and precision in legislative drafting is paramount for withstanding constitutional scrutiny. For

---

[985] *Canada (Human Rights Commission) v Taylor*, [1990] 3 SCR 892, 1990 CarswellNat 1030 at para 81 (WL).

[986] *Crouch v Snell*, 2015 NSSC 340 at paras 115, 137.

example, if a law were drafted to target 'online harassment', how would 'online harassment' be defined such that digital platforms could act on that definition in their content moderation decisions, with desired results? Moreover, how would the definition exclude powerful or privileged individuals who consider themselves harassed under circumstances where they are merely confronted with legitimate criticism from historically marginalized voices, and prevent those individuals from turning the law against the very groups it is meant to protect? One potential response to the latter would be to explicitly build in principles of substantive equality and recognition of inherent power imbalances between individuals from historically marginalized groups and individuals from sociopolitically dominant groups. This is just one example, however, illustrating issues of definition, interpretation, workability, and unintended consequences that must be carefully thought through.

In addition to subject matter, another element of legislation targeting TFGBV may include considering the range of potential responses that platforms can leverage to address TFGBV and similarly harmful expression. For example, where less extreme content is concerned, or content that could not be constitutionally taken down by law, platforms might instead restrict to private view, downrank, remove sharing capabilities on, stop recommending, or otherwise halt the amplification of certain content, without necessarily removing it from the platform altogether. This may influence the minimal impairment analysis under section 1.

Those engaging in online abuse would not be silenced, but 'deplatformed' on principled grounds of equality and freedom from discrimination, as well as genuine commitment to freedom of expression for all. As Renee DiResta put it, "[F]ree speech does not mean free reach. There is no right to algorithmic amplification."[987] Similarly, Joan Donovan and danah boyd note that all may "have the right to speak their minds, but not every person deserves to have their opinions amplified, particularly when their goals are to sow violence, hatred and chaos".[988] Even in the event that speech is removed from a particular platform, it should be noted, the speaker remains able to speak anywhere else on the Internet—and still without the violent repercussions that women and Black, Indigenous, and other racialized people risk facing when speaking out online.[989]

6.1.5.2. Nature of the Legislation

The reasons in *Taylor* suggest that whether a law targeting speech is remedial or punitive will impact the courts' analysis of its constitutionality. Section 13(1) of the *CHRA*—the law at issue in *Taylor*—was part of a remedial and salutary human rights regime, rather than a punitive law.[990] In its proportionality analysis, the SCC took into account that the primary aim of the *CHRA* was to provide ameliorative relief and redress to those targeted by hate and discrimination, rather than "bring the full force of the state's

---

[987] Renee DiResta, "Free Speech Is Not the Same As Free Reach" (30 August 2018), online: *Wired* <https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/>.

[988] Joan Donovan & Danah Boyd, "The case for quarantining extremist ideas" (1 June 2018), online: *Guardian*, <https://www.theguardian.com/commentisfree/2018/jun/01/extremist-ideas-media-coverage-kkk>.

[989] Zeynep Tufekci, "It's the (Democracy-Poisoning) Golden Age of Free Speech" (16 January 2010), online: *Wired* <https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/> ("Even when the big platforms themselves suspend or boot someone off their networks for violating 'community standards'—an act that does look to many people like old-fashioned censorship—it's not technically an infringement on free speech, even if it is a display of immense platform power. Anyone in the world can still read what the far-right troll Tim 'Baked Alaska' Gionet has to say on the internet. What Twitter has denied him, by kicking him off, is attention.").

[990] *Canada (Human Rights Commission) v Taylor*, [1990] 3 SCR 892, 1990 CarswellNat 1030 at paras 37, 61, 69 (WL).

power against a blameworthy individual for the purpose of imposing punishment".[991] This gives rise to two implications in the context of legal reforms to impose platform liability, or platform accountability, for TFGBV.

The first implication is that a remedial and regulatory regime that centres the experiences of those targeted by TFGBV, and is focused on prevention, expedient relief, and remedy, may be not only more effective but also more likely to pass constitutional muster. Given that such a law would place legal obligations on large technology companies, it is also important to note that financial penalties for compliance do not turn an otherwise regulatory regime into a punitive one, as confirmed in *Lemire*:

> In my view, when the penalty provisions are considered in the context of the objectives of the CHRA and its remedial scheme, they are not properly characterized as penal in nature. […] Like the financial penalties often contained in other regulatory legislation, [the fines are] designed to induce compliance with the statutory scheme in order to impose a measure of financial accountability on those in breach of section 13 [of the *CHRA*] and to deter future breaches..[992]

The second implication is that laws addressing TFGBV should not be jeopardized by bundling TFGBV with efforts to address other, less related issues through forms of platform liability or platform regulation, especially issues where the state has historically played a central role as 'antagonist of the individual'. In this light, it is potentially concerning that the Departments of Justice and Canadian Heritage are purporting to address all of "hate speech, child pornography, incitement to violence, incitement to terrorism and the non-consensual disclosure of images" in one single regulatory framework.[993] Seen through an intersectional feminist lens, this may risk putting many in the difficult position of having either to temper support for a law that addresses hate speech, child pornography, and NCDII—remedying devasting forms of abuse by private, non-state actors—or lend support to what may at least in part be an 'anti-terrorism' initiative, knowing that national security concerns have for years been an undeniable and persistent source of state abuse against individuals, particularly on grounds of race and religion. Legislative reforms that are genuinely intended to meaningfully address TFGBV must ensure they are sufficiently tailored to TFGBV and related systemic oppressions, in order to ensure a strong constitutional basis.

---

[991] *Ibid* at para 37 (WL).

[992] *Lemire v Canada (Human Rights Commission)*, 2014 FCA 18 at paras 90-91.

[993] Canada, Parliament, House of Commons, Standing Committee on Canadian Heritage, *Evidence*, 43rd Parl, 2nd Sess, No 12 (29 January 2021) at 4 (Evidence of the Hon. Steven Guilbeault, Minister of Heritage), online: <https://www.ourcommons.ca/Content/Committee/432/CHPC/Evidence/EV11074629/CHPCEV12-E.PDF>.

ISSUE SPOTLIGHT NO. 2
Criminal Justice and Law Enforcement in Platform Liability for TFGBV

Some legal reform and policy proposals to strengthen platform liability for TFGBV have called for a greater role for law enforcement and the criminal justice system. Such proposals are not without concern, while recognizing that there remains a role for the criminal law to play in addressing TFGBV.

The criminal justice system and law enforcement in Canada and in peer jurisdictions have overwhelmingly disproportionately targeted Black, Indigenous, and other racialized persons, as a result of deeply rooted systemic racism and discrimination at nearly every level.[994] In a comprehensive study of criminal justice cases in Canada involving TFGBV, and while advocating for the legitimate role of criminal law in addressing TFGBV, Bailey and Mathen note that these cases, like the rest of the criminal justice system, "reflect systemic oppressions and prejudices that lead, among other things, to the overpolicing of Indigenous and Black community members, and a greater likelihood that convictions of members of marginalized communities will result from plea bargains that tend to go unreported," in addition to "alarming rates of police refusal to press charges, leading to drastic under-representation of the prevalence and nature of sexual violence in criminal courtrooms".[995] It is thus important that platform liability proposals to address TFGBV avoid "exacerbate[ing] the negative interactions marginalized communities have with police and the state, leading to under-reporting and under-representation of their experiences of violence".[996]

Building law enforcement and the criminal justice system directly into platform regulation regimes may also empower the state to further exploit rhetoric around women's and children's safety as opportunistic cover for advancing state surveillance capabilities and expanding the legal boundaries of law enforcement's ability to interfere with people in everyday spheres of activity.[997] This is a routine public messaging strategy deployed by governments attempting to justify the introduction of new or farther reaching law enforcement powers in the context of criminal justice or national security. For example, former federal Public Safety Minister Vic Toews infamously said that critics of expansive

---

[994] See e.g., Kate Robertson, Cynthia Khoo & Yolanda Song, "To Surveil and Predict: A Human Rights Analysis of Algorithmic Policing in Canada" (2020) at 15-18, 107-108, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2020/09/To-Surveil-and-Predict.pdf>.

[995] Jane Bailey & Carissima Mathen, "Technology-Facilitated Violence Against Women & Girls: Assessing the Canadian Criminal Law Response" (2019) 97:3 The Canadian Bar Review 664 at 664, 668.

[996] *Ibid* at 668. Bailey and Mathen also state at 673 that "criminalization of certain behaviours in a social context rife with the over-policing of subordinated communities can also *undermine* equality by disproportionately exposing members of those communities to criminal sanction. Given the social equality consequences at stake, it is essential to be able to articulate clearly principled bases for concluding that criminal sanction is a just and warranted response to an emergent social problem".

[997] Ronald J Deibert et al., "Submission of the Citizen Lab (Munk School of Global Affairs, University of Toronto) to the United Nations Special Rapporteur on violence against women, its causes and consequences, Ms. Dubravka Šimonović" (2 November 2017) at 1, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2017/11/Final-UNSRVAG-CitizenLab.pdf> ("[W]e raise questions about narratives that capitalize on the vulnerability of women and girls in order to justify new powers to surveil, de-anonymize, police, and censor in the digital sphere. There is limited evidence to suggest that providing greater generalized powers to law enforcement leads to better outcomes for women or other marginalized and vulnerable groups. In some cases, doing so may also increase opportunities and technological capabilities for abuse").

'lawful access' provisions that would enable mass surveillance and mandate backdoors to encrypted communications would "either stand with us or with the child pornographers".[998] In another case, expansive digital surveillance powers proposed in Bill C-13, *Protecting Canadians from Online Crime Act*, "were presented to the Canadian public as a response to the incidents of online abuse that led to the tragic suicide of Rehtaeh Parsons. However, an independent inquiry into the police response concluded that law enforcement had already possessed the necessary search powers and grounds to investigate […]; they simply failed to use them due to a lack of training in identifying legal wrongs in a technologically mediated context."[999]

Based on the track record of Canadian law enforcement and national security activities, increasing surveillance powers on digital platforms and establishing new information-sharing arrangements between platform companies and law enforcement will in turn most likely be used to interfere with the human rights of Black, Indigenous, and other racialized people as well as equality-seeking social movements in a disproportionate and discriminatory manner.[1000] Yet, these are the very people whom platform liability frameworks for TFGBV are intended to benefit.

Moreover, law enforcement and the criminal justice system have consistently failed to actually protect—and have more often than not contributed to the revictimization and further traumatization of—women as victims and survivors of gender-based and sexual violence, abuse, and harassment, including intimate partner and dating violence, whether technology-facilitated or not. In these cases, law enforcement agencies neglect to use the powers they already have. This has been especially the case for Black, Indigenous, and other racialized women, or those from the 2SLGBTQQIA communities or who live with a disability or mental illness. Thus, those living at one or more intersections of race, disability, and/or sexual orientation, combined with being a woman or gender-diverse individual, are more likely to be exposed to greater harm by law enforcement, while not receiving additional help in responding to private gender-based violence, abuse, and harassment in any case. Suzor et al. state:

> While some new laws have been useful to some survivors of GBV [gender-based violence] online and off, criminal legal responses remain inconsistently applied, often along intersectional lines of structural inequality that continue to exempt wealthy, white men from consequences while posing disproportionate unintended consequences for poor women of color. Accordingly, many criminologists and antiviolence advocates are challenging the idea that the benefits of criminal legal approaches to GBV outweigh the harms of mass incarceration […].[1001]

---

[998] "Online surveillance critics accused of supporting child porn", *CBC News* (13 February 2012), online: <https://www.cbc.ca/news/technology/online-surveillance-critics-accused-of-supporting-child-porn-1.1196829>.

[999] Ronald J Deibert et al, "Submission of the Citizen Lab (Munk School of Global Affairs, University of Toronto) to the United Nations Special Rapporteur on violence against women, its causes and consequences, Ms. Dubravka Šimonović" (2 November 2017) at 3-4, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2017/11/Final-UNSRVAG-CitizenLab.pdf>.

[1000] See e.g., Kate Robertson, Cynthia Khoo & Yolanda Song, "To Surveil and Predict: A Human Rights Analysis of Algorithmic Policing in Canada" (2020) at 97-98, n 444, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2020/09/To-Surveil-and-Predict.pdf>.

[1001] Nicolas Suzor et al., "Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online" (2019) 11:1 Policy and Internet 84 at 90 (in-text citations omitted).

Uncritical calls to strengthen police powers tend to ignore the ways they have been used selectively to punish, disenfranchise, and oppress Black, Indigenous, and other racialized communities, among other marginalized populations, while simultaneously under-policing and ignoring—or empathizing with and treating with undue leniency—perpetrators from dominant social classes, i.e., white men.[1002] Moreover, members of law enforcement themselves number among perpetrators of TFGBV. "[T]here are unfortunately many examples where such individuals have leveraged state surveillance tools in order to stalk former partners […] Within the intelligence community, this form of abuse of power is so common that it has its own name: LOVEINT ['love intelligence']."[1003]

At the same time, the criminal justice system, to the extent we continue to contend with its existence and current form, remains a significant part of the overall constellation of legal and other responses required to address such a pervasive, multifaceted, and devastating problem as TFGBV.[1004] As Bailey and Mathen state, "Behaviours such as nonconsensual distribution of intimate images should be understood not only as individual harms, but also as public wrongs violating sexual integrity, autonomy and equality […] Any systemic refusal or failure to apply criminal law would suggest that such attacks, and the crucible of subordination in which they are incubated, are not to be understood as harms worthy of public sanction."[1005]

Platform liability law should hold platform companies accountable in a way that mitigates risk of the above impacts of discriminatory criminalization flowing through to marginalized users. For example, criminal liability could apply to platforms and their owners who deliberately or knowingly generate or encourage TFGBV—i.e., "purpose-built platforms", such as 'The Dirty'. The concern is not with criminalizing platform companies or their owners, who have tended to be from the most privileged and powerful social groups;[1006] it is the high risk of damaging impact on individual users who belong to historically marginalized groups that are already subject to discriminatory over-criminalization.

[1002] "Beyond practices of neglect in relation to Indigenous and racialized crime victims, under-policing can take a very different form – 'favoritism toward an offending class.' Given limited law enforcement resources, racial profiling, as a manifestation of over-policing directed toward Indigenous and racialized populations, can entail the under-policing of [w]hite people who are engaged in criminal activity. The OHRC describes these race-specific patterns of law enforcement as preferential under-policing." (footnotes omitted). "Policy on eliminating racial profiling in law enforcement" (20 June 2019) at 2.2, online: *Ontario Human Rights Commission* <http://www.ohrc.on.ca/en/policy-eliminating-racial-profiling-law-enforcement>. While this observation emerges from the context of police racism, preferential under-policing of white offenders based on race is compounded by the same based on gendered under-policing of men for TFGBV.

[1003] Ronald J Deibert et al, "Submission of the Citizen Lab (Munk School of Global Affairs, University of Toronto) to the United Nations Special Rapporteur on violence against women, its causes and consequences, Ms. Dubravka Šimonović" (2 November 2017) at 4-5, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2017/11/Final-UNSRVAG-CitizenLab.pdf>; see also Joseph Cox, "Military, FBI, and ICE Are Customers of Controversial 'Stalkerware'", *Vice Motherboard* (23 February 2018), online: <https://motherboard.vice.com/en_us/article/ywqqkw/mili-tary-fbi-and-ice-are-customers-of-controversial-stalkerware>; Alina Selyukh, "NSA staff used spy tools on spouses, ex-lovers: watchdog", *Reuters* (27 September, 2013), online: <https://www.reuters.com/article/us-usa-surveillance-watchdog/nsastaff-used-spy-tools-on-spouses-ex-lovers-watchdogidUSBRE98Q14G20130927>; and Letter from Dr George Ellard, Inspector General, National Security Agency Central Security Service, to Senator Charles E Grassley (11 September 2013), online: *National Security Agency Central Security Service* <https://www.nsa.gov/news-features/press-room/statements/assets/files/grassley-letter.pdf>.

[1004] Jane Bailey & Carissima Mathen, "Technology-Facilitated Violence Against Women & Girls: Assessing the Canadian Criminal Law Response" (2019) 97:3 The Canadian Bar Review 664 at 666.

[1005] *Ibid* at 666, 673.

[1006] See e.g., Liza Mundy, "Why Is Silicon Valley So Awful to Women?", *Atlantic* (April 2017), online: <https://www.theatlantic.com/magazine/archive/2017/04/why-is-silicon-valley-so-awful-to-women/517788/>; Priya Anand

6.2. Challenges with Platform Liability for User Expression

Holding digital platforms liable for harmful speech and conduct in which their users engage raises a host of legal, conceptual, and logistical complications. This has been evidenced by the wide range and varying extents of success and failure of such attempts over the past several years across multiple contexts in jurisdictions around the world, and before that, in the context of online intermediaries such as Internet service providers. These challenges have been discussed at great length by other intermediary liability and platform regulation experts and digital rights organizations elsewhere,[1007] and an in-depth analysis of them is beyond the scope of this report. However, this section will provide a brief overview of three major challenges that should be taken into consideration in legal reform efforts to hold digital platforms liable for their users engaging in online violence, abuse, and harassment.

The three issues are: the risk of overbroad and wrongful takedowns of legal speech; the wide variability of platforms and the wide variability of harms; and potential issues with entrenching privatized governance of public discourse and privatized speech regulation.

6.2.1. Wrongful Takedowns of Legitimate Expression

Laws that incentivize online platforms to remove unlawful content without countervailing incentives to leave up legitimate or beneficial content will most likely result in overbroad and wrongful removal of beneficial, lawful, or otherwise legitimate online expression.[1008] Research abounds that demonstrates this impact, though much of it concerns the copyright context in particular.[1009] This is largely due to

---

& Sarah McBride, "For Black CEOs in Silicon Valley, Humiliation Is a Part of Doing Business", *Financial Post* (16 June 2020), online: <https://financialpost.com/pmn/business-pmn/for-black-ceos-in-silicon-valley-humiliation-is-a-part-of-doing-business>; Dominic Rushe, "Twitter's diversity report: white, male and just like the rest of Silicon Valley", *Guardian* (25 July 2014), online: <https://www.theguardian.com/technology/2014/jul/25/twitter-diversity-white-men-facebook-silicon-valley>; and Arielle Pardes, "Yet Another Year of Venture Capital Being Really White" (29 December 2020), online: *Wired* <https://www.wired.com/story/venture-capital-2020-still-really-white/>.

[1007] See e.g., Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harvard Law Review 1598; Michael Karanicolas, "Squaring the Circle Between Freedom of Expression and Platform Law" (2019-2020) 20 Journal of Technology Law & Policy 177; Daphne Keller, "Who Do You Sue? State and Platform Hybrid Power over Online Speech" (2019), online (pdf): *Hoover Institution* <https://assets.documentcloud.org/documents/5699593/Who-Do-You-Sue-State-and-Platform-Hybrid-Power.pdf>; Karl Bode, "That's A Wrap: Techdirt Greenhouse Content Moderation Edition" (16 September 2020), online: *Techdirt* <https://www.techdirt.com/articles/20200915/09054045310/thats-wrap-techdirt-greenhouse-content-moderation-edition.shtml>; and Mike Masnick, "Protocols, Not Platforms: A Technological Approach to Free Speech" (21 August 2019), online: *Knight First Amendment Institute at Columbia University* <https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech>.

[1008] Mike Masnick, "Rights Groups Demand Facebook Set Up Real Due Process Around Content Moderation" (15 November 2018), online: *Techdirt* <https://www.techdirt.com/articles/20181113/17312841045/rights-groups-demand-facebook-set-up-real-due-process-around-content-moderation.shtml> ("[W]hile there are all sorts of concerns about content moderation, the number of false positives that lead to 'good' content being taken down is staggering. Lots of people like to point and laugh at these, but any serious understanding of content moderation at scale has to recognize that when you need to process many many thousands of requests per day, often involving complex or nuanced issues, many, many mistakes are going to be made".)

[1009] Daphne Keller, "Empirical Evidence of Over-Removal by Internet Companies Under Intermediary Liability Laws: An Updated List" (8 February 2021), online: *The Center for Internet and Society at Stanford Law School* <https://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws>.

platform companies erring on the side of caution and being unwilling to risk legal liability in the event of incorrectly deciding the legality of a particular post.[1010] It may also be due to the implementation of automated content moderation tools in order to rapidly assess high volumes of content, which then inevitably err based on inability to parse context, for example.[1011]

The intermediary liability regime under the US *Digital Millennium Copyright Act* (DMCA) is perhaps the foremost case study in years of widespread wrongful removals of legitimate content without due process.[1012] In their review of empirical research concerning wrongful removals by online platforms, particularly under notice-and-takedown frameworks, Daphne Keller and Paddy Leerssen note that platforms "have removed information ranging from journalism and videos documenting police brutality in Ecuador [...] to media coverage of fraud investigations in the United States [...] to criticism of religious organizations [...] to scientific reporting".[1013] Another study found that most platforms complied with copyright takedown notices submitted on famous works of literature that had clearly already entered the public domain.[1014] Citron writes of similar consequences associated with incentivizing efforts to remove hate speech.[1015]

In addition to erroneous takedowns by the platform's internal decision-making or the fallibility of automated tools, enacting platform liability regimes with built-in content takedown mechanisms also enables bad faith actors to exploit such processes as a silencing weapon against others' speech. Rebecca Tushnet relates: "If you had asked me ten years ago, I would have been skeptical that a person would pretend to be another person's parent, a police officer, or a lawyer representing a copyright claimant in order to get another person's account closed. I no longer need to believe in these things—I've seen them."[1016] In 2017, the Vancouver Aquarium attempted to use Canadian copyright law to

---

[1010] Ronald J Deibert et al, "Submission of the Citizen Lab (Munk School of Global Affairs, University of Toronto) to the United Nations Special Rapporteur on violence against women, its causes and consequences, Ms. Dubravka Šimonović" (2 November 2017) at 13, online (pdf): *Citizen Lab* <https://citizenlab.ca/wp-content/uploads/2017/11/Final-UNSRVAG-CitizenLab.pdf> ("Imposing liability on platforms and other intermediaries for user-generated content frequently leads to overbroad censorship. Platforms faced with a choice between assuming the potential liability of a user as their own and preemptively removing contested content more often than not err on the side of content removal").

[1011] *Ibid* ("Premising liability immunities for third party content in this manner also encourages automation of takedown responses—particularly by central intermediaries who are used by billions of individuals around the world and who therefore face large volumes of allegedly infringing content with no incentive to conduct case-by-case assessments of the underlying legitimacy of allegations." (footnotes omitted)).

[1012] See Lumen, online: <https://lumendatabase.org/> (formerly "Chilling Effects"); and "Takedown Hall of Shame", online: *Electronic Frontier Foundation* <https://www.eff.org/takedowns>.

[1013] Daphne Keller & Paddy Leerssen, "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation" in Nathaniel Persily & Joshua A Tucker, eds, *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge: Cambridge University Press, 2020) 220 at 222 (in-text citations omitted).

[1014] *Ibid* at 239.

[1015] Danielle Keats Citron, "What to Do about the Emerging Threat of Censorship Creep on the Internet" (28 November 2017) at 4, online (pdf): *CATO Institute* <https://object.cato.org/sites/cato.org/files/pubs/pdf/pa-828.pdf> ("As more expression is deemed to violate TOS agreements, more expression will be deleted. When content is reported as hate speech, the likely response will be removal. Removal of reported content would forestall criticism and would be cheaper than the cost of complying with new laws").

[1016] Rebecca Tushnet, "Content Moderation in an Age of Extremes" (2019) 10 Case Western Reserve Journal of Law, Technology & the Internet 1 at 3.

remove a critical documentary from the Internet.[1017] A major study of notice-and-takedown practices found that bad faith notices were common:

> Nearly every OSP [online service provider] recounted stories of deliberate gaming of the DMCA takedown process, including to harass competitors, to resolve personal disputes, to silence critics, or to threaten the OSP or damage its relationship with its users. Although the proportion of problematic requests varied by type of OSP, every OSP told stories of takedowns that ignored fair use defenses or that targeted non-infringing material. Several echoed one respondent's view that "many copyright complaints…would obviously qualify as fair use; others are complete fabrications to remove content considered undesirable to the filer."[1018]

Moreover, wrongful takedowns disproportionately impact members of marginalized communities. Examples from Facebook alone include the following incidents:

- suspending a high-profile Black Lives Matter activist for posting a racist and abusive email he received;[1019]
- deleting a user's post in which she described an incident where a stranger hurled racist profanity at her and her children in a grocery store;[1020]
- removing a photo of two fully-clothed men kissing, stating that it violated guidelines on "nudity or graphic or sexually suggestive content" and sparking outcries that the company was homophobic (for not flagging or removing photos of heterosexual couples kissing);[1021]
- deleting a central organizing page for Egyptian protestors for violating its 'real-name' policy;[1022]

[1017] Katie Sykes, "Opinion: Aqua-gag — How the Vancouver Aquarium abuses copyright law to silence criticism", *Vancouver Sun* (27 April 2016), online: <https://vancouversun.com/opinion/aquagag-how-the-vancouver-aquarium-abuses-copyright-law-to-silence-criticism>; see also *Vancouver Aquarium Marine Science Centre v Charbonneau*, 2017 BCCA 395 (overturning the lower court decision awarding the Vancouver Aquarium an injunction against the filmmaker's documentary criticizing the aquarium's cetacean program, on the basis of copyright infringement for having filmed or photographed parts of the aquarium's interior).

[1018] Jennifer M Urban, Joe Karaganis & Brianna Schofield, "Notice and Takedown in Everyday Practice" (2017) UC Berkeley Public Law Research Paper No 2755628 at 40, online: *SSRN* <https://ssrn.com/abstract=2755628>.

[1019] Sam Levin, "Facebook temporarily blocks Black Lives Matter activist after he posts racist email", *Guardian* (12 September 2016), online: <https://www.theguardian.com/technology/2016/sep/12/facebook-blocks-shaun-king-black-lives-matter>.

[1020] Tracey Jan & Elizabeth Dwoskin, "A white man called her kids the n-word. Facebook stopped her from sharing it", *The Washington Post* (31 July 2017), online: <https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html?utm_term=.451805b729db>.

[1021] John Hudson, "The Controversy Over Facebook's Gay Kissing Ban Isn't Over", *The Atlantic* (22 April 2011), online: <https://www.theatlantic.com/technology/archive/2011/04/controversy-over-facebooks-gay-kissing-ban-isnt-over/349921/>; Kate Crawford & Tarleton Gillespie, "What is a flag for? Social media reporting tools and the vocabulary of complaint" (2016) 18(3) new media & society 410 at 411 ("So began a public controversy in which Facebook was accused of hypocrisy and homophobia, with critics noting that gay kisses were being flagged and removed while straight kisses went unremarked.").

[1022] Rebecca MacKinnon, *Consent of the Networked: The Worldwide Struggle for Internet Freedom* (New York: Basic Books, 2013) at 151-52.

- deleting an environmental protest page with over 800,000 members;[1023] and
- removing historically significant photos such as 'Napalm Girl' as well as political art.[1024]

Given all of the above, a keen sensitivity to the issue of wrongful removals has been firmly entrenched in those who work on, or advocate for, civil liberties and human rights in the context of intermediary liability, freedom of expression, and related domains of scholarship and expertise. This is for good reason.[1025]

At the same time, care must be taken not to let the mistakes and scars of past and present intermediary liability battles in copyright cast a shadow that impedes efforts to address online violence and abuse against women, girls, and other marginalized identities.[1026] The stakes between the two contexts are incomparable. The equities must be weighed differently. Zarizana Abdul Aziz elaborates:

> The stage set between internet intermediaries and violence against women victims/survivors cannot be further removed from the stage set between copyright concerns and internet intermediaries. Unlike intellectual property protection, which involves big corporations with limitless funds pursuing violators and internet intermediaries and influencing governments, victims/survivors of online violence are everyday women. The high cost of litigation and such formidable opponents as internet intermediaries with resources that rival states can combine to defeat victims/survivors at the outset. These obstacles are especially acute for women who already face greater challenges in accessing justice, such as poor women, female teenagers, younger women and sexual minorities. It also has the effect of bringing more unwanted attention and can prompt recurring instances of the violation, since courts are not always willing to shield the victims/survivors by giving them anonymity.[1027]

In sum, victims/survivors of TFGBV must not be given short shrift with an incorrectly struck balance that results in too weak or ineffective a platform liability regime to uphold their rights to equality, privacy, and freedom of expression. The effect would penalize vulnerable Internet users for the fact that the

---

[1023] *Ibid* at 155.

[1024] Brigette Supernova, "Facebook's Most Famous Banned Images" (9 September 2016), online: *Daily Beast* <https://www.thedailybeast.com/facebooks-most-famous-banned-images>.

[1025] This may also partly explain why, as Keller and Leerssen observe, the overwhelming majority of empirical research into content moderation practices focuses on examining the rates, implications, and consequences of wrongful takedowns of content, with much less focus on illuminating the same for 'wrongful leave-ups' of online abuse, hate speech, and related harmful expression on digital platforms. See Daphne Keller & Paddy Leerssen, "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation" in Nathaniel Persily & Joshua A Tucker, eds, *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge: Cambridge University Press, 2020) at 220.

[1026] Cynthia Khoo, "CUSMA: No one-size solution to platform liability" (July/August 2020), online (pdf): *Monitor* at 11 <https://www.policyalternatives.ca/sites/default/files/uploads/publications/National%20Office/2020/06/CCPA%20Monitor%20July%20Aug%20WEB.pdf>. ("The point is that defamation, copyright, technology-facilitated violence and other user-generated issues tied to digital platforms each require their own separate and contextualized legal and policy analysis of the most suitable approach to liability. Each analysis can make reference to, but should be ultimately independent of, the analysis in other areas of law. This mitigates the risk that incorrect, misguided or objectionable approaches to platform liability in one area will cascade into others, resulting in further poor law and policy.").

[1027] Zarizana Abdul Aziz, "Due Diligence and Accountability for Online Violence against Women" (31 July 2017), *APC Issue Paper* at 19, online (pdf): <https://www.apc.org/sites/default/files/DueDiligenceAndAccountabilityForOnlineVAW.pdf>.

balance was struck badly a decade or more earlier for a completely unrelated set of (well-resourced) commercial interests, when governments conceded disproportionately aggressive and punitive intermediary liability regimes to the legacy film, music, and publishing industries for the sake of commercial copyright.

6.2.2. Platform Liability Cannot Be One-Size-Fits-All

To mitigate risks of being under-inclusive, over-inclusive, or simply unsuitable, an effective platform liability regime must account for the wide range of digital platforms, which vary in size, nature, purpose, technical infrastructure, staff, content, and community. In addition to the classifications of artisanal, community, and industrial content moderation models that Robyn Caplan set out, Rebecca Tushnet reminds that platforms may also be categorized into profit or not-for-profit, with examples of the latter including Wikipedia and an NGO founded by Tushnet, the Organization for Transformative Works (OTW), which runs the well-known fanfiction website Archive of Our Own (AO3).[1028] Additionally, laws and policies aimed at digital platforms would likely not be appropriate for other kinds of intermediaries that operate at different, lower levels of the Internet stack, such as domain name registrars and Internet service providers.[1029]

By implementing legislation that has been drafted primarily in contemplation of large, dominant, commercial platforms such as Facebook and Google, but which is then imposed on all platforms regardless of relevant distinctions, the paradoxical result may be entrenchment of such dominant platforms' monopolies, as they will be the only ones equipped to comply with the legislation. As Tushnet says, "it is important not to treat YouTube as a model for the internet at large—unless all we want from the internet is YouTube".[1030] Using the EU Copyright Directive as an example, Tushnet explains that a highly controversial content filter proposal "takes a solution that has benefits for a few big copyright owners and big internet services and demands its imposition on other intermediaries—most of which don't have a big infringement problem in the first place and many of which couldn't continue to operate if they had to bear the costs of developing and constantly updating a filtering system".[1031]

The law must also take into account the existence of 'purpose-built' platforms. These are not platforms of general application such as Twitter or Instagram, which ostensibly serve general and varied purposes and, relative to certain platforms, only 'incidentally' facilitate TFGBV. These 'general use' platforms stand in contrast to other platforms that may lie closer to the publication end of the spectrum, such as 'The Dirty', due to a higher level of involvement in actively soliciting, facilitating, and featuring only a certain kind of content that constitutes TFGBV. Thus, platform liability regimes for 'platforms of general

---

[1028] Rebecca Tushnet, "Content Moderation in an Age of Extremes" (2019) 10 Case Western Reserve Journal of Law, Technology & the Internet 1 at 5 ("[S]ome entities, like the OTW [Organization for Transformative Works], don't resemble the profit-seeking model at which most regulatory and governance proposals are directed. Other online entities, such as those that participate in the domain name system, have very different functions and abilities than the websites and apps most people think of as 'the internet.' If we don't keep these variances in mind, we are unlikely to get the results we seek.").

[1029] *Ibid*.

[1030] *Ibid* at 11.

[1031] *Ibid* at 9. Tushnet continues at 9, "Ironically, because Europe is hostile to Facebook and YouTube, it has adopted a solution that ensures that Facebook and YouTube will continue to dominate, since they are the ones most likely to survive filtering and licensing requirements" (footnotes omitted).

application' would likely not be appropriate or sufficiently robust to address platforms explicitly dedicated to perpetuating, encouraging, or disseminating TFGBV.

A final consideration related to the nature of different platforms is whether some might give rise to a higher standard of responsibility and accountability than others, by virtue of their relationship with users. Jack Balkin, for example, has proposed that some Internet intermediaries ought to be deemed "information fiduciaries" in certain contexts:

> We should think of these kinds of online service providers, in short, as special-purpose information fiduciaries. The nature of their services should guide our judgments about what kinds of duties it is reasonable to impose. We should connect the kinds of duties that information fiduciaries have to the kinds of services they provide. What is unexpected or seems like a breach of trust will depend on the kind of service that entities provide and what we would reasonably consider unexpected or abusive for them to do.[1032]

Such a proposal goes beyond the duty of care proposed by the UK *Online Harms White Paper*, as it would not require platforms to meet an established reasonable standard of care, but rather, to commit to acting in their users' best interests at all relevant times. Tushnet warns that the concept of information fiduciaries and other platform regulation proposals "tend to lock in the idea that large online spaces where people engage with one another will be privatized and profit-seeking", to the detriment of both smaller commercial platforms and non-governmental, non-profit, public interest platforms "who don't want to build something advertising-driven".[1033] If a model of fiduciary duty were applied, it may also be prudent to build in substantive equality considerations ensuring salutary effects for historically marginalized and more vulnerable groups in particular, to prevent unintended consequences or exploitation of the model against them in ostensibly fulfilling a 'fiduciary duty' to more powerful and privileged users of online platforms.

A comprehensive and context-sensitive platform liability regime might address the above set of considerations and concerns in at least two ways. First, the regime could operate on a principles-based sliding scale approach, such as a standard of care that would adjust to accommodate what would be considered reasonable in light of a specific platform's nature and circumstances. This is not to say that different standards of *liability* should apply; marginalized users of smaller or less influential platforms are as deserving of equality rights and freedom of expression as are marginalized users of larger or more dominant platforms. Moreover, in no case should platforms be exempt from responding to the most egregious or harmful types of TFGBV once they are brought to their attention, if a specific legal obligation has been established for platforms to address TFGBV. A reasonable standard of care approach may make more sense where the legal obligation more closely resembles a broad duty of care that leaves some degree of discretion to the platforms in fulfilling their legal obligations.

Second, the liability framework could set out clear and precise definitions as to what criteria a platform would have to meet in order for the framework to apply, and this could include granular applications of specific sets of provisions that would apply. Examples of this approach include the tiered approaches in the *Online Harms White Paper: Full government response to the consultation* in the United Kingdom

---

[1032] Jack M. Balkin, "Information Fiduciaries and the First Amendment" (2016) 49(4) UC Davis Law Review 1183 at 1229.

[1033] Rebecca Tushnet, "Content Moderation in an Age of Extremes" (2019) 10 Case Western Reserve Journal of Law, Technology & the Internet 1 at 15.

and in the European Union's proposed *Digital Services Act*. The definitions and criteria might also apply to the *functions* a platform carries out for users, as opposed to *entity category*, recognizing that some platforms may engage in multiple functions simultaneously, or that what appear to be similar platforms in form are not so in substance, for the purposes of applying legal obligations to address TFGBV.

6.2.3. Privatized Regulation of Speech and Public Discourse

Holding digital platforms liable for user content and behaviour raises several issues associated with what may amount to formalizing privatized governance of public discourse, depending on the specific platform regulation model implemented. The extent to which privatized enforcement raises issues may depend on the extent to which a platform's content moderation decisions are part of voluntary internal processes or 'co-regulatory' frameworks, or the extent to which platform decisions are compelled through a regulatory order, explicit legal obligations, or a law imposing direct liability. In the latter cases, the enforcement would not be privatized, *per se*, though there may remain questions tied to the amount of discretion platforms are given in fulfilling such obligations or orders.

Potential issues include the risk of privatized censorship; potential 'laundering' of state policy through unaccountable or opaque agreements with digital platforms, which raises rule of law considerations; the risk of entrenching national or global 'content cartels'; and ongoing transparency and accountability issues associated with content moderation-related collaboration and standardization between major digital platforms, as well as between collaborating platforms and governments.

First, many have questioned the wisdom or desirability of putting digital platforms in a position where they are legally obligated to make consequential determinations over what has traditionally been exclusively within the purview of government and the judiciary, i.e., questions of speech and public discourse. As former UN Special Rapporteur David Kaye wrote regarding platform regulation and user-generated content, "Complex questions of fact and law should generally be adjudicated by public institutions, not private actors whose current processes may be inconsistent with due process standards and whose motives are principally economic."[1034] Without carefully designed legal frameworks that include robust oversight, transparency, and accountability mechanisms, the fear is that "privatised enforcement system[s] encouraged under the [EU Code of Conduct on countering illegal hate speech and similar agreements] would lead to private censorship", where "censorship measures are delegated to private entities" such as digital platform companies.[1035] As a result, "the ultimate arbiters of the proper limits on fundamental rights might ultimately be algorithms or other forms of artificial intelligence deployed by platforms who presumably lack judicial training, not to mention cross-border accountability."[1036]

Second, Bailey points out that privatized, market-driven forces—such as those at the foundation of corporate technology companies—cannot be relied upon to make decisions that necessarily advance

---

[1034] David Kaye, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (2018 thematic report on content regulation), 2018, A/HRC/38/35, at para 17.

[1035] Eugénie Coche, "Privatised enforcement and the right to freedom of expression in a world confronted with terrorism propaganda online" (2018) 7 Internet Policy Review 4 at 4, citing United Nations General Assembly, Human Rights Council (May 2011), *Report on the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression* (A/HCR/17/27) at para 45.

[1036] Shaheen Shariff & Karen Eltis, "Addressing Online Sexual Violence: An Opportunity for Partnerships between Law and Education" (2017) 27 Education & Law Journal 99 at 110.

human rights and the public good, particularly when it comes to matters of equality and freedom from discrimination. The following statement about Internet service providers (ISPs) applies as well, if not more so, to digital platforms such as social media companies:

> [T]he private market has an unimpressive historic record in correcting discrimination based on personal characteristics such as race, gender, and sexual identity. Unfortunately, this record is arguably consistent with rational behaviour by suppliers in seeking to meet consumer preferences. To the extent that there is demand for hateful content sufficiently widespread to sustain profitable economic activity, rational suppliers acting in their own economic self-interest will supply product to meet that demand. […] Unfortunately, economic rationality may dictate that some ISPs will refuse to restrict content in relation to the least powerful members of society, reasoning that these groups are least able to bring economic pressure to bear.[1037]

Third, Evelyn Douek has raised concerns with the rising popularity and development of what she terms "content cartels: arrangements between platforms to work together to remove content or actors from their services without adequate oversight".[1038] Such arrangements often involve a shared centralized hash database of banned content, shared content moderation tools (thus universalizing decisions on specific pieces of content across platforms), or shared centralized standards and criteria for what content to leave up or take down. Examples range from the relatively well-defined and uncontroversial, such as collaborative agreements to remove child sexual abuse material (with Cleanfeed being one such collaborative effort in Canada[1039]), to more complicated and problematic 'content cartel' initiatives, such as the Global Internet Forum to Counter Terrorism (GIFCT).

Since its start between Facebook, Microsoft, YouTube, and Twitter in December 2016, the GIFCT "has grown to include a dozen platforms and engagement with more than 120 tech companies" and "has swelled from focusing on ISIS and al-Qaeda to preventing 'terrorists and violent extremists from exploiting digital platforms.'"[1040] According to Courtney Radsch, the GIFTCT exemplifies key issues common to the centralized content moderation model that forms the basis of 'content cartels': unclear and contested definitions regarding impugned content (e.g., what constitutes terrorist speech or hate speech);[1041] lack of independent review or audit of removed content to ensure legitimate expression is

---

[1037] Jane Bailey, "Private Regulation and Public Policy: Toward Effective Restriction of Internet Hate Propaganda" (2003) 49 McGill Law Journal 59 at 95-96 (footnotes omitted).

[1038] Evelyn Douek, "The Rise of Content Cartels" (2020) at 6, online: *Knight First Amendment Institute at Columbia University* <https://s3.amazonaws.com/kfai-documents/documents/704838d2ec/3.23.2021_Douek_MW--To-Print-.pdf>.

[1039] "Cleanfeed Canada", online: *Cybertip.ca*, <https://www.cybertip.ca/app/en/projects-cleanfeed> ("How does Cleanfeed Canada work? Cybertip.ca receives complaints from Canadians regarding websites potentially hosting child pornographic images. Child pornography websites meeting the necessary criteria for Cleanfeed are amassed on the Cleanfeed Canada distribution list. Cybertip.ca provides that list in a secure manner to participating ISPs (participation is voluntary). The ISPs' filters automatically prevent access to addresses on the list. There is essentially no "human" intervention on the part of participating ISPs. ISPs do not have input into creating the list nor knowledge of what is contained on it.").

[1040] Courtney Radsch, "GIFCT: Possibly the Most Important Acronym You've Never Heard Of" (30 September 2020), online: *Just Security* <https://www.justsecurity.org/72603/gifct-possibly-the-most-important-acronym-youve-never-heard-of/>.

[1041] Evelyn Douek, "The Rise of Content Cartels" (2020) at 27-28, online: *Knight First Amendment Institute at Columbia University* <https://s3.amazonaws.com/kfai-documents/documents/704838d2ec/3.23.2021_Douek_MW--To-Print-.pdf>.

not removed; opaque decision-making and lack of transparency, accountability, or oversight,[1042] including by civil society groups and historically marginalized communities disproportionately impacted by such policies; and the ability to internationally implement a given country's domestic content moderation standards or speech laws, if the rule is incorporated into GIFCT.[1043]

Another concern with 'content cartels' is that wrongful takedowns or biased content moderation standards may rapidly proliferate or simultaneously occur across all major platforms at once, if they result from a centralized content moderation database.[1044] This is even more disquieting when combined with the lack of recourse or appeal mechanisms for content that has been wrongfully identified as violating the standards of a given 'cartel', entered into the shared centralized database, and then taken down across all platforms which are members of the agreement.[1045] Both Douek and Radsch make recommendations for transparency, accountability, oversight, and due process measures that may mitigate the above-mentioned concerns with centralized content moderation between private platforms and governments, while retaining the benefits of such approaches.[1046]

Fourth, outsourcing far-reaching decisions on what constitutes legitimate or permissible speech across digital platforms to platform companies themselves, raises troubling questions around rule of law and the role of the state in regulating speech and public discourse. On one level, such arrangements make it possible for governments to 'launder' speech-related policy decisions by embedding them into content moderation policies centralized between platform companies in a 'content cartel', without having to account for such decisions either to the electorate or to the courts.[1047] The issues that Michael

---

[1042] See e.g., Heidi Tworek, "Social Media Councils" (28 October 2019), online: *Centre for International Governance Innovation* <https://www.cigionline.org/articles/social-media-councils> ("For now, the GIFCT remains mostly a mystery to those outside the companies or specific civil society organizations and governments who cooperate with the forum. To give a few examples, we do not even know where the database is housed. The GIFCT has no provision for third-party researcher access. We do not know if additions to the database by one company are ever disputed by another or if there are even mechanisms to resolve such a dispute.").

[1043] Courtney Radsch, "GIFCT: Possibly the Most Important Acronym You've Never Heard Of" (30 September 2020), online: *Just Security* <https://www.justsecurity.org/72603/gifct-possibly-the-most-important-acronym-youve-never-heard-of/>.

[1044] Evelyn Douek, "The Rise of Content Cartels" (2020) at 14, online: *Knight First Amendment Institute at Columbia University* at 14 <https://s3.amazonaws.com/kfai-documents/documents/704838d2ec/3.23.2021_Douek_MW--To-Print-.pdf>. ("This creates the danger that biases in one platform's data will find their way into how other platforms moderate. In the case of Perspective, researchers have raised concerns that the tool has a disparate impact on already marginalized communities.").

[1045] *Ibid* at 24 and 36.

[1046] *Ibid* at 34 ("Ultimately, the answer should depend on an empirical inquiry into factors such as the prevalence of that category of content; the accuracy of the relevant technology; the cost and practicality of small platforms developing similar tools; the relevant risk of harm; and, especially, the contestability of the category definition and whether it implicates speech, such as political speech, that is ordinarily highly protected. More research is needed for a true assessment of social welfare costs and benefits. This requires greater openness from companies (with a nudge from regulators, if necessary). In the meantime, in these cases, ad hoc, opaque cartelization should not be encouraged."); see also, generally, *ibid* at 33-36, and Courtney Radsch, "GIFCT: Possibly the Most Important Acronym You've Never Heard Of" (30 September 2020), online: *Just Security* <https://www.justsecurity.org/72603/gifct-possibly-the-most-important-acronym-youve-never-heard-of/>.

[1047] Daphne Keller & Paddy Leerssen, "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation" in Nathaniel Persily & Joshua A Tucker, eds, *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge: Cambridge University Press, 2020) 220 at 225.

Karanicolas raises with government 'jawboning'[1048] apply with equal force where governments are part of or have influence over platform content cartels:

> [I]n the context of restrictions on speech, this tactic can be problematic, insofar as it removes any opportunity to question whether the new rules are consistent with bedrock freedom of expression principles, since traditional avenues of judicial appeal do not apply in the same way to private sector enforcement decisions. Similarly, if the new restrictions are unpopular, the public is denied a meaningful opportunity to express their displeasure at the ballot box. There is rarely a clear and visible line which connects private sector policy shifts to the government's complaints. While this dynamic is characteristic of all jawboning campaigns, since they present a fuzzier target for opposition than the passage of new legislation, the politically sensitive nature of restrictions on speech, and the centrality of freedom of expression to the political process, mean that it is particularly concerning in this context.[1049]

Bailey points out an additional major problem in the other direction: not only may content cartels and the privatization of speech regulation allow for opaque state interference without accountability, but it can also allow for the government's abdication of its human rights obligations in this arena. Specifically, it is possible that "reliance on private market solutions in the context of hate propaganda would divest public authorities of their responsibility to safeguard Canadian public policy, with no degree of certainty that private market responses would serve national and international collective commitments to equality and diversity."[1050]

At the same time, Bailey also suggests that conceding some degree of reliance on platform intermediaries is a necessary compromise,[1051] given the limitations of public regulation and the particular position and power that digital platforms hold today as the 'new governors' of speech. A legal framework that imposes accountability to a certain extent—including robust transparency measures—would appropriately reflect the extent to which digital platforms have "have become unwitting arbiters of the global public interest".[1052] Dunn, Lalonde, and Bailey state:

> [S]ince social media companies are the providers of some of the dominant spaces for public discourse and social interaction, their increasing impact on people's everyday lives arguably renders them quasigovernmental. With that shift in power should come increased responsibility for social media companies to create and maintain,

---

[1048] 'Jawboning' refers to a process in which "platforms are pressured through threats of regulation to shift their broader approach to moderating content in order to bring it into line with categories that governments might seek to target": Michael Karanicolas, "Squaring the Circle Between Freedom of Expression and Platform Law" (2019-2020) 20 Journal of Technology Law & Policy 177 at 186.

[1049] Michael Karanicolas, "Subverting Democracy to Save Democracy: Canada's Extra-Constitutional Approaches to Battling 'Fake News'" (December 2019) 17 Canadian Journal of Law & Technology 200 at 216.

[1050] Jane Bailey, "Private Regulation and Public Policy: Toward Effective Restriction of Internet Hate Propaganda" (2003) 49 McGill Law Journal 59 at 94.

[1051] *Ibid* at 80, 97.

[1052] Shaheen Shariff & Karen Eltis, "Addressing Online Sexual Violence: An Opportunity for Partnerships between Law and Education" (2017) 27 Education & Law Journal 99 at 109.

> accountably and transparently, safe and respectful online spaces that facilitate girls' equal participation, rather than their victimization.[1053]

To reiterate the passage above, responsibility should follow power, regardless of formal legal categories. If platform companies are powerful enough that they ought to be bound by obligations to uphold the right to freedom of expression, then they are, by the same token, powerful enough that they ought to be bound by obligations to uphold the right to equality and freedom from discrimination.

6.3. Additional Challenges in Addressing Platformed TFGBV

In addition to potential constitutional vulnerabilities of a platform liability regime, and difficulties related to the unique characteristics of digital platforms and their role in society, there remain several additional challenges related to mitigating, preventing, or eliminating platformed TFGBV in practice. This section will briefly touch on each of these challenges in turn.

The first challenge is what might be considered a seeming pattern of oversight when it comes to addressing TFGBV in its own right, as a form of platform-related individualized and systemic harm that specifically impacts women and girls as a group. For example, the Facebook Civil Rights Audit does not include the words 'sexism' or 'misogyny' anywhere in the report, despite stating it intended to encompass civil rights issues across the board and examine harms to individuals and communities based on all legally protected characteristics.[1054] The *Online Harms White Paper: Full Government Response to the consultation* in the United Kingdom does not mention 'misogyny', 'sexism', 'gender', or the right to equality, and mentions 'women' only once[1055]—despite gender equality groups and women's rights organizations participating in the consultation, and the original White Paper noting these issues. However, it does include copious references and reassurances throughout regarding the rights to privacy and freedom of expression.

In addition, central debates shaping intermediary and platform liability laws historically have largely been dominated by copyright lobbyists, national security and intelligence communities, and law enforcement on the one side who advocated for greater control over Internet intermediaries and platform companies; and digital rights or 'Internet freedom' advocates and civil liberties defenders on the other side, who primarily focused on protecting the right to freedom of expression and the right to privacy. As Jonathan Zittrain describes, the first of what he deems the "three eras" of Internet governance was characterized by "a classic libertarian ethos of the preservation of rapidly-growing individual affordances in speech [...] against encroachment by government censorship or corporate pushback motivated by the disruption of established business models."[1056] Without having elevated the right to equality alongside the rights to privacy and freedom of expression, this left little room for individuals and groups who were victimized by anonymous abusers, for example, yet were equally

---

[1053] Suzanne Dunn, Julie S. Lalonde, and Jane Bailey, "Terms of Silence Weaknesses in Corporate and Law Enforcement Responses to Cyberviolence against Girls" (2017) 10(2) Girlhood Studies 80 at 86-87.

[1054] "Facebook's Civil Rights Audit - Final Report" (8 July 2020) at 5, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf>.

[1055] *Ibid* at 35.

[1056] Jonathan Zittrain, "Three Eras of Digital Governance" (23 September 2019) at 1-2, online: *SSRN* <https://ssrn.com/abstract=3458435>.

opposed to strengthening state surveillance or aggressive copyright enforcement measures chilling historically marginalized or critical expression.

The second challenge is the severe, widely- and repeatedly-documented failures and ignorance of justice system actors when it comes to both TFGBV and gender-based violence generally.[1057] This includes the judiciary, police officers, and other members of law enforcement. Among these actors, victim-blaming, victim responsibilization, and dismissal and trivialization of sexual, gender-based, and intimate partner violence, abuse, and harassment remain prevalent. Adding layers of technological illiteracy on top of pre-existing patriarchal dynamics do not improve the situation.[1058]

The third challenge is the lack of internal platform data regarding their content moderation policies and practices, and metrics and statistics providing a more detailed picture of the prevalence and exact nature of the various kinds of TFGBV that occurs across digital platforms. There is also a lack of empirical research regarding 'wrongful *leave-ups*' of violent and abusive content that has been flagged and reported to platforms, in contrast to the abundance of research on wrongful takedowns in the context of copyright notices and online censorship.[1059]

The fourth challenge is that common business-oriented themes are continually raised in opposition to imposing platform liability or stronger platform accountability for online violence and abuse. However, these arguments do not address or outweigh the importance of addressing TFGBV or what is at stake in permitting TFGBV to continue flourishing in the absence of effective regulation. Specifically, such arguments include:

- consistent focus on 'innovation' as a necessary priority or automatic good,[1060] regardless of its costs or consequences, and without recognizing that the Internet and digital platforms are "no longer a fragile new means of communication that could easily be smothered in the cradle by

---

[1057] See e.g., Jessica West, "Cyber-Violence Against Women" (May 2014) at 24, 26-27, online (pdf): *Battered Women's Support Services* <http://www.bwss.org/wp-content/uploads/2014/05/CyberVAWReportJessicaWest.pdf>; Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Taking Action to End Violence against Young Women and Girls in Canada: Report of the Standing Committee on the Status of Women* (March 2017) at 56-57, 88-89 (Chair: Marilyn Gladu); Canada, Parliament, House of Commons, Standing Committee on the Status of Women*, Evidence*, 42nd Parl, 1st Sess, No 21 (21 September 2016) at 1615 (Valerie Steeves); Shaheen Shariff & Karen Eltis, "Addressing Online Sexual Violence: An Opportunity for Partnerships between Law and Education" (2017) 27 Education and Law Journal 99 at 111-114; and Cynthia Khoo, Kate Robertson & Ronald Deibert, "Installing Fear: A Canadian Legal and Policy Analysis of Using, Developing, and Selling Smartphone Spyware and Stalkerware Applications" (June 2019) at 165-67, online (pdf): *Citizen Lab* <https://citizenlab.ca/docs/stalkerware-legal.pdf>.

[1058] See e.g., *R v Elliott*, 2016 ONCJ 35 (discussion regarding the function of the '@' sign on Twitter).

[1059] Daphne Keller, "Empirical Evidence of Over-Removal by Internet Companies Under Intermediary Liability Laws: An Updated List" (8 February 2021), online: *The Center for Internet and Society at Stanford Law School* <https://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws> ("Many of these studies were conducted by academics or advocates with a particular interest in protecting user free expression and ensuring that legal content remains available online. One day I hope we will see more data from the other side — advocates for rightsholders, defamation plaintiffs, or other groups harmed by online content that violates their legal rights. That could help build a more complete picture of the over-removal issue as well as any related under-removal problem — intermediaries failing to remove content when notified, even though applicable law requires removal.").

[1060] Danielle Keats Citron & Neil M Richards, "Four Principles for Digital Expression (You Won't Believe #3!)" (2018) 95 Washington University Law Review 1353 at 1375 ("More recent evangelists have emphasized Silicon Valley's 'disruptive innovation,' its capacity to continually replace old business models with new ones. Implicit is the belief that disruption is either intrinsically a good thing or that "innovation" tends to produce new good things rather than new bad ones.").

overzealous enforcement of laws and regulations applicable to brick-and-mortar businesses";[1061]

- overemphasis on 'scale' and inability to achieve 'perfect' content moderation, though many platforms have yet to consistently capture even the lowest hanging fruit (e.g., unambiguously and substantively harmful expression and behaviour, such as NCDII, or posts celebrating or encouraging sexual violence against women) and there remains ample room for even imperfect regulation given the cost of inaction;[1062] and
- the idea that digital platforms are "damned if they do, and damned if they don't",[1063] as if it matters not at all who is doing the damning and why, in the context of implementing measures to decrease misogynistic hate speech and other forms of platformed TFGBV.

The fifth challenge is the perennial and unavoidable fact that at the end of the day, while TFGBV is very much tied to and shaped by its technological context, its root causes are not technological at all. Violence against women and girls is a societal, cultural, and political problem, and those who hold misogynistic and sexist views, or who are complacent in the face of them, will continue to find ways to perpetrate and perpetuate gender-based violence and abuse regardless of what technologies are available. Thus, alongside legal reform efforts targeting digital platform accountability and liability, work must continue across multiple spheres of society, within and outside of the law, to "address misogyny, racism, homophobia, and other intersecting oppressions that have been used as tools to keep women down, to silence them, and to keep them out of the public sphere".[1064] As Jane Bailey said to the House of Commons Standing Committee on the Status of Women:

> Meaningfully addressing the disproportionate targeting of girls and young women for sexualized [TFGBV] […] requires nothing short of social transformation. That's what it's about. As a friend of mine said, "Yes, you're talking about ending the patriarchy, so good luck." That's okay. That's what we're talking about: ending the patriarchy.[1065]

Legislative and other reform efforts must thus simultaneously work towards addressing the above challenges in forming effective responses to platformed TFGBV.

[1061] Danielle Keats Citron & Benjamin Wittes, "The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity" (2017) University of Maryland Francis King Carey School of Law Legal Studies Research Paper at 20, citing *Fair Housing Council v Roommates.com*, 521 F.3d 1157, 1164 (9th Cir. 2008).

[1062] See e.g., "Masnick's Impossibility Theorem: Content Moderation At Scale Is Impossible To Do Well" (20 November 2019), online: *Techdirt* <https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml>; and Tarleton Gillespie, "'The Scale Is Just Unfathomable'" (1 April 2018) online: *Logic* <https://logicmag.io/scale/the-scale-is-just-unfathomable/>.

[1063] Lauren Feiner, "Facebook, Twitter CEOs will have to answer to Senate Republicans after Biden New York Post story controversy" (15 October 2020), online: *CNBC* <https://www.cnbc.com/2020/10/15/facebook-twitter-ceos-set-to-answer-to-senate-after-reducing-ny-post-story-distribution.html>.

[1064] Canada, Parliament, House of Commons, Standing Committee on the Status of Women, *Evidence*, 42nd Parl, 1st Sess, No 20 (16 June 2016) at 1650 (Jane Bailey).

[1065] *Ibid*.

7. Recommendations

This report has developed six overarching priorities to guide law reform, which were synthesized from the law and literature reviewed throughout this report. The recommendations set out in this Part are animated by those priorities to address platform liability for technology-facilitated gender-based violence, abuse, and harassment (TFGBV).

While there is a role for all levels of government, other decision-makers, and platform companies themselves to play in addressing TFGBV, these priorities and recommendations are aimed primarily at the Canadian federal government. This reflects the report's focus on legislative reform at the federal level.

Before presenting the priorities and recommendations, it bears repeating that TFGBV is not wholly a new phenomenon. It is a technological evolution of traditional intersecting systems of oppression, including sexism, racism, colonialism, ableism, transphobia, and homophobia. All of these pre-date digital platforms and the Internet itself. Systemic oppressions and violence against women, girls, and gender-diverse people would not disappear even if all digital platforms were shut down tomorrow. This injustice will continue unabated so long as these root causes are not substantively addressed at systemic and institutional levels.

Addressing the root causes of systemic oppression requires social, cultural, and political change. The core harms of TFGBV will not be eradicated unless all levels of government and civil society also act decisively to end *other forms of* gender-based violence, abuse, and harassment. As with TFGBV, the law plays an important but ultimately limited role in the broader matrix of meaningful solutions, and must be contextualized as such to avoid complacency when it comes to pursuing non-legal solutions to TFGBV and gender-based violence, abuse, and harassment.

The recommendations set out here concern platform liability for TFGBV *alone*. They are proposed with the specific context and concerns of TFGBV in mind, including the fundamental right to equality and other human rights of, in addition to the overall wellbeing of, those impacted by TFGBV. The recommendations should thus not be seen as arguments for platform liability in other, unrelated areas in which the federal government has also expressed an interest in platform regulation, such as media industry funding, copyright, or Canadian cultural policy.

Finally, the recommendations also only apply to digital platforms as defined in Section 3.1.1 ("What Are Digital Platforms?") of this report, i.e., application-layer and content-layer intermediaries such as social media platforms and video-sharing websites. They are not necessarily intended to apply to Internet intermediaries that operate at a more infrastructural level, such as Internet service providers, which have been excluded from the scope of this report.

7.1. Priorities for Law Reform in Platform Liability for TFGBV

The **first priority** is to recognize that there is indeed a need for law reform to address platformed TFGBV. Regulating or placing certain kinds of liability on digital platforms is appropriate and necessary, given their role in facilitating TFGBV. However, creating a platform liability framework must be done thoughtfully and with a clear focus on TFGBV specifically, while building in substantive equality and

intersectionality principles. Any platform liability framework enacted must be human rights-centred, principled, and proportionate to the specific objectives at the heart of the regime. Where TFGBV is concerned, a more robust legal response would be justified given its devastating and systemic harms to historically marginalized groups.

The **second priority** is to recognize that proportionate limitations on freedom of expression are constitutionally justified, both to uphold the right to equality and freedom from discrimination, as well as to give full effect to core values underlying the right to free expression. This approach is consistent with Supreme Court of Canada jurisprudence. For women, girls, and individuals from intersecting historically marginalized and systemically oppressed communities, TFGBV is a pervasive and devastating component of sustained inequality. This priority also requires recognizing, as Canadian courts have, that such groups are as vulnerable to private abuses of power as they are to state abuses of power. For that reason, there is room for the state to legislate limitations on certain freedoms to address systemic discrimination, violence, and abuse by private, non-state actors.

The **third priority** is to guarantee that legal reforms that address TFGBV build in victim/survivor-centered, trauma-informed, and intersectional feminist perspectives. This must include substantive consultation with those impacted by TFGBV—notably, members of historically marginalized and vulnerable groups. This will be essential to guarding against the adoption of legal approaches that are inconsistent with the goals, aspirations, and lived experiences of members of these communities.

The **fourth priority** is to ensure expedient, practical, and accessible remedies for those targeted by TFGBV, particularly where it causes clear and immediate harm, as with the non-consensual distribution of intimate images (NCDII). For some instances of TFGBV, requiring a court order to support a platform takedown request is unrealistic and unworkable; the damage would already be done by the time the individual targeted was able to obtain such an order. The focus should be on providing effective remedial relief and support to those subjected to TFGBV. Moreover, to be accessible and effective, remedies must provide options for individuals who do not wish to engage, come into contact with, or have their information passed on to law enforcement or the criminal justice system.

The **fifth priority** is to provide due process mechanisms to users who wish to contest platforms' content moderation decisions (whether a decision to leave up or take down content). These must be made available by platform companies themselves, with an appeal process before an independent TFGBV-specialized regulator. Such processes acknowledge the complexity of platform regulation and content moderation, as well as the beneficial impacts of the Internet, including for historically marginalized and systemically oppressed groups, while safeguarding users' freedom of expression where it may be inadvertently unduly infringed.

The **sixth priority** is to require transparency from platform companies regarding their content moderation policies and decisions, as well as the outcomes of such policies and decisions concerning TFGBV. Without more and better data providing insights into how such policies and decisions are made and implemented, governments, regulators, and the public will be stymied in influencing how the general public is governed and affected by digital platforms.

7.2. Recommendations

This report proposes 14 recommendations for the Canadian federal government to implement. They are organized into the following categories:

- recommendations emphasizing the importance of centering human rights, substantive equality, and intersectionality in legal reforms, particularly as it impacts victims/survivors of TFGBV;
- specific legislative reforms that the federal government should enact, establishing new legal regimes and a new TFGBV-specialized agency;
- certain legal obligations that the government should place on digital platform companies to enhance regulators' and the public's ability to hold them accountable for TFGBV; and
- specific areas of TFGBV-related research, education, and training that the government should support through funding.

7.2.1. Centering Human Rights, Substantive Equality, and Intersectionality

1. **Apply a principled human rights-based approach to platform regulation and platform liability, including giving full effect to the right to equality and freedom from discrimination.** Such an approach would be rooted in Canadian human rights and constitutional law, in addition to Canada's obligations under international human rights instruments. This involves understanding that the *Charter of Rights and Freedoms* creates a non-hierarchical matrix of rights, where giving full effect to the right to equality and freedom from discrimination constitutionally justifies proportionate limitations on other rights, such as freedom of expression.
2. **Ensure that legislation addressing TFGBV integrates substantive equality considerations and guards against exploitation by members of dominant social groups to silence expression by members of historically marginalized groups.** Groups and individuals in power have often used seemingly salutary law to silence members of historically marginalized communities from speaking out, such as in the case of defamation law and victims/survivors of sexual assault. Similarly, abusive users have often gamed and manipulated platforms' content moderation features to silence or shut down the accounts of users from historically marginalized groups. Platform companies themselves have regularly implemented content moderation policies and decisions that failed to take historical context and substantive equality into account, which resulted in under-removal of abusive content and over-removal of content that spoke out against such abuse. Any proposed legislation must avoid a harmful formal-equality approach (treating all users and circumstances the same, regardless of context or social location), and additionally, must account for and build in safeguards against the high likelihood of abuse of process by those with power and privilege, so that the new system cannot be used as another tool to perpetuate further acts of TFGBV.

3. **When pursuing legislative or other means of addressing TFGBV, consult substantively with and take into account the perspectives and lived experience of victims, survivors, and those broadly impacted by TFGBV.** This must include intersectional considerations such as the intersecting impacts of racial discrimination or transphobia. For example, mechanisms such as 'real name' policies and identity verification have been shown to operate against historically marginalized and vulnerable individuals. This includes those who have escaped or are hiding from situations of intimate partner violence, sex workers who rely on pseudonymity for safety, or activists and human rights defenders living under authoritarian regimes. Those who have been targeted or otherwise impacted by TFGBV have valuable and hard won insights to share about potential consequences of regulation that may be overlooked by government actors and other potential stakeholders who have not been negatively impacted by TFGBV.

7.2.2. Legislative Reforms

4. **Establish a centralized expert regulator for TFGBV specifically, with a dual mandate: a) to provide legal remedy and support to individuals impacted by TFGBV on digital platforms, including regulatory and enforcement powers; and b) to develop research and provide training and education on TFGBV to the public, relevant stakeholders, and professionals.** Recommendations about specific features of the TFGBV-specialized agency include the following:

    a. **Mandate:** Expressly define the regulator's mandate to be focused on TFGBV. This must clearly articulate that TFGBV is rooted in, and includes, all forms of intersecting systemic oppressions, such as misogyny, racism, colonialism, homophobia, transphobia, and ableism. Women, girls, and gender-diverse individuals may be simultaneously targeted based on other characteristics protected under equality and non-discrimination law—for example, ethnicity, disability, and/or socioeconomic status. The mandate may go beyond strictly gender-based harms, to include technology-facilitated violence, abuse, and harassment that is not based on gender but based on being Black, Indigenous, or otherwise racialized, for instance. The regulator and associated legal framework must at all times recognize that individuals who belong to two or more historically marginalized communities are targeted in ways particular to the *intersection* and distinct from the experiences of individuals who belong solely to any one historically marginalized group. The unifying principle that should apply to constrain the boundaries of the mandate is that the sole focus is on addressing technology-facilitated violence, abuse, and harassment which targets members of historically marginalized groups, with the core objective of advancing substantive equality and upholding these groups' human rights. The legislation must prohibit any 'mission creep' that would expand the regulator's mandate or functions to other issue areas that the government may be interested in addressing through platform regulation.

    b. **Definition of TFGBV:** Clearly and specifically define the types of behaviours that constitute TFGBV, based on the intersectional understanding of the term described in (a), and which are therefore within the purview of the regulator to address. Ensure that the behaviours included provide an 'intelligible standard' by which to identify content that is and is not captured by the law.

c. **Remedial, Adjudicative, and Enforcement Function:** Set up the regulator as a remedial and adjudicative complaints body and create a resolution process available to individuals being subjected to TFGBV (as defined in the statute). The resolution process must prioritize speed, practicality, and accessibility for those individuals. The regulator should provide both legal remedies and solutions outside the legal system where appropriate. Under no circumstances should police or law enforcement be informed or involved without the express and informed consent of the victim/survivor. The regulator should be granted powers to provide declaratory relief and issue orders to platform companies that fall under the legislation, enforced through administrative penalties. Individuals being victimized by TFGBV must have access to the agency's resolution process and support systems even if they have not yet used the platforms' internal processes to address abuse. Individuals who have already undergone a platform's internal process and wish to contest the platform's decision may appeal the decision to the regulator, which will then begin an adjudicative process that results in a binding decision on the platform. Both the individual who submitted a complaint *and* the person whose content is the subject of the complaint must be able to appeal the platform's decision—whether the decision was to take down *or* leave up the content.

d. **Training, Education, and Research Function:** Establish the regulator with a robust training, education, and research wing, parallel to and with as much if not more funding and resources than the remedial, adjudicative, and enforcement wing. This function of the regulator would involve providing a range of training and education resources to members of the public; to community-based organizations and frontline support workers addressing TFGBV, gender-based violence, and intimate partner and dating violence; and to law enforcement, legal professionals, schools, and other relevant institutions. The agency would also be responsible for consulting historically marginalized groups impacted by TFGBV and frontline organizations serving them, as well as liaising with platform companies, to develop best practices for industry, support regulatory compliance, and ensure that the regulator and its processes are meeting the needs of victims/survivors of TFGBV. In addition, part of the agency's mandate would be to conduct or commission and publish further research regarding TFGBV, of the kind described in Recommendation 14 below.

e. **Expertise and Capacity:** Staff the agency with personnel who are well-versed in TFGBV or related issues. Individuals in an executive, management, or frontline support role must have prior expertise and/or experience in addressing TFGBV, intimate partner violence, racial injustice, and/or other forms of systemic oppression and how they can be furthered through technology, and appropriately supporting those impacted by TFGBV. The regulator must be sufficiently resourced to build further internal expertise and capacity regarding all aspects of TFGBV, including the technosociological aspects of digital platforms, the way platform features are exploited and gamed by users to perpetrate TFGBV, and the lived experiences of those subjected to and impacted by TFGBV—both online and offline (including understanding the increasing meaninglessness of such a dichotomy).

f. **Consultation:** Consult extensively—in setting up this agency, its mandates, and its processes—with historically marginalized groups, those who have been or are impacted

by TFGBV, technology and human rights experts, gender equality advocates, community-based groups, and lawyers and researchers who specialize in TFGBV.

g. **Oversight and Statutory Review:** Put in place oversight and accountability mechanisms for the regulatory body itself, and include statutorily mandated periodic reviews of the governing legislation, to ensure that it is meeting its victim/survivor-centered mandate.

h. **Sequestered from Law Enforcement:** Prohibit the regulatory body from being used as a conduit for automatically transmitting information to law enforcement agencies. Any transmission of information must be done with the express informed consent of the victim/survivor, and only under certain circumstances clearly delineated in the legislation—for example, where the regulator has reasonable grounds to believe that the conduct at issue may constitute a criminal offence. The regulator must also have a legal duty to evaluate the situation based on principles of substantive equality and intersectionality. Any automatic 'off-ramp' to law enforcement or data sharing will guarantee that the body becomes inaccessible and unavailable to many who may need it the most, due to heightened risks of discrimination and state abuse related to engagement with the criminal justice system, for members of historically marginalized groups.

5. **Enact one or more versions of the current 'enabler' provision in subsections 27(2.3) and 27(2.4) of the *Copyright Act*, adapted to specifically address different forms of TFGBV, including 'purpose-built' platforms.** Recommendations for specific aspects of the provision(s) include the following:

    a. Draft the provisions to capture 'purpose-built' platforms that exist predominantly to host, solicit, generate, and/or facilitate TFGBV by users.

    b. Clearly and specifically define what constitutes TFGBV for the purposes of being captured by the legislation, taking into account intersectional considerations.

    c. Consider beginning with enabler provisions that capture only the most clearly defined and easily identifiable forms of TFGBV with pressing substantive harms, such as NCDII and expression that constitutes hate speech under current civil and criminal laws.

    d. The provisions might attribute liability in one of two ways, where a platform is found to have met the test for being an 'enabler' of TFGBV as defined in the legislation:

        i. Direct liability for the underlying offence (e.g., applying existing criminal liability for NCDII, or applying criminal or statutory human rights liability for hate speech, as if the platform were the speaker); or

        ii. A new 'enabler liability' specific to the provision. This may be preferred only in situations where the underlying user activity does not already constitute a civil action or criminal offence, but collectively amounts to systemic harm requiring a legal response. This would justify targeting the platform for institutional liability even if the individual users would not be liable individually.

6. **Enact a law that allows for victims/survivors of TFGBV to obtain immediate removal of certain clearly defined kinds of content from a platform *without* a court order, such as NCDII.** Not requiring victims/survivors to obtain a court order would take into account the practical reality of TFGBV, as well as its devastating and human rights-violating impacts. Requiring a court order would be completely unworkable in providing timely and meaningful remedies to victims/survivors in practice (especially when combined with ongoing access-to-justice concerns).

7. **Ensure that legislation to address TFGBV focuses solely on TFGBV (including intersectional considerations)—do not dilute, compromise, or jeopardize the constitutionality of such legislation by 'bundling' TFGBV with other issues that the government may wish to also address through platform regulation.** Such other issues may require alternative approaches and attract different analyses of constitutionality under the *Charter of Rights and Freedoms*. Examples might include disinformation, terrorism outside of white supremacist extremism, or non-TFGBV-related defamation. Most of these issues, at best, do not primarily engage the right to equality, and at worst, introduce a high risk of state action that threatens equality. Their respective contexts involve legally significant departures from the context of TFGBV, impacting the constitutional analysis of a given limitation on platform-facilitated user expression. This includes differences in the equities and the nature of the relationship between the state and individuals impacted by the law.

7.2.3. Legal Obligations for Platform Companies

8. **Require platform companies to provide to users *and non-users* clearly visible, easily accessible, plain-language complaint and abuse reporting mechanisms to expediently address and remedy instances of TFGBV.** These complaint procedures and content moderation processes should also include due process mechanisms, such as appealing a decision to remove or leave up content, subject to victim- and survivor-centred considerations. Making platform resolution processes available to *non*-users is critical because individuals targeted by TFGBV may not themselves be users of platforms where the abuse is occurring. Moreover, such individuals should not be deemed subject to a platform's terms of use if they access a platform's resolution services to respond to TFGBV. Platforms' data deletion and retention policies must centre the needs of victims/survivors of TFGBV, including, for example, offering the option of total deletion of NCDII across the platform *and* any parent, sibling, or subsidiary platform companies where the NCDII is also found (to reduce the 'whack-a-mole' burden on victims/survivors), or disabling public access to content but retaining it on the back end for evidentiary purposes where the individual wishes, in contemplation of potential legal action. Where an individual has opted for total deletion, the platform should provide them with a formal incident report that documents details of the complaint for evidentiary purposes for the person's records and in case they decide to proceed with legal action.

9. **For 'purpose-built', 'enabling', or otherwise TFGBV-dedicated platforms, and where a clearly delineated threshold of harm is met, provide that an order to remove specific content on one platform will automatically apply to any of that platform's parent, subsidiary, or sibling platform companies where the same content also appears.** The purpose of this power is to reduce the burden on victims/survivors of having to undergo multiple resolution processes to obtain a remedy on a case-by-case basis, where the same

substantively harmful content is involved and where time may be of the essence, such as in the case of NCDII. It also aligns accountability with those who commercially benefit from such content regardless of which of their platforms is involved. That this remedy is only available by way of an order through the regulator, requires meeting a threshold of harm, and is limited to a set of platforms that by definition excludes 'platforms of general application', is to safeguard against the possibility that the remedy is misused to efficiently shut down and silence expression by members of historically marginalized groups.

10. **Require platform companies to undergo independent audits (which could be conducted by the new TFGBV agency) and publish comprehensive annual transparency reports.** These reports should provide qualitative information and granular data in both human- and machine-readable formats. The data should be broken down by demographics (particularly gender and race) to the extent possible, regarding the platform's internal content moderation policies and practices, and regarding the prevalence of and efforts to address TFGBV, as well as the results of those efforts. In drafting this requirement, the government should consult current literature and experts regarding transparency reports in the fields of platform regulation, content moderation, and algorithmic accountability.

11. **When determining legal obligations for digital platforms, account for the fact that platforms vary dramatically in size, nature, purpose, business model (including non-profit), extent of intermediary role, and user base.** This does not mean that different platforms should be held to different standards of liability—marginalized users of smaller or less influential platforms are as deserving of equality and freedom of expression as are marginalized users of larger or more dominant platforms. Rather, it means that it may be appropriate to adopt an element of flexibility and context-sensitivity in establishing *what* platforms are required to do to fulfill any established regulatory obligations. Consider as well regulating by *function* as opposed to *entity category*, as some digital platforms may otherwise fall into multiple categories if they offer a variety of intermediary services to users.

7.2.4. Research, Education, and Training

12. **Fund, make widely available, and mandate (where appropriate) education resources and training programs in TFGBV, which include information on how to support those who are subjected to TFGBV.** These resources must be developed in collaboration with those who have subject matter expertise and/or lived experience with TFGBV. These resources should be provided to members of the public; to community-based organizations and frontline support workers addressing TFGBV, gender-based violence, and intimate partner and dating violence; and to law enforcement, legal professionals, schools, and other relevant institutions. People who access the resources should learn about, for instance, technological literacy; the broader social context in which TFGBV is grounded; preventing TFGBV; challenging or refraining from victim-blaming; the lived experiences of those impacted by TFGBV; and providing a trauma-informed and victim/survivor-centred response in cases of TFGBV. This recommendation applies with particular force to police agencies and law enforcement, and such education and training should be mandatory for these entities at both the federal and provincial /territorial levels. The new TFGBV-specialized agency could be responsible for specialized education and training aimed at actors within the legal system, in addition to broader public education and

training, though funding must also go to community-based organizations and others who are qualified to create and provide training and education resources to the public or other groups.

13. **Fund frontline support workers and community-based organizations working to end, and supporting victims/survivors of, gender-based violence, abuse, and harassment, specifically to enhance their internal expertise, resources, and capacity to support those impacted by TFGBV (which often accompanies gender-based violence and abuse).** In addition, fund community-based organizations to systematically track incidents of TFGBV over time in order to evaluate the impacts of relevant laws and other response systems to TFGBV. The new TFGBV-specialized regulator could administer such funding, in partnership with community-based organizations.

14. **Fund further empirical, interdisciplinary, and law and policy research by TFGBV scholars, other TFGBV experts, and community-based organizations on TFGBV and the impacts of emerging technologies on those subjected to TFGBV.** In the context of platform liability for TFGBV specifically, such research might begin with a focus on the prevalence and causes of 'wrongful leave-ups' of reported content constituting TFGBV, in contrast to research that focuses on 'wrongful takedowns' of reported content that was not abusive. Research in this area could also involve collecting further empirical data on the impacts of different platform liability models on the experiences of historically marginalized groups subjected to TFGBV. The new TFGBV-specialized regulator could be tasked with setting up and administering a research grants program similar to the Contributions Program at the Office of the Privacy Commissioner of Canada, as well as commissioning research from subject-matter experts to inform further law reform, policy-making, and future government responses to TFGBV.