Centre for International Governance Innovation

Digital Policy Hub

Digital Policy Hub – Working Paper

The Digital Ecology of Hate: Technology, Policy and Online Fields

Sophie Xiaoyi Liu

Fall 2024 cohort

About the Hub

The Digital Policy Hub at CIGI is a collaborative space for emerging scholars and innovative thinkers from the social, natural and applied sciences. It provides opportunities for undergraduate and graduate students and post-doctoral and visiting fellows to share and develop research on the rapid evolution and governance of transformative technologies. The Hub is founded on transdisciplinary approaches that seek to increase understanding of the socio-economic and technological impacts of digitalization and improve the quality and relevance of related research. Core research areas include data, economy and society; artificial intelligence; outer space; digitalization, security and democracy; and the environment and natural resources.

The Digital Policy Hub working papers are the product of research related to the Hub's identified themes prepared by participants during their fellowship.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

Partners

Thank you to Mitacs for its partnership and support of Digital Policy Hub fellows through the Accelerate program. We would also like to acknowledge the many universities, governments and private sector partners for their involvement allowing CIGI to offer this holistic research environment.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Key Points

- The rise of digital platforms has transformed communication but also enabled the spread of hate speech and divisive content. Algorithms optimized for engagement have created feedback loops that reinforce polarizing behaviours and narratives, highlighting the tension between user engagement and harm reduction.
- Pierre Bourdieu's concepts of *habitus* and *field* provide insight into the interplay between platform structures and user behaviour. Platforms, as digital fields, condition user dispositions through algorithms and content prioritization, fostering echo chambers and reinforcing harmful patterns.
- Existing legal frameworks, often outdated in the context of digital technologies, struggle to address the rapid proliferation of online hate speech. Recent efforts, such as Canada's Bill C-63, aim to introduce regulatory mechanisms to reshape platform practices and promote accountability.
- While artificial intelligence (AI) is a critical tool for moderating content at scale, it struggles with nuanced context, low-resource languages and biases. Effective moderation requires hybrid models that integrate AI efficiency with human oversight to ensure fairness and accuracy.
- This working paper suggests promoting algorithmic transparency, strengthening global legal frameworks, investing in inclusive AI innovations, enhancing accountability through robust oversight and enabling users with education and tools to challenge hate speech and misinformation effectively.

Introduction

The rise of digital platforms has revolutionized how individuals communicate, engage and organize, offering unprecedented opportunities for connection. However, these platforms have simultaneously facilitated the proliferation of hate speech, misinformation and extremism, leading to significant real-world consequences (Boulianne and Lee 2022; Ganesh and Bright 2020; Zhang and Davis 2022). Online harassment targeting marginalized communities and the organization of hate-fuelled events exemplify how the internet has become a critical battleground in the fight against hate (Citron 2014; Inara Rodis 2024; Matamoros-Fernández 2017).

Hate speech — defined broadly as abusive or threatening communication directed at individuals or groups based on characteristics such as race, religion, ethnicity or gender (Bartlett et al. 2014; Felmlee, Inara Rodis and Zhang 2020; Howard 2019; Paz, Montero-Díaz and Moreno-Delgado 2020) — carries profound societal and legal implications. It fosters divisions, incites violence and undermines democratic values (Gelber and McNamara 2015; Piazza 2020). While governments, social media companies and civil society actors increasingly recognize the urgency of addressing hate speech, significant challenges remain. These challenges include difficulties in defining hate speech across cultural contexts, regulating global platforms and balancing the protection of free expression with the imperative to prevent harm (Banks 2010; Brown 2018; Silva et al. 2021).

Pierre Bourdieu's (1984) concepts of *habitus* and *field* provide a compelling sociological framework for understanding the dynamics of online hate speech. For Bourdieu, *fields*

are structured spaces where norms, rules and power dynamics shape social action. Within these fields, *habitus* refers to the deeply ingrained dispositions and practices that individuals develop through their interactions with social structures (ibid.). Online platforms function as digital fields, shaping user behaviour by establishing norms and expectations through algorithms, content moderation policies and structural designs (Ignatow and Robinson 2017; Julien 2015; Levina and Arriaga 2014). These algorithms, as mediators of platform dynamics, determine what content is prioritized and visible, conditioning user behaviour to align with engagement-driven incentives (Anderson 2013; Burrell and Fourcade 2021; Tufekci 2018).

Users' dispositions, or habitus, are conditioned by their interactions within these digital fields (Bourdieu 1984). Exposure to algorithmically amplified divisive content reinforces behaviours such as participation in echo chambers and engagement with polarizing narratives (Cinelli et al. 2021). These echo chambers can create a feedback loop where users' actions further entrench the structural priorities of the platform, perpetuating harmful behaviours (Bakshy, Messing and Adamic 2015; Flaxman, Goel and Rao 2016). However, users are not entirely passive within these fields. Through collective action, such as organizing campaigns for transparency or engaging in counter-speech initiatives, users can challenge and reshape platform norms and practices (Obermaier 2024; Poole, Giraud and de Quincey 2021). Such actions demonstrate the reciprocal relationship between structure and agency, as habitus reshapes the field while the field continues to condition habitus.

Regulatory frameworks introduce another dimension to this dynamic by acting as external forces capable of reconfiguring the digital field. Germany's Network Enforcement Act (NetzDG) is a notable example, mandating content moderation and platform accountability, thereby reshaping both platform behaviour and user expectations (Tworek and Leerssen 2019). Similarly, calls for algorithmic transparency and ethical design reflect broader societal efforts to recalibrate the relationship between platforms and users, fostering digital fields that prioritize inclusivity over engagement metrics (Couldry and Mejias 2019). The legal frameworks for addressing hate speech are further discussed in the case studies section.

This working paper examines the interplay between technology, policy and legal frameworks in addressing hate speech. Specifically, it seeks to answer three critical questions: How do algorithms and platforms contribute to the amplification or mitigation of hate speech? What are the strengths and limitations of current legal frameworks regulating hate speech online? And how can AI-driven content moderation and enhanced transparency practices improve policy effectiveness? Using Bourdieu's framework to analyze the reciprocal dynamics between digital fields and habitus, this paper provides actionable insights for policy makers, tech companies and civil society actors striving to create safer and more equitable digital environments.

Background and Methods

This section examines existing scholarship and practices related to the role of technology in amplifying hate speech, the legal responses addressing it and innovations

in content moderation, framed through the lens of Bourdieu's concepts of *habitus* and *field*.

Technology's Role in Amplifying Hate

The intersection of technology and hate speech has garnered significant scholarly attention, with research demonstrating how digital platforms amplify and normalize hateful content (Bartlett et al. 2014; Felmlee, Inara Rodis and Zhang 2020; Gelber and McNamara 2015; Howard 2019; Paz, Montero-Díaz and Moreno-Delgado 2020; Piazza 2020). Unlike traditional platforms for hate speech, digital platforms leverage algorithms that increase the visibility of polarizing and controversial content, inadvertently amplifying hate speech with broader reach (Anderson 2013; Burrell and Fourcade 2021; Tufekci 2018). The anonymity and speed afforded by online communication exacerbate these issues. Danielle Keats Citron (2014) highlights how anonymity reduces accountability, enabling individuals to express hostility they might otherwise suppress. Similarly, Alice Marwick and Rebecca Lewis (2017) discuss how networked technologies facilitate the rapid dissemination of hateful ideologies, allowing fringe groups to reach global audiences in unprecedented ways. Beyond passive dissemination, technology actively cultivates hate. Whitney Phillips (2015) emphasizes how extremists deliberately exploit algorithmic biases on social media platforms, ensuring their messages reach susceptible audiences. This pattern aligns with the concept of "platform affordances" (Boyd 2011), which suggests that the structural design of digital technologies inherently shapes user behaviours, including the propagation of hate speech.

Bourdieu (1984) offers a useful lens for understanding these dynamics through his concept of social fields — structured spaces of positions where power and capital (economic, cultural or social) are contested. Digital platforms function as such fields, with algorithms and design features acting as structuring mechanisms that dictate norms and expectations (Ignatow and Robinson 2017; Julien 2015). Optimized for engagement, these algorithms prioritize emotionally charged and polarizing content, inadvertently promoting hate speech (Anderson 2013; Burrell and Fourcade 2021; Tufekci 2018). In these digital fields, users are conditioned to engage with content aligned with platform incentives, creating echo chambers and filter bubbles that reinforce biases and polarize discourse further (Cinelli et al. 2021). This conditioning reflects Bourdieu's concept of habitus, wherein individuals' dispositions are shaped by their environment, fostering behaviours conducive to divisive engagement.

Collectively, this body of literature underscores the pressing need for updated policies and enhanced platform accountability to address the technological amplification of hate.

AI-Driven Content Moderation

AI is increasingly used to moderate hate speech at scale, shaping the digital field through automated processes. Tools such as Google's Perspective API and Twitter's machine learning models analyze text for toxicity, flagging potentially harmful content (Nahmias and Perel 2021). While these tools provide efficiency, they often struggle with context. For instance, AI frequently misinterprets sarcasm, satire or cultural references, leading to false positives and negatives (Molina and Sundar 2022).

A significant limitation of AI moderation is its inability to address hate speech in low-resource languages, leaving certain communities more vulnerable to harm. Studies emphasize the bias inherent in AI models trained predominantly on English data sets (Abid, Farooqi and Zou 2021; Bender et al. 2021). Efforts to develop multilingual

AI systems, such as Facebook's natural language processing tools, show promise but require significant investment (Goyal et al. 2022). This disparity reflects the uneven conditioning of habitus across linguistic and cultural contexts within the global digital field (Siapera 2022).

Human moderators remain essential for providing the contextual understanding necessary to evaluate nuanced cases (Roberts 2019). However, the toll on moderators' mental health highlights the ethical challenges of relying on human oversight to mitigate the limitations of AI systems (Gray and Suri 2019). This interplay between human agency and algorithmic structure underscores Bourdieu's notion of the dialectical relationship between habitus and field (Gillespie 2020; Udupa et al. 2022).

Accountability and Transparency

Accountability mechanisms influence platform norms by introducing external oversight into the digital field. Transparency reports from platforms such as Facebook, Twitter and YouTube offer insights into moderation practices, but often lack detail about error rates, appeals outcomes and regional disparities (Gillespie 2018; Klonick 2018; Roberts 2019). Critics argue that these reports, while useful, are insufficient for holding platforms accountable and call for independent audits to verify their accuracy (Vogus and Llansó 2021).

Facebook's Oversight Board, launched in 2020, represents an effort to embed external accountability within the platform's field. The board reviews disputed moderation decisions and issues policy recommendations. However, there is critique of its limited scope and lack of enforcement power, suggesting that independent oversight mechanisms require broader authority to reshape platform fields effectively (Wong and Floridi 2023).

Calls for algorithmic transparency and ethical design reflect broader societal efforts to recalibrate the digital field. Nick Couldry and Ulises A. Mejias (2019) emphasize the need for platforms to prioritize inclusivity and public accountability over engagement-driven metrics. Such efforts aim to create fields that align with democratic values, fostering user dispositions that reject divisive content and promote inclusivity (Gulati 2023).

The role of technology and policy in addressing hate speech emerges as a critical focus, as demonstrated across the three case studies that follow. These cases reveal how technology's design choices and operational practices can either amplify harmful content or serve as tools for mitigation, depending on the regulatory and accountability frameworks in place.

Case Study 1: TikTok and Platform Governance in the United States

Over the past two decades, platform governance in the United States has evolved in response to the rapid growth of social media and digital platforms. Initially, platforms such as Facebook, YouTube and Twitter were governed by broad protections under section 230 of the Communications Decency Act (1996), which shields platforms from liability for user-generated content while allowing them to moderate content in good faith. However, the rise of online misinformation, hate speech, and concerns about

privacy and national security have prompted a re-evaluation of these frameworks at both state and federal levels.

At the federal level, regulatory efforts have often focused on transparency, accountability and user protection. While section 230 remains foundational, successive administrations have explored reforms. Under the first Donald Trump administration, the Executive Order on Preventing Online Censorship in May 2020 directed the Federal Communications Commission to investigate whether platforms were engaging in unfair practices by moderating content with political bias. It also required federal agencies to assess their advertising spending on platforms accused of censorship. Furthermore, the administration proposed narrowing section 230 protections, specifically targeting platforms perceived to suppress conservative viewpoints. President Joe Biden's administration later took a different approach, emphasizing the need to combat misinformation and foster accountability for harmful content. For example, Biden advocated for reforms that would increase platform responsibility for moderating harmful content, such as misinformation related to public health and elections.

At the state level, laws vary significantly. States such as California have enacted the California Consumer Privacy Act, targeting privacy rights and data security. Conversely, Florida and Texas have introduced legislation aimed at preventing perceived censorship of conservative viewpoints, highlighting the fragmented and often politically charged nature of state-level governance.

The US Congress has played an active role in platform governance, intervening through a combination of legislative proposals, hearings and public inquiries into big tech's practices. Bills such as the SAFE TECH Act and the EARN IT Act aimed to refine section 230, introducing measures to hold platforms accountable for harmful content. High-profile congressional hearings with CEOs of major tech companies such as Facebook, Google, Twitter and TikTok have scrutinized platform practices, ranging from content moderation to algorithmic transparency and data privacy.

Under this pressure, platforms introduced reforms, including enhanced moderation policies, transparency reports, AI tools to identify hate speech and user data safeguards. However, the extent of the improvement remains uneven. While platforms have made strides in removing harmful content, critics argue that systemic issues persist, such as the algorithmic amplification of polarizing content (Cinelli et al. 2021; Daniels 2018; Gillespie 2018; Noble 2018; Schmitt et al. 2018; Tufekci 2018).

TikTok, owned by the Chinese company ByteDance, has faced intense scrutiny in the United States under the banner of national security. Concerns centre on the potential misuse of user data by the Chinese government, alongside the platform's role in amplifying hate speech and misinformation (Vergun 2023). These concerns culminated in significant actions. In 2023, President Biden signed a bill, the No TikTok on Government Devices Act, into law, granting authority to ban TikTok on federal devices. Some states, such as Montana, moved to ban the platform outright within their jurisdictions with Senate Bill 419. These actions signal the prioritization of platform governance when framed as a national security issue, and heightened scrutiny suggests that with sufficient political will, regulatory frameworks can address data privacy, as well as the amplification of potentially harmful content

Case Study 2: The Freedom Convoy Protests in Canada

The Freedom Convoy protests of 2022 underscore how social media platforms such as Facebook and Telegram acted as digital fields facilitating the organization and amplification of hate speech (Osman 2022). These platforms provided spaces where symbols of hate and targeted harassment against racialized communities and public officials were normalized (Daniels 2018; Ganesh and Bright 2020). Driven by priorities of engagement and virality, these platforms allowed harmful content to spread with minimal intervention (Anderson 2013; Burrell and Fourcade 2021; Tufekci 2018), highlighting unchecked dynamics that prioritize user interaction over content moderation (Gillies, Raynauld and Wisniewski 2023; *PressProgress* 2022).

This digital field shaped the collective habitus of users, fostering dispositions that legitimized hate speech under the guise of free expression. Bourdieu's framework illustrates how these platforms not only reflect societal power dynamics but also reinforce them. In the context of the Freedom Convoy protests, platform dynamics amplified narratives of intolerance and hostility, further marginalizing dissenting voices and deepening societal divisions (Askanius, Molas and Amarasingam 2024; Roy and Gandsman 2023).

Proposed legislation such as Bill C-63 aims to disrupt these dynamics through regulatory mechanisms designed to reconfigure the digital field. The bill proposes a Digital Safety Commissioner and mandated transparency requirements to establish new norms prioritizing harm reduction over engagement-driven practices. These measures seek to reshape user behaviours toward rejecting hate speech and promoting inclusivity. However, the effectiveness of these interventions hinges on robust enforcement and holding platforms accountable for their role in shaping harmful societal behaviours (Tenove and Tworek 2024).

Enforcement is a critical concern addressed by Bill C-63. The bill empowers the Digital Safety Commissioner to oversee compliance with new digital safety standards, ensuring platforms promptly remove harmful content. These compliance measures include imposing penalties for non-compliance, which aims to compel platforms to take swift action against hate speech and related content. By establishing clear guidelines and timelines for content moderation, Bill C-63 aims to eliminate the ambiguity that has allowed harmful narratives to persist unchecked. Moreover, Bill C-63 enhances enforcement through transparency requirements. Platforms must disclose their content moderation practices, detailing how they identify and remove harmful content and explaining their decision-making processes. This transparency aims to foster accountability and public trust, ensuring platforms uphold digital safety standards consistently and effectively (Gillies, Raynauld and Wisniewski 2023).

While these measures represent significant strides toward curbing online harm, challenges remain in ensuring universal compliance and addressing emerging forms of digital threats. Effective implementation will require continuous adaptation to evolving digital landscapes and close collaboration between regulatory bodies, platforms and civil society (Tenove and Tworek 2024).

Case Study 3: Germany's NetzDG and the European Union's Digital Services Act

Germany's NetzDG (2017) provides a compelling example of how regulatory frameworks can mitigate harm in digital spaces while striving to preserve essential freedoms. The requirement to remove unlawful content within 24 hours of notification fundamentally shifted platform behaviour, compelling companies to prioritize the timely moderation of harmful content. This intervention significantly reduced exposure to hate speech, fostering a digital environment where such behaviour became less visible and socially acceptable. Additionally, NetzDG introduced measures to ensure platform accountability through mandatory transparency reporting. Platforms must publish biannual reports detailing the number of complaints received, the volume of content removed and the processes used for moderation (Schulz 2018). These reports provide public scrutiny and encourage consistency in enforcement practices, illustrating how transparency mechanisms can strengthen platform governance.

At the same time, NetzDG incorporates safeguards to protect free expression. Judicial oversight allows users to challenge content removal decisions, offering a mechanism to address over-censorship and ensuring that legitimate discourse is not unduly stifled (Tworek and Leerssen 2019). These safeguards are critical for preserving democratic values within the digital field. However, the implementation of NetzDG has revealed complexities in enforcement. Platforms, seeking to avoid penalties, have often pre-emptively removed borderline content, raising concerns about over-censorship. This highlights the tension noted by Bourdieu, as overly rigid field structures can suppress agency and creativity among participants.

The critical insights from NetzDG's implementation underscore its dual impact: while effectively recalibrating the digital field to reduce harm, it also demonstrates the risks of stifling legitimate expression. The law's transparency requirements and judicial safeguards serve as valuable mechanisms for balancing moderation with freedom. Ultimately, NetzDG offers a nuanced blueprint for how regulatory measures can mitigate harm while preserving essential freedoms, emphasizing the need for a dynamic approach to digital governance that evolves with societal needs.

NetzDG also provides a compelling example of a national framework that complements broader regulatory efforts such as EU Digital Services Act (DSA). NetzDG coexists with the DSA by focusing on national enforcement while adhering to EU-wide principles. For example, the DSA provides a harmonized framework for defining and addressing harmful content, ensuring consistency across member states, while NetzDG tailors these principles to Germany's legal and cultural context. Both frameworks incorporate safeguards for free expression, with NetzDG allowing judicial oversight to challenge content removal decisions (ibid.). This safeguard aligns with the DSA's requirement for transparent appeals processes, ensuring users can contest moderation actions.

Findings and Conclusion

Across these case studies, Bourdieu's concepts of habitus and field offer a valuable lens through which to understand the dynamic interplay between users, digital platforms and evolving regulatory structures. The design and operational logics of platforms — such as prioritizing emotionally charged, engagement-driven content — condition users' behaviours and perceptions, often amplifying hate speech and divisive narratives. In this sense, the field exerts a structuring force on habitus, guiding users toward patterns of interaction that may reinforce harmful norms and exclude alternative viewpoints.

Regulatory frameworks and collective user actions emerge as critical avenues to disrupt these entrenched cycles. Germany's NetzDG serves as an instructive example of how legal mandates can recalibrate digital fields. By compelling platforms to prioritize content moderation, transparency and accountability, the intervention shifts the structural conditions under which users operate. Furthermore, NetzDG's enforced removal timelines and required transparency reports, as well as the DSA's harmonized EU-wide standards, illustrate how state-level and supranational regulations can realign platform behaviours.

The efficacy of such regulatory interventions also depends on supporting user agency. Although deeply influenced by the structural logics of platforms, users are not entirely passive. The provision of greater algorithmic transparency, appeals processes and user control tools can create opportunities for individuals and communities to resist dominant platform imperatives. This reciprocal relationship between field and habitus underscores that neither regulation nor user action alone can fully address the complexities of hate speech online. Instead, meaningful solutions lie at the intersection of well-enforced regulatory frameworks, intentional platform design choices, effective AI-driven moderation tools and empowered user communities. Together, they can reconfigure the digital field, guiding habitus toward dispositions that uphold democratic values, foster constructive discourse and, ultimately, contribute to safer and more equitable digital environments.

Recommendations

To address the amplification of hate speech and promote accountability, this working paper proposes a set of targeted policy interventions that align technology, policy and ethical practices. These recommendations aim to foster safer digital environments by aligning technological systems with ethical practices, strengthening legal accountability and enabling users to contribute to a more equitable online landscape.

- **Recommendation 1: Strengthen legal frameworks and global cooperation.** Effective hate speech regulation requires cohesive legal frameworks that balance enforcement with free expression. International standards, such as the UN Rabat Plan of Action, should harmonize definitions of hate speech and enforcement practices to foster consistency across jurisdictions. Regional regulations such as the European Union's DSA can serve as adaptable models, emphasizing transparent appeals mechanisms and tailoring to local cultural contexts.

- **Recommendation 2: Enhance accountability through transparency.** Platforms must adopt comprehensive transparency measures to ensure accountability in content moderation. Detailed transparency reports should include data on flagged content, error rates, appeals outcomes and regional disparities. Independent oversight bodies with enforcement authority should monitor platform practices and ensure compliance. Platforms can further build public trust by implementing real-time updates on flagged and moderated content.
- **Recommendation 3: Guide users through education.** Users play a vital role in combatting hate speech. Digital literacy campaigns should educate individuals on identifying and addressing hate speech and misinformation, with a focus on vulnerable populations. Simplified reporting tools should make it easier for users to flag harmful content and receive timely responses. Counter-speech initiatives that promote positive, inclusive messaging should be supported to challenge hate narratives effectively.

Acknowledgements

I am grateful to Alex He and Caleigh Wong for their generous and insightful comments.

About the Author

Sophie Liu is a Digital Policy Hub doctoral fellow and Ph.D. candidate in sociology at the University of British Columbia, specializing in law and society, race and migration. Her dissertation focuses on Canadian society's response to hate, including online hate expression. Her particular areas of interest include digital platform regulation, pathways to justice for individuals impacted by hate, and the role of hate-crime and hate-incident data in policy making. She employs diverse methodologies to explore these issues, including survey experiments, in-depth interviews and content analysis.

Works Cited

AAbid, Abubakar, Maheen Farooqi and James Zou. 2021. "Large language models associate Muslims with violence." *Nature Machine Intelligence* 3: 461–63. https://doi.org/10.1038/s42256-021-00359-2.

Anderson, Christopher W. 2013. "Towards a sociology of computational and algorithmic journalism." *New Media & Society* 15 (7): 1005–21. https://doi.org/10.1177/1461444812465137.

Askanius, Tina, Bàrbara Molas and Amarnath Amarasingam. 2024. "Far-right extremist narratives in Canadian and Swedish COVID-19 protests: a comparative case study of the Freedom Movement and Freedom Convoy." *Behavioral Sciences of Terrorism and Political Aggression:* 17 (2): 164–84. https://doi.org/10.1080/19434472.2024.2340492.

Bakshy, Eytan, Solomon Messing and Lada A. Adamic. 2015. "Exposure to ideologically diverse news and opinion on Facebook." *Science* 348 (6239): 1130–32. https://doi.org/10.1126/science.aaa1160.

Banks, James. 2010. “Regulating hate speech online.” *International Review of Law, Computers & Technology* 24 (3): 233–39. http://dx.doi.org/10.1080/13600869.2010.522323.

Bartlett, Jamie, Jeremy Reffin, Noelle Rumball and Sarah Williamson. 2014. *Anti-social Media*. London, UK: Demos. https://demos.co.uk/wp-content/uploads/2014/02/DEMOS_Anti-social_Media.pdf.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. https://doi.org/10.1145/3442188.3445922.

Boulianne, Shelley and Sangwon Lee. 2022. “Conspiracy Beliefs, Misinformation, Social Media Platforms, and Protest Participation.” *Media and Communication* 10 (4): 30–41. https://doi.org/10.17645/mac.v10i4.5667.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgement of Taste*. Translated by Richard Nice. Cambridge, MA: Harvard University Press.

Boyd, Danah. 2011. “Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications.” In *A Networked Self: Identity, Community, and Culture on Social Network Sites*, edited by Zizi Papacharissi, 47–66. New York, NY: Routledge.

Brown, Alexander. 2018. “What is so special about online (as compared to offline) hate speech?” *Ethnicities* 18 (3): 297–326. https://doi.org/10.1177/1468796817709846.

Burrell, Jenna and Marion Fourcade. 2021. “The Society of Algorithms.” *Annual Review of Sociology* 47: 213–37. https://doi.org/10.1146/annurev-soc-090820-020800.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi and Michele Starnini. 2021. “The Echo Chamber Effect on Social Media.” *Proceedings of the National Academy of Sciences* 118 (9). https://doi.org/10.1073/pnas.2023301118.

Citron, Danielle Keats. 2014. *Hate Crimes in Cyberspace*. Cambridge, MA: Harvard University Press.

Couldry, Nick and Ulises A. Mejias. 2019. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Redwood City, CA: Stanford University Press.

Daniels, Jessie. 2018. “The Algorithmic Rise of the ‘Alt-Right.’” *Contexts* 17 (1): 60–65. https://doi.org/10.1177/1536504218766547.

Felmlee, Diane, Paulina Inara Rodis and Amy Zhang. 2020. ‘‘Sexist Slurs: Reinforcing Feminine Stereotypes Online.’’ *Sex Roles* 83: 16–28. https://doi.org/10.1007/s11199-019-01095-z.

Flaxman, Seth, Sharad Goel and Justin M. Rao. 2016. “Filter Bubbles, Echo Chambers, and Online News Consumption.” *Public Opinion Quarterly* 80 (S1): 298–320. https://doi.org/10.1093/poq/nfw006.

Ganesh, Bharath and Jonathan Bright. 2020. “Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation.” *Policy & Internet* 12 (1): 6–19.

Gelber, Katharine and Luke McNamara. 2015. “Evidencing the harms of hate speech.” *Social Identities* 22 (3): 324–41. https://doi.org/10.1080/13504630.2015.1128810.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.

— — —. 2020. “Content moderation, AI, and the question of scale.” *Big Data & Society* 7 (2). https://doi.org/10.1177/2053951720943234.

Gillies, Jamie, Vincent Raynauld and Angela Wisniewski. 2023. “Canada is No Exception: The 2022 Freedom Convoy, Political Entanglement, and Identity-Driven Protest.” *American Behavioral Scientist*. https://doi.org/10.1177/00027642231166885.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán and Angela Fan. 2022. “The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation.” *Transactions of the Association for Computational Linguistics* 10: 522–38. https://doi.org/10.1162/tacl_a_00474.

Gray, Mary L. and Siddharth Suri. 2019. Ghost Work: *How to Stop Silicon Valley from Building a New Global Underclass*. Boston, MA: Houghton Mifflin Harcourt.

Gulati, Rishi. 2023. “Meta’s Oversight Board and Transnational Hybrid Adjudication — What Consequences for International Law?” *German Law Journal* 24 (3): 473–93. https://doi.org/10.1017/glj.2023.34.

Howard, Jeffrey W. 2019. “Free Speech and Hate Speech.” *Annual Review of Political Science* 22 (1): 93–109. https://doi.org/10.1146/annurev-polisci-051517-012343.

Ignatow, Gabe and Laura Robinson. 2017. “Pierre Bourdieu: theorizing the digital.” I*nformation, Communication & Society* 20 (7): 950–66. https://doi.org/10.1080/1369118X.2017.1301519.

Inara Rodis, Paulina d. C. 2024. “The Managed Response: Digital Emotional Labor in Navigating Intersectional Cyber Aggression.” *Social Psychology Quarterly* 87 (1): 1–21. https://doi.org/10.1177/01902725231166637.

Julien, Chris. 2015. “Bourdieu, Social Capital and Online Interaction.” *Sociology* 49 (2): 356–73. https://doi.org/10.1177/0038038514535862.

Klonick, Kate. 2018. “The New Governors: The People, Rules, and Processes Governing Online Speech.” Harvard Law Review 131 (6): 1598–670. https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/.

Levina, Natalia and Manuel Arriaga. 2014. “Distinction and Status Production on User-Generated Content Platforms: Using Bourdieu’s Theory of Cultural Production to Understand Social Dynamics in Online Fields.” *Information Systems Research* 25 (3): 468–88. https://doi.org/10.1287/isre.2014.0535.

Marwick, Alice and Rebecca Lewis. 2017. *Media Manipulation and Disinformation Online*. New York, NY: Data & Society Research Institute. https://datasociety.net/library/media-manipulation-and-disinfo-online/.

Matamoros-Fernández, Ariadna. 2017. “Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube.” *Information, Communication & Society* 20 (6): 930–46. https://doi.org/10.1080/1369118X.2017.1293130.

Molina, Maria D. and S. Shyam Sundar. 2022. “When AI moderates online content: effects of human collaboration and interactive transparency on user trust.” *Journal of Computer-Mediated Communication* 27 (4). https://doi.org/10.1093/jcmc/zmac010.

Nahmias, Yifat and Maayan Perel. 2021. “The Oversight of Content Moderation by AI: Impact Assessments and their Limitations.” *Harvard Journal on Legislation* 58 (1): 145–94. https://journals.law.harvard.edu/jol/wp-content/uploads/sites/86/2021/02/105_Nahmias.pdf.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: New York University Press.

Obermaier, Magdalena. 2024. "Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech." *New Media & Society* 26 (8): 4785–807. https://doi.org/10.1177/14614448221125417.

Osman, Laura. 2022. "Social media tools were key to 'Freedom Convoy' protest, expert tells inquiry." CBC News, November 29. www.cbc.ca/news/politics/social-mediea-convoy-protests-emergencies-act-inquiry-1.6668543.

Paz, María Antonia, Julio Montero-Díaz and Alicia Moreno-Delgado. 2020. "Hate Speech: A Systematized Review." *SAGE Open* 10 (4). https://doi.org/10.1177/2158244020973022.

Phillips, Whitney. 2015. *This is Why We Can't Have Nice Things: Mapping the Relationship between Online Trolling and Mainstream Culture*. Cambridge, MA: MIT Press.

Piazza, James A. 2020. "Politician hate speech and domestic terrorism." *International Interactions* 46 (3): 431–53. https://doi.org/10.1080/03050629.2020.1739033.

Poole, Elizabeth, Eva Haifa Giraud and Ed de Quincey. 2021. "Tactical interventions in online hate speech: The case of #stopIslam." *New Media & Society* 23 (6): 1415–42. https://doi.org/10.1177/1461444820903319.

PressProgress. 2022. "Meet the Extremists and Social Media Influencers at the Centre of the Far-Right Siege of Ottawa." *PressProgress*, February 8.https://pressprogress.ca/meet-the-extremists-and-social-media-influencers-at-the-centre-of-the-far-right-siege-of-ottawa/.

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

Roy, Mélissa and Ari Gandsman. 2023. "Polarizing figures of resistance during epidemics. A comparative frame analysis of the COVID-19 freedom convoy." *Critical Public Health* 33 (5): 788–802. https://doi.org/10.1080/09581596.2023.2284633.

Schmitt, Josephine B., Diana Rieger, Olivia Rutkowski and Julian Ernst. 2018. "Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube: Recommendation Algorithms." *Journal of Communication* 68 (4): 780–808. https://doi.org/10.1093/joc/jqy029.

Schulz, Wolfgang. 2018. "Regulating Intermediaries to Protect Privacy Online – the Case of the German NetzDG." HIIG Discussion Paper 2018-01. January. Berlin, Germany: Humboldt Institute for Internet and Society. www.hiig.de/wp-content/uploads/2018/07/SSRN-id3216572.pdf.

Siapera, Eugenia. 2022. "AI Content Moderation, Racism and (de)Coloniality." *International Journal of Bullying Prevention* 4: 55–65. https://doi.org/10.1007/s42380-021-00105-7.

Silva, Leandro, Mainack Mondal, Denzil Correa, Fabrício Benevenuto and Ingmar Weber. 2021. "Analyzing the Targets of Hate in Online Social Media." *Proceedings of the International AAAI Conference on Web and Social Media 10* (1): 687–90. https://doi.org/10.1609/icwsm.v10i1.14811.

Tenove, Chris and Heidi Tworek. 2024. "What Lessons Did Canada Learn Before Creating Its Online Harms Bill?" *Tech Policy Press*, March 12. www.techpolicy.press/what-lessons-did-canada-learn-before-creating-its-online-harms-bill/.

Tufekci, Zeynep. 2018. "YouTube, the Great Radicalizer." Opinion, *The New York Times*, March 10.www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.

Tworek, Heidi and Paddy Leerssen. 2019. “An Analysis of Germany’s NetzDG Law.” Transatlantic Working Group. April 15. Amsterdam, the Netherlands: Institute for Information Law. www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf.

Udupa, Sahana, Antonis Maronikolakis, Hinrich Schütze and Axel Wisiorek. 2022. “Ethical Scaling for Content Moderation: Extreme Speech and the (In)Significance of Artificial Intelligence.” June 9. Cambridge, MA: The Shorenstein Center on Media, Politics and Public Policy. https://shorensteincenter.org/ethical-scaling-content-moderation-extreme-speech-insignificance-artificial-intelligence/.

Vergun, David. 2023. “Leaders Say TikTok Is Potential Cybersecurity Risk to U.S.” U.S. Department of Defense, April 6. www.defense.gov/News/News-Stories/Article/Article/3354874/leaders-say-tiktok-is-potential-cybersecurity-risk-to-us/.

Vogus, Caitlin and Emma Llansó. 2021. *Making Transparency Meaningful: A Framework for Policymakers*. Washington, DC: The Center for Democracy & Technology. December. https://cdt.org/wp-content/uploads/2021/12/12132021-CDT-Making-Transparency-Meaningful-A-Framework-for-Policymakers-final.pdf.

Wong, David and Luciano Floridi. 2023. “Meta’s Oversight Board: A Review and Critical Assessment.” *Minds and Machines* 33: 261–84. https://doi.org/10.1007/s11023-022-09613-x.

Zhang, Xinyi and Mark Davis. 2022. “E-extremism: A conceptual framework for studying the online far right.” *New Media & Society* 26 (5): 2954–70. https://doi.org/10.1177/14614448221098360.