Working papers in Applied Linguistics and Linguistics at York 2, 1-12

Content moderation as language policy: Connecting commercial content moderation policies, regulations, and language policy

Mandy Lau[1]
York University, Toronto, Canada

Abstract: Commercial content moderation removes harassment, abuse, hate, or any material deemed harmful or offensive from user-generated content platforms. A platform's content policy and related government regulations are forms of explicit language policy. This kind of policy dictates the classifications of harmful language and aims to change users' language practices by force. However, the de facto language policy is the actual practice of language moderation by algorithms and humans. Algorithms and human moderators enforce which words (and thereby, content) can be shared, revealing the normative values of hateful, offensive, or free speech and shaping how users adapt and create new language practices. This paper will introduce the process and challenges of commercial content moderation, as well as Canada's proposed Bill C-36 with its complementary regulatory framework, and briefly discuss the implications for language practices.

Keywords: language policy; commercial content moderation; social media regulation; platform moderation.

1 Introduction

The harmful proliferation of online hate has seeped into public consciousness, particularly as many societies face increased white supremacy hate crimes and right-wing extremist violence both online and offline. In response, governments are hurriedly crafting regulations to control the potential for harm on social media platforms. For example, Germany has increased the purview of its 2017 Network Enforcement Act (also known as NetzDG). The Act, which regulates the removal of hate speech off platforms, now requires platforms to report illegal content to law enforcement (Lomas, 2020). Simultaneously, the European Commission is also progressing with legislation regulating hate speech within the Digital Services Act, proposed in December 2020 (European Commission, 2021). Canada is following suit, prioritizing a new bill to combat harmful online content in the first 100 days of the newly re-elected Federal Government (Liberal Party of Canada, 2021). The proposed Bill C-36 makes amendments to the Criminal Code, the Youth Criminal Justice Act, and the Canadian Human Rights Act. It is complemented by a new regulatory framework that further directs social media platforms in how they address harmful content (Government of Canada, 2021b). Social media platform companies built on user-generated content (UGC) already grapple with implementing effective content moderation policies and practices. These platforms, such as Facebook, YouTube, or TikTok, employ algorithms and humans to identify, downrank, and remove content deemed harmful according to the company's content

[1] *Corresponding author: laumandy@yorku.ca*

 ISSN 2564-2855

policy. Content moderation practices include using language markers, such as removing posts consisting of or tagged with words that are classified as harmful.

Language policy as a discipline is concerned with the language practices of language users, language users' beliefs and ideologies about the value of languages, and language management—how some members of the language community force or encourage changes to how or which language is used by the greater community (Spolsky, 2012). From this perspective, a state's regulatory framework and a platform's content policy can be viewed as forms of explicit language policy. This kind of policy dictates the classifications of harmful language and aims to change users' language practices by force. However, the de facto language policy (Shohamy, 2006) is the actual practice of language moderation by algorithms and humans. Algorithmic systems and human moderators enforce which words (and thereby, content) can be shared, revealing the normative values of what is considered hateful, offensive, or free speech and shaping how users adapt and create new language practices. This paper will introduce the process and challenges of commercial content moderation, as well as the proposed Bill C-36 with its complementary framework, and briefly discuss the implications for language practices.

2 What is commercial content moderation?

Commercial content moderation removes harassment, abuse, hate, or any material deemed harmful or offensive from user-generated content (UGC) platforms. Also known as platform moderation, commercial content moderation operates at an across-the-platform scale, in contrast to community-based moderation led by volunteer administrators of early bulletin board systems or online forums (Gorwa et al., 2020; Roberts, 2019). Moderation practices combine both human and computational approaches, just as content can be created and uploaded by humans and computational systems (i.e., bots or deep fakes).

A platform's content moderation process typically follows this pathway:

1. *Establishment of policy*

   A content policy is established, which may or may not be shared with the public. A public-facing content policy (sometimes referred to as community guidelines) may be shared with human users during content upload to verify adherence to policy.
2. *Identification of harmful material*

   Identification of potentially harmful material occurs at the time of content upload or after. It includes both automated and human methods.
   a. Automated/Top-down methods (Gorwa et al., 2020):
      i. Matching methods: Also known as fingerprinting or hash matching, the identified content is matched against a collection of known harmful content. A positive match is flagged.
      ii. Predictive methods: The identified content is classified (i.e., harassment, toxic, terrorist content). This approach may use filtering methods (i.e., screen for prohibited words or phrases, where the lists of words are curated) or machine learning approaches (i.e., language classifiers trained on a corpus of texts annotated by human reviewers). Content classified as problematic is flagged.
   b. Human flagging/Bottom-up: Users flag harmful content and report it to the UGC platform.

3. *Content review*

   Human content moderators review flagged content. This step is skipped in some instances involving automated methods. For example, Facebook uses automated methods to "proactively…take action on the content automatically", particularly for content categorized as "most real-world harm," such as terrorist or self-harm content (Zuckerberg, 2018). Facebook reported that automated removal of posts only occurs when the system's "confidence level is high enough that its 'decision' indicates it will be more accurate than [their] human reviewers" (Bickert & Fishman, 2018).

4. *Governance outcome*

   A decision is made by the human content moderator or the automated system. Content may be permitted, suppressed (by being downranked, rendered less visible, or muted; also known as shadow banning, stealth banning, or ghost banning), geoblocked, or deleted. Deleted content that violates laws may be reported to law enforcement or intelligence agencies based on legal frameworks. Implementation of the decision may occur immediately at upload or within a specified timeline, such as within 24 hours of identification. In many cases, moderation decisions extend to entire accounts, which may be flagged and removed entirely, or "whitelisted," where high-profile users are given more leniency and exemptions from moderation policy. For example, Facebook's cross-check program claims to provide high-profile users (i.e., celebrities, athletes, politicians) additional quality control to avoid false-positive identification of harmful content (Horwitz, 2021; Meta, 2021).

5. *Notification*

   Some UGC platforms notify the content creator or user who flagged the content of the review decision. For whitelisted VIP users such as those in Facebook's cross-check program, they may be provided a "self-remediation window" of 24 hours to delete the content themselves (Horwitz, 2021). For platforms without clear notification processes, content creators typically discover that their content cannot be found or has less than expected exposure, or users discover that the content they flagged remains on the platform.

6. *Appeal*

   Some UGC platforms, such as Facebook, have a secondary review process to appeal to leave up or remove content. Facebook also established a semi-independent oversight board to review further appeals on a case-by-case basis (Meta, 2021; Oversight Board, 2021).

2.1 The challenges of commercial content moderation

Commercial content moderation is a crucial commodity of platforms; it enables useability for users, attracts advertisers, and appeases governance stakeholders while limiting corporate liability (Cobbe, 2020; Gillespie, 2018; Gorwa et al., 2020). In sum, it protects both the brand and its shareholder interests. However, it would be an understatement to say that moderating content based on language and labelled images is highly complex. Language-using is a social practice in which meaning is negotiated within specific socio-historical contexts and relationships (Cameron, 1995, p. 2; Street, 2005). The act of languaging is not straightforward telementation; ideas are often expressed through affective or metaphorical modes, via tone, gestures, jokes, memes, sarcasm, irony, or idioms (Cameron, 1995; Cobbe, 2020; Harris, 1987). Content moderation through language is thus a technical problem and a sociopolitical phenomenon. To interpret meaning through decoding a string of texts requires a shared understanding of a stable set of codes and a fixed sociopolitical context. But language does not exist in a neutral vacuum. Constructing

meaning requires negotiation and contestation. Whether human or machine, moderation is contextually situated. Effective moderation requires knowing a word's meaning in the contexts of the content, the content creator, the intended audience, the norms of the platform, and the laws from where the content is shared (Roberts, 2019).

Given this complexity, the task of content moderation may be best left up to humans who are superior to machines in terms of cognitive flexibility and creativity. However, moderating harmful content subjects humans to psychological trauma, including increased risks for mental illness, suicide, and acceptance of disinformation and malinformation[2] (Newton, 2019; Roberts, 2019). There are also scalability limitations since an overwhelming amount of content is generated online. For example, within every minute in 2021, an average of 240,000 photos are shared on Facebook, 695,000 stories on Instagram, 575,000 tweets on Twitter, and 500 hours of content uploaded on YouTube (Statista, 2021; Jenik, 2021). Even a company as well-resourced as Facebook can only screen about five percent of all daily postings (Langvardt, 2018). Already, this works out to three million posts (text updates, videos, images) daily (Koetsier, 2020).

Oppressive labour conditions also exacerbate the problem. Much of a tech company's content moderation is outsourced either to companies with access to a cheaper labour force, many in the Global South (Newton, 2019; Roberts, 2019), or to platform workers engaging in hidden labour, or what Gray and Suri (2019) termed *ghost work*. These workers are evaluated against strict time and accuracy metrics and are only given seconds to click and evaluate a post. They often do not have much access to feedback or resources aside from the company policies (Gray & Suri, 2019; Langvardt, 2019; Roberts, 2019). These conditions create a perfect storm fraught with errors.[3]

Automation via algorithms may reduce the psychological trauma afflicted on human moderators and potentially scale-up moderation practices. However, scaling up means that all content is subjected to moderation, even before it is uploaded, thereby increasing the reach of techno-surveillance. Gorwa et al. (2020) also note three political issues concerning algorithmic content moderation. They include:

1. *Decisional transparency:* Machine learning conveniently blackboxes how automated systems flag content and make decisions. It makes it more difficult for third-party auditors and researchers to vet the system and enables tech companies to experiment with new tweaks without oversight.
2. *Justice:* Algorithmic systems depend on classification data that have demonstrated representational harms against marginalized peoples, including women (Gonen & Goldberg, 2019), racialized people (Manzini et al., 2019), transgender people (Dias Oliva, Antonialli & Gomes, 2021), people of the Muslim faith (Abid et al., 2021), and the poor (Eubanks, 2018), disproportionately impacting their language practices and the content

[2] Malinformation is the publication of private, genuine material (original or altered) intended to produce harm. Disinformation is non-genuine, false material fabricated and disseminated to produce harm (i.e., conspiracy theories). These differ from misinformation, in which false content is shared unintentionally, by mistake (Wardle, 2018).

[3] Facebook admitted that 10% of moderation decisions are errors (Zuckerberg, 2018). This is estimated to be around 300,000 decision mistakes daily (Koetsier, 2020). As of September 28, 2021, 18 decisions were made by the Facebook Oversight Board: 11 moderation decisions were overturned, 6 were upheld, and 1 was a non-decision (Oversight Board, 2021). Error rates are also not evenly distributed; the rate of false-positive identification of terrorist content is at 77% for Arabic languages (Simonite, 2021).

they post. Content posted by marginalized peoples is also more prone to false-positive identification as harmful material. Further, marginalized users are more often subjected to online policing and trolling (whereby trolls falsely flag their content), effectively moderating their voices out (Bender et al., 2021; Benjamin, 2019; Noble, 2018).

3. *Depoliticization:* Should moderation systems successfully operate as part of the background infrastructure of UGC platforms, any "questionable" content would be made invisible. Political contestation would also be erased from view. Trust in automation's capacity for scientific objectivity and its opaque design will reduce political negotiation.

In essence, this type of statistical classification system branded as content moderation AI will have the effect of scaling up opacity and discrimination while limiting democratic oversight and contestation.

3 The Canadian context: The proposed Bill C-36 and regulatory framework

Bill C-36: An Act to amend the Criminal Code and the Canadian Human Rights Act and to make related amendments to another Act (hate propaganda, hate crimes and hate speech) (2021) amends the Criminal Code, the Youth Criminal Justice Act, and the Canadian Human Rights Act, adding new definitions and a peace bond on hatred, hate speech, hate propaganda offences, and hate crimes. The Bill was proposed before the 2021 Federal election was called, passing its first reading in June 2021 before the Government dissolved, and is one of the legislations tabled for the first 100 days of the newly re-elected Liberal Government (Curry, 2021). The first iteration of the proposed complementary legislative and regulatory framework was shared via a public consultation from July to September 2021 (Government of Canada, 2021a). Through a discussion guide and a technical paper, the Government of Canada (2021b) outlines how "online communication service providers (OCSP)" would address harmful content, identified by five types:

1. Child sexual exploitation content
2. "Terrorist" content
3. Content that incites violence
4. Hate speech
5. Non-consensual distribution of intimate images (NCDII)

The Act holds OCSPs accountable for identifying and making inaccessible harmful content to persons in Canada. It sets out rules for OCSPs, such as the requirement to publish content moderation policies, address flagged content within a 24-hour time frame (including notification to the content flagger and content creator, content removal or reporting to law enforcement), establish procedures for recourse, and file data reports to regulation bodies regularly (Government of Canada, 2021c). Those who fail to comply face a monetary penalty of up to 3% of the OCSP's gross global revenue or ten million dollars and must commit to a compliance agreement (Government of Canada, 2021c). Further failure to comply may result in indictable offences, increased fines (up to 5% of gross global revenues or twenty-five million dollars), and having the service blocked in Canada (Government of Canada, 2021c). New regulatory bodies and boards will be established to oversee and enforce the proposed framework, including a Digital Safety Commissioner, a Digital Safety Commission, a Digital Recourse Council of Canada, and an Advisory board (Government of Canada, 2021c).

Researchers at Citizen Lab have criticized the proposed framework as "vague, ambiguous, and in some cases contradictory" within a consultation process that is "grossly inadequate" (Khoo

et al., 2021, p.2-3). They note that the framework's scope is "overly broad and incoherent," as each of the five categories of harmful content implicates different laws, rights, and risks and requires unique mitigations and interventions (Khoo et al., 2021, p.6). The broad commitment to reducing online harms without any clarity in its implementation can be seen in this example of establishing timelines to address harmful content after it has been flagged:

> The Act should provide that... 'expeditiously' is to be defined as twenty-four (24) hours from the content being flagged, or such other period of time as may be prescribed by the Governor in Council through regulations.
>
> ...the Governor in Council may prescribe through regulations different timelines for different types or subtypes of harmful content...the new timeframes could be either extended or shortened from the timeframe provided...
>
> (Government of Canada, 2021c, Module 1(B), section 11).

Much in the document follows this format: Basically, a general principle is stated as a rule (i.e., addressing flagged content within 24 hours), then contradicts itself by stating that the Governor in Council may later prescribe new rules that may be different, specific to the type of harm (i.e., new timelines could be extended or shortened). This pattern of vagueness and lack of operational detail continues regarding the Government of Canada's statements on the OCSP's obligation to publish "clear content-moderation guidelines" (2021c, section 13), maintain records (2021c, section 14-17), and report harmful content to law enforcement (2021c, section 18-33). As Citizen Lab noted, it is difficult to engage in thoughtful public analysis, debate, and policymaking with so much ambiguity and so many contradictions (Khoo et al., 2021).

Despite lacking meaningful detail, the general idea of the framework is to hold social media companies more accountable in addressing some of the issues that surfaced publicly, such as moderation procedures and transparency in recourse policies and impacts. However, it also seeks to expand state and commercial powers for intelligence and policing, including new powers for the Canadian Security Intelligence Services (CSIS) (Government of Canada, 2021b; Government of Canada, 2021c; Khoo et al., 2021). New reporting protocols will undoubtedly enhance the state's abilities to surveil and police and increase state dependence on privatized OCSPs for surveillance and policing purposes. At best, this set-up can resemble what Crawford (2021) describes as an "uncomfortable bargain," whereby states are uninformed in making agreements with OCSPs that are out of their control, and OCSPs are inapt in taking on state functions while opening themselves up for possible constitutional liability (p. 208). At worst, it risks regulatory capture, whereby the new regulatory bodies no longer make decisions based on public interest but are conspiring to act based on the interests of the OCSPs The Act is intended to regulate.

In part, the proposed legislation is directing social media platforms to continue doing what they are already doing through their content policies. In recent years, social media companies such as Facebook have come under immense public scrutiny and criticism and are fervently protecting their brand by clamping down on internal content policies anyway (see, for example, The Wall Street Journal's Facebook Files, 2021). AI and machine learning are commonly flaunted as the solution (Canales, 2021; Jeong, 2018; Pelley, 2021; Seetharaman et al., 2021). This focus on a technical fix is also reflected in Canada's proposed regulatory framework (see Government of Canada, 2021c, Module 1(B) Section 10: "...OCSP must take all reasonable measures, which can include the use of automated systems, to identify harmful content…"). But looking to the same platform company that hosted the problem to provide an opaque technical fix only further

entrenches their power. Broadly touting AI as a mythical fix exaggerates the capabilities of what are really statistical classification systems rife with errors and inequities and cements the companies' significance in providing a necessary social service. This tech-centric narrative continues to erase the human workers who train or review moderation systems from view (Roberts, 2019; Gray & Suri, 2019). It serves to reduce accountability for over-censorship (the costly penalties incentivize removal of content, which critics say impedes freedom of expression and speech rights; Kaye, 2019). It also reduces accountability for under-censorship (for example, a goal of Facebook's cross-check program is to minimize "PR fires" by tilting moderation towards influential, popular, or newsworthy users (Horwitz, 2021)). Further, the outsourcing model provides convenient plausible deniability for the powerful. Regulators limit their responsibility to reduce social harm by outsourcing both moderation solutions and responsibility to platforms, which further outsource to smaller tech companies or platform workers, many in the Global South.

4 Battle of the words

The framing of content moderation as a technical problem to be solved by AI tech-washes the underlying structural conditions that lead to the dissemination of abusive and harmful content. The focus remains on individual users and their individual posts, and the site of contestation remains at the word level—a battle of the words.

Definitions are established to construct a fixed notion of terms. For example, in the proposed Bill C-36, *hatred* is defined as "the emotion that involves detestation or vilification and that is stronger than dislike or disdain" (Bill C-36, 2021, p. 1). *Online hate speech* is speech that is "likely to foment detestation or vilification of an individual or group of individuals based on a prohibited ground of discrimination" communicated over the Internet or "other means of telecommunication" (Bill C-36, 2021, p. 9). However, it is a fine line between detestation or vilification (hatred) and dislike or disdain (protected speech). Whether or not an online speech act is considered hateful or protected is prone to various degrees of interpretation and debate. This level of subjectivity increases ambivalence in identifying online hate speech for both victims and law enforcement. For example, a report on hate speech crimes in the EU found that hate speech is underreported, and even when it is reported, police officers are ill-equipped to recognize it (Bayer & Bárd, 2020). A parallel may be drawn with historical cases of hate propaganda laws in Canada, where subjective interpretation of the terms along with a high legal burden of proof results in low rates of criminal charges and convictions (Yang, 2017).

In her book *Verbal Hygiene* (1995), Cameron notes the indeterminacy of language, "the impossibility of ever definitively pinning down what a particular utterance means" (p. 24). The indeterminate quality of language destabilizes the legal definitions of hate speech. However, Cameron also notes that the indeterminacy of language reveals its flexibility, "which enables us to use it in novel situations to mean an infinite number of things" (1995, p. 24). The flexible quality of language enables individual users to avoid content moderation or conceal their activities.

Narrow technical moderation solutions are easily gamed and navigated around simply by changing the spelling of words, substituting letters with symbols or words, or the use of emojis. For example, malicious actors can continue to spread harmful material and dodge positive identification by appropriating innocuous language. This can be seen in how child abuse and pornography continue to spread on Instagram via hashtags containing cheese and pizza emojis (Andrews, 2020). On the opposite end of the spectrum, political activists adopt new practices to avoid being falsely blocked. For example, Facebook's ban on the word *Taliban* during the Taliban's takeover of Afghanistan resulted in the suspensions of many anti-Taliban activists based

in Afghanistan and Pakistan. To circumvent this, some activists will use the word *Taliban* with a different spelling order (Nazari, 2021). However creative, this still reduces the reach of activists and journalists and has the effect of self-censorship. Many have reportedly deactivated their accounts or stopped writing about the Taliban (Nazari, 2021). More subtle and partially hidden ways of circumventing moderation are seen in online pro-eating disorder communities. On Instagram, users stopped using hashtags (i.e., *#promia*, *#size00*, *#thyghgap*, and *#thinspiration*) to connect within the community because it was a primary method of moderation (Gerrard, 2019). Instead, they use profile biographies, which are not typically moderated by algorithms, to post signals, such as target weights, ambiguous sounding diet plans, or participation in fasting games (Gerrard, 2019).

In formal learning, educators and students are also contending with word-level classification algorithms. In a project on digital surveillance and privacy, university students in Hong Kong and the UK constructed learning on how their linguistic, political, and social subjectivities are embedded in how they interact with algorithmic texts (Jones, 2021). They would interact with various algorithms to get to know their features and affordances to exploit or hack the algorithms to serve their needs and wants. What is clear from their narratives is that the algorithms are still normative. For example, students noted how they began to change their music-listening behaviour for the Spotify algorithm or pay closer attention to how Turnitin (a plagiarism detection tool) evaluates their writing rather than focus on creating meaning (Jones, 2021). Similarly, CEOs learn to adjust their speech and tone to manipulate algorithmic trading (Cao et al., 2020; Wigglesworth, 2020). Effectively, algorithms teach people to associate different values with different words. Likewise, content moderation algorithms and practices will continue to shape how people communicate by assigning values to words.

5 Conclusion

Commercial content moderation is vital in addressing abuse and violence facilitated and intensified through UGC platforms. This paper introduces commercial content moderation as a global human labour network and an algorithmic statistical classification system. It lists key issues, such as the everchanging and indeterminate quality of language, oppressive human labour and market conditions, and the exaggeration and depoliticization of error-prone, opaque, and unjust AI content moderation systems. It also introduces Canada's proposed Bill C-36 and the accompanying regulatory framework, which was loosely constructed to appeal to voters, widen the reach for state surveillance, and remain congruent to what social media platform companies are already doing. In its current form, content moderation policy and practices are language policies that shape users' language behaviours. Content moderation at the word level has implications for status planning (i.e., Which languages are most resourced for moderation?), corpus planning (i.e., Which words are considered hateful or merely distasteful? Which words are politically charged and at risk for flagging? Which new words or language practices emerge in response?), and acquisition planning (i.e., What are emerging algorithmic literacies for content creators, consumers, and moderators? What languages are needed to fulfil the labour pool of moderators?). But perhaps we ought to imagine alternative architecture. The attempt to improve content moderation at the word or meaning level may be futile if it is the only tool in the toolkit to fight online dissemination and amplification of hatred, violence, and abuse. By tech-washing the problem, cloaked in a narrative of content moderation AI, we reduce the phenomenon to individuals doing bad things, one act at a time, and fail to account for how social structures uphold hatred, violence, and abuse.

Acknowledgements

I would like to thank Dr. Eve Haque for her mentorship in shaping the ideas for this paper. My research activities are also supported in part by funding from the Social Sciences and Humanities Research Council.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, *3*(6), 461–463. https://doi.org/10.1038/s42256-021-00359-2

Andrews, L. (2020, August 31). Paedophiles are using cheese and pizza emojis to communicate secretly on Instagram. *Daily Mail Online*. https://www.dailymail.co.uk/news/article-8681535/Paedophiles-using-cheese-pizza-emojis-communicate-secretly-Instagram.html

Bayer, J., & Bárd, P. (2020). *Hate speech and hate crime in the EU and the evaluation of online content regulation approaches* (PE655.135 - July 2020). European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf

Bender, E., Gebru, T., McMillan-Major, A., Shmitchell, S., & Anonymous. (2020). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, *1*(1), 271–278. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Bickert, M., & Fishman, B. (2018). *Hard questions: What are we doing to stay ahead of terrorists?* Meta. https://about.fb.com/news/2018/11/staying-ahead-of-terrorists/

Bill C-36: An Act to amend the Criminal Code and the Canadian Human Rights Act and to make related amendments to another Act (hate propaganda, hate crimes and hate speech). (2021). 1st Reading June 23, 202, 43rd Parliament, 2nd session. Retrieved from the Parliament of Canada website: https://www.parl.ca/DocumentViewer/en/43-2/bill/C-36/first-reading

Cameron, D. (1995). *Verbal hygiene*. Routledge.

Canales, K. (2021, March 25). Mark Zuckerberg said content moderation requires "nuances" that consider the intent behind a post, but also highlighted Facebook's reliance on AI to do that job. *Business Insider*. https://www.businessinsider.com/zuckerberg-nuances-content-moderation-ai-misinformation-hearing-2021-3

Cao, S., Jiang, W., Yang, B., Zhang, A. L., & Robinson, J. M. (2020). *How to talk when a machine is listening: Corporate disclosure in the age of AI*. NBER Working Paper Series. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3683802

Cobbe, J. (2020). Algorithmic censorship by social platforms: Power and resistance. *Philosophy and Technology*. https://doi.org/10.1007/s13347-020-00429-0

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Curry, B. (2021, October 1). Liberals' Parliamentary agenda lists three internet regulation bills as early priorities. *The Globe and Mail*. https://www.theglobeandmail.com/politics/article-liberals-parliamentary-agenda-lists-three-internet-regulation-bills-as/

Dias Oliva, T., Antonialli, D. M., & Gomes, A. (2021). Fighting hate speech, silencing drag queens? Artificial Intelligence in content moderation and risks to LGBTQ voices online. *Sexuality and Culture*, *25*(2), 700–732. https://doi.org/10.1007/s12119-020-09790-w

Douek, E. (2021, June 2). More content moderation is not always better. *Wired*. https://www.wired.com/story/more-content-moderation-not-always-better/

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Picador.

European Commission. (2021, October 21). *The digital services act package*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media and Society, 20*(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Ghaffary, S. (2021, August 15). The algorithms that detect hate speech online are biased against black people. *Vox*. https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 609–614. https://doi.org/10.18653/V1/N19-1061

Google Canada. (2021, November 5). *Our shared responsibility: YouTube's response to the Government's proposal to address harmful content online*. Official Google Canada Blog. https://canada.googleblog.com/2021/11/our-shared-responsibility-youtubes.html

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Government of Canada. (2021a). *Consultation closed: The Government's proposed approach to address harmful content online*. https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content.html

Government of Canada. (2021b). *Discussion guide*. https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/discussion-guide.html

Government of Canada. (2021c). *Technical paper*. https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/technical-paper.html

Gray, M.L. & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass.* Houghton Mifflin Harcourt.

Harris, R. (1987). *The language machine.* Duckworth.

Horwitz, J. (2021). Facebook says its rules apply to all. Company documents reveal a secret elite that's exempt. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353?mod=hp_lead_pos7

Jenik, C. (2021). *Technology: What happens every minute on the Internet?* World Economic Forum. https://www.weforum.org/agenda/2021/08/one-minute-internet-web-social-media-technology-online/

Jeong, S. (2018, April 13). AI is an excuse for Facebook to keep messing up. *The Verge*. https://www.theverge.com/2018/4/13/17235042/facebook-mark-zuckerberg-ai-artificial-intelligence-excuse-congress-hearings

Jones, R. H. (2021). The text is reading you: Teaching language in the age of the algorithm. *Linguistics and Education, 62*. https://doi.org/10.1016/j.linged.2019.100750

Kaye, D. (2019). *Speech police: The global struggle to govern the Internet*. Columbia Global Reports.

Khoo, C., Gill, L., & Parsons, C. (2021, September 28). *Comments on the Federal Government's proposed approach to address harmful content online*. The Citizen Lab. https://citizenlab.ca/2021/09/comments-on-the-federal-governments-proposed-approach-to-address-harmful-content-online/

Koetsier, J. (2020, June 9). Report: Facebook makes 300,000 content moderation mistakes every day. *Forbes*. https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=77b8b0da54d0

Langvardt, K. (2018). Regulating online content moderation. *Georgetown Law Journal, 106*, 1353–1388. https://perma.cc/Z48T-H3K3.

Liberal Party of Canada. (2021). *Protecting Canadians from online harms*. https://liberal.ca/our-platform/protecting-canadians-from-online-harms/

Lomas, N. (2020, June 19). Germany tightens online hate speech rules to make platforms send reports straight to the feds. *TechCrunch*. https://techcrunch.com/2020/06/19/germany-tightens-online-hate-speech-rules-to-make-platforms-send-reports-straight-to-the-feds/

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 615–621. https://doi.org/10.18653/V1/N19-1062

Meta. (2021, November 3). *Reviewing high-impact content accurately via our cross-check system*. Meta Transparency Center. https://transparency.fb.com/enforcement/detecting-violations/reviewing-high-visibility-content-accurately/

Nazari, J. (2021). Don't say 'Taliban': Facebook suppresses Afghan activists and artists leading the resistance. *The Toronto Star*. https://www.thestar.com/news/world/2021/10/03/dont-say-taliban-facebook-suppresses-afghan-activists-and-artists-leading-the-resistance.html

Newton, C. (2019, February 25). The secret lives of Facebook moderators in America. *The Verge*. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Noble, S. (2018). *Algorithms of oppression*. New York University Press. https://doi.org/10.4324/9780429399718-51

Oversight Board. (2021, November). *Board decisions*. Oversight Board. https://oversightboard.com/?page=decision

Patel, F., & Hecht-Felella, L. (2021, February 22). *Facebook's content moderation rules are a mess*. Brennan Center for Justice. https://www.brennancenter.org/our-work/analysis-opinion/facebooks-content-moderation-rules-are-mess

Pelley, S. (2021, October 4). *Whistleblower: Facebook is misleading the public on progress against hate speech, violence, misinformation*. 60 Minutes. https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-misinformation-public-60-minutes-2021-10-03/

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media.* Yale University Press.

Seetharaman, D., Horwitz, J., & Scheck, J. (2021, October 17). Facebook says AI will clean up the platform. Its own engineers have doubts. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184

Shenkman, C., Thakur, D., & Llansó, E. (2021). *Do you see what I see? Capabilities and limits of automated multimedia content analysis*. Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf

Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. Routledge.

Simonite, T. (2021, October 25). Facebook is everywhere; its moderation is nowhere close. *Wired*. https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/

Spolsky, B. (2012). *The Cambridge handbook of language policy*. Cambridge University Press.

Statista. (2021). *Media usage in an internet minute as of August 2021*. Statista. https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/

Stober, E. (2021, November 5). Google warns Canada's plan to fight online hate is 'vulnerable to abuse'. *Global News*. https://globalnews.ca/news/8353819/google-canada-online-hate-plan/

Street, B. (2005). Understanding and defining literacy. In *Paper commissioned for the Education for All Global Monitoring Report 2006*. https://doi.org/10.1201/9780849378508.ch2

The Wall Street Journal. (2021). *The Facebook files: The Wall Street Journal investigation*. The Wall Street Journal. https://www.wsj.com/articles/the-facebook-files-11631713039

Wardle, C. (2018). Information disorder: The essential glossary. *First Draft Footnotes*. https://medium.com/1st-draft/information-disorder-part-1-the-essential-glossary-19953c544fe3

Wigglesworth, R. (2020, December 5). *Robo-surveillance shifts tone of CEO earnings calls*. Financial Times. https://www.ft.com/content/ca086139-8a0f-4d36-a39d-409339227832

Woolf, M. (2021, November 21). Wrangling over language may slow online harm bill, anti-hate groups say. *CBC News*. https://www.cbc.ca/news/politics/anti-hate-online-legislation-1.6257338

Yang, J. (2017, February 27). Why hate crimes are hard to prosecute. *The Toronto Star*. https://www.thestar.com/news/gta/2017/02/27/why-hate-crimes-are-hard-to-prosecute.html

Zakrzewski, C., De Vynck, G., Masih, N., & Mahtani, S. (2021, October 24). How Facebook neglected the rest of the world, fueling hate speech and violence in India. *The Washington Post*. https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/

Zuckerberg, M. (2018). *A blueprint for content governance and enforcement*. Facebook. https://www.facebook.com/notes/751449002072082/