INCITEMENT AND THE GEOPOLITICAL INFLUENCE OF FACEBOOK CONTENT MODERATION

Sarah Koslov[*]

CITE AS: 4 GEO. L. TECH. REV. 183 (2019)

TABLE OF CONTENTS

[*] J.D. Candidate, Georgetown University Law Center, 2020; B.A, University of South Carolina, 2014. Thank you to Professor Julie Cohen for her guidance, insight, and feedback throughout the development and writing of this note. I am also grateful to Lyn Abdullah and the editors of GLTR for their thoughtful comments and assistance.

I. INTRODUCTION

Between 2015 and 2016, a new surge of violence broke out against Israeli civilians and soldiers, typified by lone actors using knives to attack at sudden and opportunistic times.[1] Most often, assailants were Palestinian men without ties to any formal organization or a history of engaging in violence.[2] During this period, attackers killed thirty-four Israelis while nearly two hundred Palestinians died trying to carry out the stabbings.[3] While there remained no clear evidence that the attacks were organized or supported by a particular faction of the Palestinian community, some Palestinians took to social media voicing encouragement and support for the attackers.[4] Opposing

[1] Peter Beaumont, *What's Driving the Young Lone Wolves Who Are Stalking the Streets of Israel?*, GUARDIAN (Oct. 17, 2015), https://www.theguardian.com/world/2015/oct/18/knife-intifada-palestinian-israel-west-bank [https://perma.cc/B36X-U6EC]; William Booth & Ruth Eglash, *Israelis Are Calling Attacks a "New Kind of Palestinian Terrorism,"* WASH. POST (Dec. 25, 2015), https://www.washingtonpost.com/world/middle_east/israelis-are-calling-attacks-a-new-kind-of-palestinian-terror/2015/12/24/e162e088-0953-4de5-992e-adb2126f1dcc_story.html [https://perma.cc/Z2X8-4YK4].

[2] Booth & Eglash, *supra* note 1.

[3] Ruth Eglash, William Booth & Darlas Cameron, *A New Kind of Terrorism in Israel*, WASH. POST, https://www.washingtonpost.com/graphics/world/israel-palestine-deaths/ (last entry Sept. 2016) [https://perma.cc/FMX6-XCWU]; *see also* Pierrick Leurent, *Palestinian "Knife Intifada" Reflects a Generation's Despair*, FRANCE 24 (June 5, 2016), https://www.france24.com/en/20160506-reporter-israel-knife-intifada-palestinian-territories-violence [https://perma.cc/J3EX-RSLL].

[4] *Is Palestinian-Israeli violence being driven by social media?*, BBC NEWS (October 22, 2015), https://www.bbc.com/news/world-middle-east-34513693 [https://perma.cc/6ZBP-7QUS].

narratives developed around the "Knife Intifada"[5] and the motivations behind it. The Israeli government blamed the violence on inciting posts on Facebook, claiming that content encouraging violence against Israelis permeated the platform. Meanwhile, Palestinians argued that the attackers acted out of frustration, desperation, and a loss of hope after years of occupation and displacement.[6]

The Israeli government's actions following the attacks illustrate how Facebook has emerged as an influential intermediary in politically conflicted territories. In the midst of the Knife Intifada, the Israeli government called on Facebook to aid it in combating lone-wolf attacks.[7] Israeli Justice Minister Ayelet Shaked explained that the government wanted Facebook "to themselves remove posts by terrorist groups and incitement to terrorism without us having to flag each individual post."[8] Frustrated that Facebook was not cooperating satisfactorily,[9] Public Security Minister Gilad Erdan appeared on television calling Facebook a "monster" and said "some of the victims' blood is on Zuckerberg's hands."[10] In addition, the Knesset began working on a new "Facebook Bill" granting law enforcement officials broad authority to seek court orders compelling Facebook to remove content based on police recommendations.[11] Facebook, which usually prefers to wield its power

[5] Beaumont, *supra* note 1.

[6] 7AMLEH, FACEBOOK AND PALESTINIANS: BIASED OR NEUTRAL CONTENT MODERATION POLICIES 7 (2018), https://7amleh.org/wp-content/uploads/2018/10/booklet-final2-1.pdf [https://perma.cc/8ZCV-2AV7]; Beaumont, *supra* note 1; Bethan McKernan, *Israel and Facebook Team Up to Combat Social Media Posts That Incite Violence*, INDEP. (Sept. 14, 2016), https://www.independent.co.uk/news/world/middle-east/israel-facebook-team-up-social-media-posts-incitement-violence-a7306436.html [https://perma.cc/6NCB-V9KT].

[7] Amar Toor, *Israel Calls Facebook a "Monster" for Not Helping To Curb Violence*, VERGE (July 4, 2016), https://www.theverge.com/2016/7/4/12092762/israel-facebook-palestinian-attacks-censorship [https://perma.cc/M476-4TB4].

[8] *Id.*

[9] Mark Bergen*, Israel's Public Security Minister Blames Facebook After Recent West Bank Attacks*, RECODE (July 3, 2016), https://www.recode.net/2016/7/3/12090532/israel-facebook-west-bank [https://perma.cc/6PL2-HU36]; David Wainer, *Israel Accuses Facebook of Complicity in West Bank Violence*, BLOOMBERG BUS. (July 3, 2016), https://www.bloomberg.com/news/articles/2016-07-03/israel-accuses-facebook-of-contributing-west-bank-violence [https://perma.cc/H7PX-RHEZ].

[10] Jonathan Lis, *Israeli Minister Slams Facebook:* "*Terror Victims' Blood Is on Zuckerberg's Hands,"* HAARETZ (July 3, 2016), https://www.haaretz.com/israel-news/israeli-minister-terror-victims-blood-is-on-zuckerberg-s-hands-1.5404675 [https://perma.cc/9SAS-T7Z7].

[11] Shoshannah Solomon, *Israel's Facebook Bill May Endanger Democracy, Company Official Implies,* TIMES ISR. (Jan. 18, 2017), https://www.timesofisrael.com/israels-facebook-bill-may-endanger-democracy-company-official-implies/ [https://perma.cc/C869-RXVG].

quietly and behind closed doors,[12] was receptive to this public pressure. By September of 2016, Facebook had reached an informal agreement with the Israeli government to work together to address incitement on its platform.

While Facebook agreed to remove content that incites violence, promotes hate speech, or involves terrorism,[13] deciding which content falls within the scope of these terms has proven a complicated task. News outlets reported that Israel and Facebook agreed to "create teams that would figure out how best to monitor and remove inflammatory content," but no additional information was given to the public.[14] This news troubled Palestinian users, who expressed concern that Facebook was "adopting Israeli policy and terminology when it comes to defining what incitement is."[15] With ninety-six percent of Palestinians reporting that their primary use for Facebook is following the news,[16] decisions around who can access and exchange information on the platform are matters of great importance. They worried that this partnership, in effect, would lead not only to the removal of unlawful speech, but also stifle legitimate forms of dissent that conflict with the Israeli government's narrative and perspective.

The stakes here are high. Facebook's decisions about content removal and profile suspensions dictate the terms for participation in the "modern public square"[17] and effectively arbitrate which narratives can reach the global public.[18] In the context of the Israeli-Palestinian conflict, Facebook functions as a proxy battlefield, where longstanding geopolitical disputes are part and

[12] Facebook relies on private ordering and self-governance in shaping policy decisions that fit its business objectives. *See* discussion *infra* Part IV; s*ee also* Tarleton Gillespie, *Platforms Are Not Intermediari*es, 2 GEO. L. TECH. REV. 198, 202 (2018) (recognizing that "[t]oo often, social media platforms discuss content moderation as a problem to be solved—and solved privately and reactively.").

[13] Toor, *supra* note 7.

[14] *Shaked: "Penny Has Dropped" for Facebook on Incitement*, TIMES ISR. (Sept. 12, 2016), http://www.timesofisrael.com/shaked-penny-has-dropped-for-facebook-on-incitement/ [https://perma.cc/F5WM-B4AZ].

[15] Matthew Ingram, *Facebook's Censorship of Palestinian Journalists Raises Serious Questions*, FORTUNE (Sept. 28, 2016), https://fortune.com/2016/09/28/facebook-censorship-palestinian/ [perma.cc link unavailable].

[16] *Palestinians Decry Increase in Arrests for "Incitement To Violence" on Social Media*, INDEPENDENT (Apr. 4, 2018), https://www.independent.co.uk/news/world/middle-east/palestinians-israel-incitement-arrests-social-media-twitter-facebook-a8288631.html [https://perma.cc/33N3-2PEP].

[17] Packingham v. North Carolina, 137 S. Ct. 1730, 1732 (2017) (stating that barring an individual from accessing social media platforms functionally denies her access to "speaking and listening in the modern public square, and otherwise exploring the vast realms of human thought and knowledge").

[18] While content ranking and curatorial preferences influence these issues, this Note will focus only on content removal and profile suspension.

parcel of content moderation decisions that directly impact the nature of political discourse.

This Note examines Facebook's content moderation practices through the lens of the Israeli-Palestinian geopolitical conflict to highlight how Facebook's Community Standards and business practices—along with its informal relationships with individual nation states—obfuscate traditional allocations of responsibility, accountability, and power in democratic societies. This Note also challenges the inevitability of this construction by first examining the elements of Facebook's current content moderation practices that leave it vulnerable to manipulation, and then identifying opportunities to address these issues. Part II considers Facebook's content moderation practices in terms of its Community Standards and compliance with local law. Part III then locates enabling factors that make Facebook susceptible to biased execution of content moderation in relation to Israeli and Palestinian narratives. Finally, Part IV analyzes several conceptual frameworks to understand the new power dynamic among Facebook, individuals, and governments, positing steps to foster greater accountability and evenhandedness on the platform. Building on this analysis, Part IV also proposes that the platform shift away from a posture of neutrality and towards a stance informed by transparency and established human rights prinples.

While the proliferation of extremism online is an important issue, this Note does not address organized terrorist activity on Facebook. Moreover, the experiences and case studies highlighted in this writing are included to illustrate the complex power dynamics at play on the platform; however, they by no means represent the only narratives or experiences impacted by Facebook's content moderation practices in the region. Finally, this Note does not stake a position on a path forward in resolving Israeli-Palestinian territorial disputes. Rather, the analysis here aims to delineate sources of power exercised on Facebook's platform and demonstrate that transparency and an orientation towards human rights can improve the role that platform governance plays in relation to geopolitical conflict.

II. AN OVERVIEW OF FACEBOOK'S COMMUNITY STANDARDS: GUIDELINES, IMPLEMENTATION, AND GOVERNANCE IN ISRAELI AND PALESTINIAN TERRITORIES

Facebook's Community Standards are guidelines that inform content removal and profile suspension decisions on the platform. Their stated goal "is to encourage expression and create a safe environment" for users to connect and communicate.[19] The Community Standards are organized into six categories: (1) violence and criminal behavior;[20] (2) safety;[21] (3) objectionable content;[22] (4) integrity and authenticity;[23] (5) respecting intellectual property;[24] and (6) content-related requests.[25] While the publicly available Community Standards serve as general guidelines, the company also circulates internal memoranda and training documents to moderators that more concretely instruct moderators' decision-making.[26] The Community Standards remain especially important because they are the primary lens through which Facebook considers content moderation decisions absent legal requirements specific to a particular regional jurisdiction. Consequently, the Community Standards are an authority cited to justify restricting speech that are distinct from public law. In an effort to understand the way content moderation decisions are made in geopolitically contested territories, Part II examines Facebook's public Community Standards, internal guidance documents, and the conflation of Facebook's private governance and its compliance with local law.

A. Facebook's Publicly Available Community Standards Are Articulated in Broad Terms to Facilitate Global Applicability, but Their Acontextual

[19] *Community Standards*, FACEBOOK (2019). https://www.facebook.com/communitystandards/introduction (accessed Oct. 30, 2019) [https://perma.cc/C75T-EWH3].

[20] *Violence and Criminal Behavior*, FACEBOOK (2019), https://www.facebook.com/communitystandards/violence_criminal_behavior (accessed Nov. 20, 2019) [perma.cc/429J-NAQB]

[21] *Safety*, FACEBOOK (2019), https://www.facebook.com/communitystandards/safety (accessed Nov. 20, 2019) [https://perma.cc/EC86-PPYK].

[22] *Objectionable Content*, FACEBOOK (2019), https://www.facebook.com/communitystandards/objectionable_content (accessed Nov. 20, 2019) [https://perma.cc/M37B-6R94].

[23] *Integrity and Authenticity*, FACEBOOK (2019), https://www.facebook.com/communitystandards/integrity_authenticity (accessed Nov. 20, 2019) [https://perma.cc/4GH7-6PDV].

[24] *Respecting Intellectual Property*, FACEBOOK (2019), https://www.facebook.com/communitystandards/respecting_intellectual_property (accessed Nov. 20, 2019) [https://perma.cc/GRB9-N3UM].

[25] *Content- Related Requests*, FACEBOOK (2019), https://www.facebook.com/communitystandards/content_related_requests (accessed Nov. 20, 2019) [https://perma.cc/M66M-7QAT].

[26] Max Fisher, *Inside Facebook's Secret Rulebook for Global Political Speech*, N.Y. TIMES (Dec. 27, 2018), https://www.nytimes.com/2018/12/27/world/facebook-moderators.html [https://perma.cc/9759-SHJH]; Nick Hopkins, *Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence*, GUARDIAN (May 21, 2017), https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence [https://perma.cc/A3VC-VC8F].

Framing Invites Subjective and Arbitrary Interpretations that Fill Policy Gaps.

The Community Standards constitute a single set of rules applied to every country in which Facebook operates, meaning that content moderators endeavor to apply the same set rules to Facebook's 2.3 billion monthly users around the world regardless of social or political context.[27] By virtue of their global applicability, Facebook's public-facing documents employ broad, flexible language. While this language is geared toward addressing myriad circumstances, its utility is limited by the lack of context-specific insight needed to implement evenhanded, consistent enforcement.[28]

The Community Standards' high level of generality and lack of detail leave content moderators ill-equipped to navigate difficult questions in discrete circumstances. The Standards, for example, provide that the platform will "remove content that expresses support or praise for groups, leaders, or individuals involved" in terrorist or criminal activities or organized violence.[29] It is not clear, however, how these rules might apply to posts that support or praise non-violent and non-terrorist humanitarian aid efforts that also share a connection to extremist or terrorist groups. For instance, Facebook's guidelines do not provide insight as to how a post that praises programs funded by the Holy Land Foundation, a Muslim charity that supports international relief programs affiliated with Hamas, might fare under its Community Standards.[30] These complicated dynamics require context-specific knowledge, which Facebook has not accounted for in its content moderation practices (at least insofar as they have been explained to the public).

The lack of formal standardization is also apparent when considering Facebook's range of possible consequences for violating the Community Standards. Facebook states that users who violate Community Standards may

[27] Casey Newton, *The Trauma Floor: The Secret Lives of Facebook Moderators in America*, VERGE (Feb. 25, 2019), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [https://perma.cc/2PCE-722C].

[28] This applies both to helping users understand the requirements for compliance to be able to engage on the platform and to instructing moderators on which posts are impermissible in particular circumstances.

[29] *Violence and Criminal Behavior*, FACEBOOK (2019), https://www.facebook.com/communitystandards/violence_criminal_behavior (accessed Oct. 30, 2019) [https://perma.cc/K4GQ-S2G6].

[30] Fazia Patel & Rachel Levinson-Walsman, *Facebook and Free Speech*, JUST SECURITY (May 24, 2018), https://www.justsecurity.org/56864/facebook-takedown-rules-raise-questions-require-transparency/ [https://perma.cc/D25F-QX38]; *see generally* Holy Land Found. for Relief and Dev. v. Ashcroft, 333 F.3d 156 (D.C. Cir. 2003); Laurie Goodstein, *U.S. Muslims Taken Aback by a Charity's Conviction*, N.Y. TIMES (Nov. 25, 2008), https://www.nytimes.com/2008/11/26/us/26charity.html [https://perma.cc/EQA6-P49P].

experience different consequences depending on the severity of the offense and the user's "history on the platform."[31] Here, history seems primarily to refer to the frequency with which an account has violated the Community Standards.[32] However, that explanation does not consider how seemingly random or inconsistent enforcement actions may influence frequency, nor does it account for whether erroneous content flagging or removals contribute to one's history on the platform.

In a similar vein, Facebook tells users that, based on severity, the platform may provide content to law enforcement in an effort "to prevent real-world harm."[33] The guidelines, though, do not elaborate on what such instances might entail, other than that law enforcement may be engaged where Facebook deems it is appropriate. This objective is not inherently bad: indeed, one could argue this demonstrates responsible platform governance. However, in the context of Israel and Palestine, provisions like this one remain a source of concern in balancing individuals' safety and autonomy. An effort by Facebook officials to elaborate upon the circumstances in which they coordinate with law enforcement would help mitigate speculation and suspicion discussed later in this analysis.

B. Facebook's Internal Guidelines Reveal an Evolving, Patchwork Approach to Content Moderation that Contributes to Uneven Enforcement of the Community Standards.

While the public Community Standards lack culture-specific considerations, leaked internal guidelines illustrate ways that double standards and biases can shape content moderation patterns.[34] Although Facebook's internal content moderation guidelines are more specific than publicly available resources, content moderators report that internal materials are frequently changed on an ad hoc basis.[35] Moderators have also claimed that there is not a reliable, uniform handbook, "master file[,] or overarching guide"

[31] *Community Standards*, supra note 19.

[32] *Id.*

[33] *Community Standards Enforcement Report*, FACEBOOK (2019), https://transparency.facebook.com/community-standards-enforcement (accessed Oct. 30, 2019) [https://perma.cc/K4AJ-F748?type=image].

[34] Fisher, *supra* note 26.

[35] Newton, *supra* note 27 ("While official policy changes typically arrive every other Wednesday, incremental guidance about developing issues is distributed on a near-daily basis."); *see also* Fisher, *supra* note 26.

to reference as new changes are announced.[36] Consequently, for any given enforcement decision, moderators find that they have several sources of "truth to consider" from within the organization's internal guidance documents, which contributes to inconsistent moderation outcomes.[37] These constantly evolving and malleable guidelines contribute to the dissimilar treatment of similar content posted on the platform.[38] A 2017 ProPublica investigative report found that training materials for Facebook employees "banned posts that praise the use of 'violence to resist occupation of an internationally recognized state.'"[39] It is noteworthy that the inverse scenario, praise of violence against those under occupation, is not expressly banned. Facebook reasoned that it adopted this rule because it did not want content moderators "to be in a position of deciding who is a freedom fighter."[40] While the platform reported that it dropped the rule in 2017, its procedural and substantive approach to content moderation reveals how it can be manipulated by either individual preferences or majority-rule mentality.[41]

At a macro level, the platform's fluid approach may seem justifiable due to the challenges inherent to content moderation at this scale or the need to rapidly respond to unanticipated, emerging conflicts. However, absent culturally and politically specific considerations, policy gaps leave an opening for bias to go unchecked. For instance, Facebook's 2017 training materials instructed moderators to identify hate speech using a simple formula: "protected category + attack = hate speech."[42] Here, protected categories included race, gender, and religion, but the guidelines give greater latitude to

[36] Fisher, *supra* note 26 ("Facebook says the files are only for training, but moderators say they are used as day-to-day reference materials."); *see also* Newton, *supra* note 27 ("Often, this guidance is posted to Workplace, the enterprise version of Facebook that the company introduced in 2016. Like Facebook itself, Workplace has an algorithmic News Feed that displays posts based on engagement. During a breaking news event, such as a mass shooting, managers will often post conflicting information about how to moderate individual pieces of content, which then appear out of chronological order on Workplace. Six current and former employees told me that they had made moderation mistakes based on seeing an outdated post at the top of their feed.").

[37] Newton, *supra* note 27.

[38] Fisher, *supra* note 26. For research in the context of Israeli and Palestinian experiences, see 7AMLEH, *supra* note 6.

[39] Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children*, PROPUBLICA (June 28, 2017), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms [https://perma.cc/A9TJ-RP8B].

[40] *Id.*

[41] *Id.*; *see also* Julie E. Cohen, *Law for the Platform Economy*, 51 U.C. DAVIS L. REV. 133, 151 (2017) ("[P]latform affordances for volatility, polarization, and relativization are easily manipulated for malicious or simply self-interested purposes.").

[42] Angwin & Grassegger, *supra* note 39.

posts that only refer to subsets of protected categories. Thus, white men enjoy greater protection than female drivers or black children because gender and race are both protected, but occupation and age are not.[43] As such, these policies may afford fewer protections to vulnerable groups than those it provides to less frequent targets of hate speech. Because hate speech is difficult to define acontextually, the internal guidelines can exacerbate, and potentially magnify, power disparities, through their allocation of protection and censorship.

C. The Complex Legal Landscape Within Israeli and Palestinian Territories Creates Incentives for Facebook to Streamline Content Moderation Through Their Community Standards, Disproportionately and Negatively Impacting Specific Communities.

The legal complexity and geopolitical nuances specific to Israeli and Palestinian governance over physical territories create incentives for Facebook to rely on its own policies in making governance decisions in the region. Facebook's content moderation practices flow through two intersecting channels: (1) the Community Standards outlined above; and (2) the enforcement of public law specific to the region in which the platform operates. Put differently, the platform removes content or suspends accounts either because the account violated the Community Standards or because it violated local law.

Further, the variety of legal regimes governing Palestinian territories create a complex and challenging web of legal obligations. Palestinians living in Eastern Jerusalem are subject to laws enforced by Israeli civil courts, which rely on Articles from the penal code to address "incitement to violence and terrorism."[44] Meanwhile, Palestinian communities under Israeli occupation in regions like Gaza and the West Bank are subject to Israeli military law, along with the laws of the Palestinian Authority.[45] Israeli military law provides for greater prosecutorial discretion, expanding charges of incitement to include "those who express sympathy with terrorist activities," which grants the military authority to take action against conduct that may otherwise seem lawful.[46] Economic efficiency and ease of application create strong incentives for Facebook to rely on its own Community Standards rather than apply local

[43] *Id.*

[44] 7AMLEH, *supra* note 6.

[45] *Id.*

[46] *Id.*

law.[47] While conflating the Community Standards with legal requirements streamlines Facebook's compliance burdens, it also undermines the protections properly afforded in a democratic society.

Patterns of inconsistency in applying the Community Standards reveal how power dynamics permeate platform governance. In examining Facebook's documents advising moderators on differentiating between hate speech and political expression, investigators found that "at least in some instances, the company's hate-speech rules tend[ed] to favor elites and governments over grassroots activists and racial minorities."[48] The Community Standards' lack of specificity or situational awareness leave significant policy determinations vulnerable to arbitrariness or the magnification of existing power struggles. As Professor Hannah Bloch-Wehba explains, because the public "lacks key information about how, when, and at whose direction platform governance is taking place, it is extremely difficult for the outside observer to discern what is going on, and to distinguish private action from government pressure."[49]

III. CONSIDERING THE ENABLING FACTORS: THE CHALLENGES OF ALLOCATING RESPONSIBILITY AND AGENCY TO A PRIVATE PLATFORM IN EFFORT TO PROTECT INDIVIDUAL EXPRESSIVE INTERESTS

Democratic institutions and norms are underpinned by procedural and substantive modes of accountability that aim to hold those empowered to

[47] Decisions to remove content that does not violate Community Standards can stem from Facebook's desire to placate the governments that can regulate or penalize them, whereas protecting legitimate but unpopular speech may inspire backlash or negative PR. "The easiest, cheapest, and most risk-avoidant path for any technical intermediary is simply to process a removal request and not question its validity. A company that takes an 'if in doubt, take it down' approach to requests may simply be a rational economic actor." Daphne Keller, *Empirical Evidence Of "Over-removal" By Internet Companies Under Intermediary Liability Laws*, CTR. INTERNET & SOC'Y STAN. L. SCH. (Oct. 12, 2015), http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws [https://perma.cc/K5BC-SL7T].

[48] Angwin & Grassegger, *supra* note 39; *see also* Fisher, *supra* note 26 (reporting that "Facebook instructed moderators to 'look out for' the phrase 'Free Kashmir'—though the slogan, common among activists, is completely legal.").

[49] Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27, 79 (2019).

govern accountable to those living under their rule of law.[50] The Israeli government is a parliamentary democracy with legislative, judicial, and executive branches.[51] While Facebook is not a sovereign or democratic institution, the company continuously positions and prides itself as a facilitator of free speech and proponent of democratic values.[52] However, the coordination and amorphous relationship between Facebook and Israel reconceptualizes the exercise of governing power and challenges existing democratic modes of accountability. In connection to Israeli and Palestinian users, suspected double standards, lack of procedural transparency, algorithmic bias, and covert coordination all undermine trust and accountability among users in relation to the platform and the government.

A. Suspected Double Standards in Enforcement of the Community Standards Increase Concerns of Targeted Censorship and Undermine Users' Trust in Facebook.

Absent oversight, content removal and account suspensions can function as extensions of state action hidden under the aegis of Community Standards. For example, the agreement between Israel and Facebook neutralized government efforts to pass formal legislation, but it also marked an increase in content removal and account suspensions among Palestinian users.[53] Neither Facebook nor Israeli officials provided details about what their cooperation agreement entailed, though both parties indicated a preference for Facebook's voluntary removal of content over formal state

[50] "[I]dealistic pronouncements about the redemptive power of democratic politics and democratic constitutionalism have become increasingly difficult to credit. But certain high-level constraints on institutional behavior—and in particular the principles of separation of powers, procedural due process, and public reason—have commanded widespread adherence in democratic societies and have limited arbitrary exercises of official power." JULIE E. COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM 219 (2019) (ebook).

[51] *A Free People in Our Land: Israel as a Parliamentary Democracy*, ISR. MINISTRY OF FOREIGN AFF. (Apr. 1, 2005), https://mfa.gov.il/mfa/aboutisrael/state/pages/a%20free%20people%20in%20our%20land-%20israel%20as%20a%20parliamentary%20democracy.aspx [https://perma.cc/QL3B-ZQHE].

[52] Mark Zuckerberg, A Conversation on Free Expression with Mark Zuckerberg at Georgetown University (Oct. 17, 2019).

[53] Bethan McKernan, *Facebook "Deliberately Targeting" Palestinian Accounts After Meeting with Israeli Government, Rights Groups Say*, INDEPENDENT (Oct. 24, 2016), https://www.independent.co.uk/news/world/middle-east/israel-palestine-facebook-activist-journalist-arrests-censorship-accusations-incitement-a7377776.html [https://perma.cc/AV37-NGND].

action.[54] Tensions came to a head when seven prominent Palestinian journalists found that Facebook suspended their personal accounts. Facebook claimed that the accounts were reported for violating the site's community standards and mistakenly suspended, but Palestinian outlets perceived the incident as related to Israel's recent push to combat incitement. The journalists noted that, in addition to their account suspensions, Facebook removed content from their news pages, including material that was non-political.[55] In response to public outcry, Facebook reinstated the journalists' profiles and apologized for the "error" without offering insight as to how such a mistake occurred.[56] This is significant because the platform's unilateral decision to remove content and suspend user profiles limited the journalists' ability to communicate with the public. Yet, by citing the Community Standards rather than local legal strictures to make these decisions, interested parties had no established means by which to interrogate Facebook's reason for removing the content any further.

Facebook's equivocal response about the journalists' profile suspension raised concerns among Palestinian Facebook users and activists about bias and censorship of their perspectives.[57] Absent evidence to the contrary, Palestinian advocates viewed these events as "a dangerous escalation" of Facebook's approach to content moderation.[58] The platform defines hate speech as "anything that directly attacks people based on what are known as their 'protected characteristics'—race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity, or serious disability or disease."[59] This definition covers hate speech directed at both Palestinian and Israeli individuals; however, some journalists point to enforcement double standards as indicative of Facebook's bias against Palestinian users, claiming that "[c]alls by Israelis for the killing of Palestinians are commonplace on Facebook, and largely remain

[54] *Why Facebook and Google Are Complying with Israel to Delete Certain Content*, FORTUNE (Sept. 12, 2016), http://fortune.com/2016/09/12/facebook-google-israel-social-media/ [https://perma.cc/3QY9-YQ39].

[55] Patel & Levinson-Walsman, *supra* note 30.

[56] Amar Toor, *Facebook Accused of Censoring Palestinian Journalists*, VERGE (Sept. 26, 2016), https://www.theverge.com/2016/9/26/13055862/facebook-israel-palestinian-journalists-censorship [https://perma.cc/Y2YQ-AMD4].

[57] Ylenia Gostoli, *Is Facebook Neutral on Palestine-Israel Conflict?*, AL JAZEERA (Sept. 26, 2016), https://www.aljazeera.com/news/2016/09/facebook-neutral-palestine-israel-conflict-160921115752070.html [https://perma.cc/898Z-YJCC].

[58] 7AMLEH, *supra* note 6, at 1.

[59] Richard Allan, *Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?*, FACEBOOK NEWSROOM (June 27, 2017), https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/ [https://perma.cc/S7LQ-NYPV].

undisturbed."[60] Palestinian users, meanwhile, experience disproportionate levels of content removal and find that violent rhetoric or hateful posts by Israelis against Palestinians frequently go undetected by content moderators.[61]

Facebook's use of the Community Standards in these instances also demonstrates the ramifications of exercising soft power. In the context of regulating inciting and terroristic speech online, the distinction between state and private actors grew increasingly blurry after the Israeli government publicly exerted substantial political pressure on Facebook to remove content pursuant to Facebook's own Community Standards rather than in terms of what is required by Israeli law.[62] In so doing, it raised suspicions that the Community Standards functioned as a veil for state censorship.

[60] Glenn Greenwald, *Facebook Says It Is Deleting Accounts at the Direction of the U.S. and Israeli Governments*, INTERCEPT (Dec. 30, 2017), https://theintercept.com/2017/12/30/facebook-says-it-is-deleting-accounts-at-the-direction-of-the-u-s-and-israeli-governments/ [https://perma.cc/4JGP-3YXQ]; *see also Is Palestinian-Israeli Violence Being Driven by Social Media?,* BBC NEWS (October 22, 2015), https://www.bbc.com/news/world-middle-east-34513693 (stating that inciteful posts by Israelis increased after Israeli-Palestinian violence) [https://perma.cc/5CGM-A9QM]; Ofra Edelman, *Internet Incitement Against Arabs in Israel on the Rise*, HAARETZ (Oct. 13, 2015), https://www.haaretz.com/.premium-anti-arab-internet-incitement-rising-1.5408041 ("Since the latest wave of violence began there has been a sharp rise in the number of statements inciting to violence against Arabs, and in the number of Facebook pages expressing extremist right-wing positions receiving 'likes,' according to experts on monitoring Internet discourse.") [https://perma.cc/P9RG-URKH].

[61] One report found that, in 2018, hate speech against Palestinians was published on social media platforms "every 66 seconds." 7AMLEH, #ASHTAG PALESTINE 2018: AN OVERVIEW OF DIGITAL RIGHTS ABUSES OF PALESTINIANS 16 (2019), https://7amleh.org/wp-content/uploads/2019/03/Hashtag_Palestine_English_digital_pages.pdf [https://perma.cc/X29E-R9KD]. *See also The Index of Racism and Incitement in Israeli Social Media 2018: An Inciting Post Against Palestinians Every 66 Seconds*, 7AMLEH (Mar. 11, 2019), https://7amleh.org/2019/03/11/the-index-of-racism-and-incitement-in-israeli-social-media-2018-an-inciting-post-against-palestinians-every-66-seconds/ [https://perma.cc/C6MV-8YXN]; Ruth Eglash, *An Arab, a Jew and a Facebook Post: How Similar Words Are Treated Differently*, WASH. POST (July 15, 2016), https://www.washingtonpost.com/world/middle_east/an-arab-a-jew-and-a-facebook-post-how-similar-words-are-treated-differently/2016/07/14/e346ef1c-47a4-11e6-8dac-0c6e4accc5b1_story.html?utm_term=.e48623dc1772 [https://perma.cc/NM2M-X5A8].

[62] Toor, *supra* note 7.

Geopolitical indeterminacy tilts in favor of those who already have power and influence,[63] which, in this case, translates to compliance with Israel's definition of incitement. For example, Facebook received a request from the Israeli State Attorney's Office asking it to remove two pages that the Minister of Defense attributed to an unlawful association.[64] While Facebook determined that the pages themselves did not violate the Community Standards, the platform nevertheless decided to restrict access to both pages in Israel based on the request from the Israeli State Attorney's Office.[65] Here, geofencing within the region translates to restrictions on Palestinian access to public discourse within their communities. This case illustrates how Facebook's desire to comply with governmental requests for removal can result in expanding the legal reach and authority of state actors without firm grounding in law. In these circumstances, the content may well have been inciting or dangerous. However, the removed content could have simply been something the Israeli government found disagreeable. Facebook has strong incentives to remove posts when requested to avoid public blame or formal regulation by a government. Consequently, the coordination between government and digital platforms magnifies the power exercised by state actors while insulating it from public scrutiny.

B. Facebook's Opacity Eschews Democratic, Procedural Norms That Establish the Scope of Permissive Forms of Public Expression.

Procedural transparency of government actions serves an important role in democratic society, separate and distinct from substantive evaluations

[63] *See* Cohen, *supra* note 41, at 220. ("Network-and-standard-based legal-institutional arrangements connect protocol and policy directly to one another and eliminate separation between them. Within such arrangements, the point of mandated standardization is exactly to specify the kinds of flows that must, may, and may not travel via the network. The policy is the standard and vice versa. Power over one translates directly into power over the other. Under background conditions of vastly unequal geopolitical power, that equivalence sets up the two interlocking dynamics that produce policy hegemony . . . And policy hegemony is power that may be exercised without regard for the basic, high-level rule-of-law constraints that obtain in more traditional institutional settings."); *See also* Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children*, PROPUBLICA (June 28, 2017), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms ("at least in some instances, the company's hate-speech rules tend[ed] to favor elites and governments over grassroots activists and racial minorities."). [https://perma.cc/A9TJ-RP8B].

[64] *Content Restrictions Based on Local Law*, FACEBOOK TRANSPARENCY (2019), https://transparency.facebook.com/content-restrictions (accessed Nov. 26, 2019) [https://perma.cc/B3JN-XY4Q].

[65] *Id.*

of permissible forms of expression and dissent. Put differently, inquiry into the procedural adequacy of state action is analytically distinguishable from debate over substantive line drawing between protected and unprotected speech. In contrast to Facebook's removal of content and account suspensions, the Israeli government took tangible action against Palestinian poet Dareen Tatour in response to a series of her social media posts by arresting and charging her for incitement of terrorism and support of terrorist organizations.[66] The charges relied on three posts, including a poem about the Knife Intifada, along with references to her opposition to the Israeli occupation. Her indictment contained an interpretation of her poem, specifically including the lines "I will not succumb to the 'peaceful solution' / Never lower my flags / Until I evict them from my land."[67] Although there was fervent dispute around how the poem's language should be interpreted and contextualized throughout the trial, Tatour was convicted and sentenced to several months in prison for inciting violence and supporting a terrorist organization based on her social media posts.[68]

The government's decision to convict Tatour based on her social media posts sparked criticism and scrutiny from international literary groups, advocates, and individuals from around the world.[69] Critics of her conviction connected this case to an ongoing pattern of Israeli arrests of Palestinians for terrorism-related charges based on social media posts; others pointed out a double standard, asserting that Jewish Israeli poets that made similar, explicit calls for violence on social media do not face legal consequences.[70] Despite the many controversial aspects of this case, Tatour's experience highlights the value of transparency as a function of assigning responsibility and accountability, even in circumstances where public expression is punishable. This stands in contrast to instances where limitations on speech and the justification for removing it from the public square remain hidden.

[66] *Offline: Dareen Tartour*, ELECTRONIC FRONTIER FOUND. (2018), https://www.eff.org/offline/dareen-tatour [https://perma.cc/5D3H-3892].

[67] Noa Shpigel, *Israeli Arab Poet Dareen Tatour Gets Five-month Sentence for Incitement on Social Media*, HAARETZ (July 31, 2018), https://www.haaretz.com/israel-news/israel-hands-palestinian-poet-dareen-tatour-five-month-prison-sentence-1.6335232 [https://perma.cc/KU8W-G28W].

[68] *Dareen Tatour: Israeli Arab poet convicted of incitement*, BBC NEWS (May 3, 2018), https://www.bbc.com/news/world-middle-east-43990577 [https://perma.cc/F2E2-U374].

[69] Israel: End judicial proceedings and ensure Dareen Tatour's immediate release, PEN INT'L (July 2, 2018), https://pen-international.org/news/israel-end-judicial-proceedings-and-ensure-dareen-tatours-immediate-release [https://perma.cc/3JF9-SVHH]; *Artists all over the world express their solidarity by works of art based on the poem "Resist,"* FREE DAREEN, https://freedareentatour.org/poems [https://perma.cc/A8MA-3BK7].

[70] Electronic Frontier Found., *supra* note 66.

Importantly, the arrest, trial, and ensuing debate reveal the subjective nature of interpreting words like *incitement* and *terrorism*. While many people disagree about the scope of protection for expressions of dissent speech, hate speech, inciteful speech, and harassment in public discourse, a cornerstone of democracy is protecting and permitting expressions of legitimate dissent (though the line of legitimacy is often drawn in different places). The conclusion that Tatour's poem social media posts fit within the meaning of incitement and support terrorism turns on subjective judgments reflecting a set of cultural norms and priorities. To that end, the inquiry as to whether Tartour's speech appropriately fits within the scope of these charges depends on which narrative one chooses to accept. Moreover, the governing body that decides which expressive content satisfies the definitions of incitement and terrorism simultaneously has authority to determine the boundaries of permissible speech within which one can engage in the public sphere.

Throughout Tatour's case, the process of trial, adjudication, and conviction provided mechanisms requiring transparency around the charges that informed the government's restrictions on speech: prosecutors had to posit charges and support them with corresponding facts.[71] This furnished the public with tools to identify the disputed content, interrogate the legitimacy of the Government's charges, and follow the process of adjudication. With this information, individuals could make judgments about the government's decisions as they related to their own political sensibilities and advocate around the issue. Transparency does not diminish the real and significant consequences that Tatour experienced for posting her poems. There is, however, an ability to trace which parties exercised power and the institutions where accountability, critique, and recourse might be directed. After the trial, advocacy around Tatour's conviction resulted in an early release from her sentence.[72]

The procedural transparency highlighted in Tatour's case stands in contrast to the processes involved in content moderation decisions on Facebook's platform. Facebook's intervention in content removal and profile suspensions displaces conventional sources of transparency and process because its actions exist outside the formal exercise of law. While these decisions may be uncoordinated and ad hoc, the platform functions as a vehicle by which fundamental democratic rights can be exercised or suppressed. Consequently, content moderation decisions implicate important democratic functions and carry with them the capacity to reshape the dynamics of public

[71] This analysis makes no evaluative judgments concerning the government's substantive interpretation and application of law in adjudicating these charges.

[72] *Israel: Poet Dareen Tatour Released from Prison*, PEN INT'L (Sept. 20, 2018), https://pen-international.org/news/israel-poet-dareen-tatour-released-from-prison [https://perma.cc/WN6S-JVQK].

discourse. This creates a new mechanism through which Palestinian and Israeli authorities can vie for influence and an area where Facebook can allow political pressures and other soft influences to color the lens through which it interprets its own Community Standards.

C. Algorithmic Flagging and Artificial Intelligence (AI) Tools Heighten, Rather Than Mitigate, Unequal Levels of Suspicion and Surveillance.

Facebook uses a combination of algorithms and human oversight to moderate activity on its platform. The scale of content posted to Facebook requires the use of algorithms to organize and rank posts on individuals' newsfeeds, and content moderation algorithms are also used to flag and proactively remove problematic content before it is reported as an issue by a user.[73] Facebook uses these algorithms to identify, flag, and remove hate speech quickly and efficiently, yet the efficacy of these algorithms depends on the material used to train the algorithm, along with the scope of the algorithm's application and monitoring of its outputs. In the context of Israeli and Palestinian experiences on Facebook, algorithmic flagging and AI tools heighten—rather than mitigate—geopolitical tensions based on the limitations of machine learning and minimal oversight.

Bias can surface in many ways, and the nuances and challenges of content moderation relevant to Israeli and Palestinian users present difficult questions that algorithms may be fundamentally ill-suited to resolve. Studies on algorithmic fairness, accountability, and transparency suggest that mathematical formulas and algorithmic outputs may not always be appropriate tools to evaluate issues heavily dependent on social or historical context or questions of fairness.[74] In terms of both computer science and cultural sensitivity, a one -size-fits-all approach to content moderation creates a fundamentally flawed process by which to tackle thorny, context-specific issues in contested territories. And, at its core, Facebook's application of its flagging algorithms may not provide adequate sensitivity to minority perspectives or account for cultural nuances specific to contentious dynamics. The constitutive qualities of inciting speech cannot be distilled or determined in a vacuum because they vary based on sensitivities informed by a multitude

[73] FACEBOOK, FACEBOOK'S CIVIL RIGHTS AUDIT PROGRESS REPORT 7–8 (2019), https://fbnewsroomus.files.wordpress.com/2019/06/civilrightaudit_final.pdf [https://perma.cc/C9LE-SJTF]; Nathaniel Gleicher, *Removing Bad Actors from Facebook*, FACEBOOK NEWSROOM (June 26, 2018), https://newsroom.fb.com/news/2018/06/removing-bad-actors-from-facebook/ [https://perma.cc/8TYK-GHMU].

[74] *See generally* Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, *in* PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, TRANSPARENCY, AND ACCOUNTABILITY 2019 59 (forthcoming), https://dl.acm.org/citation.cfm?id=3287598 [https://perma.cc/7RBG-PMVG].

of factors, including history, setting, and circumstance. Thus, universally applied algorithms are likely to produce erroneous outcomes in this area, either missing or mischaracterizing speech in an array of fact-specific circumstances.

Algorithmic bias contributes to the sense of unequal treatment and targeting that Palestinians report to experience on social media. For instance, Israeli law enforcement arrested a Palestinian man based on an error in Facebook's automatic translation program.[75] In that case, a construction worker in the West Bank posted a picture of himself alongside a bulldozer with the caption "يصبحهم" or "'yusbihuhum," which translates as "good morning" in Arabic.[76] Facebook's algorithm translated this caption as saying "hurt them," in English or "attack them," in Hebrew.[77] This translation, coupled with the fact that bulldozers had previously been weaponized as vehicles for hit-and-run terrorist attacks,[78] resulted in the man's arrest for posting a picture to his private Facebook account.[79] Police officers detained the man after receiving notification about the post, though at no point before his arrest did "any Arabic-speaking officer read the actual post."[80] After several hours of questioning, officers realized the mistake and released him.

Here, bias permeates content moderation in a manner that reinforces existing power disparities in three ways. First, the translation algorithm used to regulate and flag content initiated punitive measures before the translation was checked for accuracy. Second, the heightened and suspicious surveilling of Palestinian users on Facebook increased incidents of flagging that can lead to arrest. Put differently, the algorithm serves as a source of confirmation bias for suspicious moderators, rather than a tool to objectively identify and respond to content inciting harm. Third, even after the man was released and

[75] Alex Hern, *Facebook Translates "Good Morning" into "Attack Them," Leading to Arrest*, GUARDIAN (Oct. 24, 2017), https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest [https://perma.cc/3T3X-WCYE].

[76] *Id.*

[77] *Id.*

[78] *Id.*

[79] Yuval Noah Harari, *Why Technology Favors Tyranny*, ATLANTIC (Oct. 2018), https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/ [https://perma.cc/TYF5-VS5L].

[80] *Id.*

Facebook apologized for the translation error,[81] the original post remained removed from the platform after the incident.[82] This demonstrates how biased outcomes, even when proven erroneous, can further marginalize minority voices.

D. Secret Flows of Information Between Private and Public Actors for the Purposes of Surveillance and Arrest Thwart Procedural Oversight and Legal Guardrails that Are Normatively Desirable and Functionally Expected.

Facebook's coordination with law enforcement has the potential to magnify police power in targeted and politically significant ways. In the translation incident discussed above, it is notable that Facebook and Israeli law enforcement's actions leading up to the arrest were nebulously comingled. News coverage on this story stated that the police "were notified"[83] of the post, though it is not clear which actor can be assigned responsibility for what action. It is unclear, for example, whether Facebook flagged the post and reported the content to law enforcement based on the translation, which is a possibility discussed in the Community Standards, or if Israeli law enforcement used its own algorithms and flagging tools that rely on the platform's translation capabilities. All that is publicly known about Facebook's involvement with the arrest discussed above is that the translation algorithm's mistake catalyzed the arrest. These circumstances raise questions concerning both the frequency and freeness by which information flows between Facebook and state actors.

Content management and moderation are essential functions of Facebook's business. Yet, the nature and amount of data sharing or surveillance the platform does on behalf of states frustrates traditional allocations of accountability and trust. Adalah, the Legal Center for Arab Minority Rights in Israel, reported that, in 2016, "82 percent of those arrested for incitement-related offenses were Palestinian citizens, whereas only 18

[81] Gizmodo received the following statement from an engineering manager at Facebook: "Unfortunately, our translation systems made an error last week that misinterpreted what this individual posted. Even though our translations are getting better each day, mistakes like these might happen from time to time and we've taken steps to address this particular issue. We apologize to him and his family for the mistake and the disruption this caused." Sidney Fussell, *Palestinian Man Arrested After Facebook Auto-Translates "Good Morning" as "Attack Them*," GIZMODO (Oct. 23, 2017), https://gizmodo.com/palestinian-man-arrested-after-facebook-auto-translates-1819782902 [https://perma.cc/9QTB-DY67].

[82] Harari, *supra* note 79.

[83] Yotam Berger, *Israel Arrests Palestinian Because Facebook Translated "Good Morning" to "Attack Them,"* HAARETZ (Oct. 22, 2017), https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427 [https://perma.cc/RJY4-Y9A9].

percent were Israeli Jewish citizens."[84] Despite documented concerns from both sides of the conflict about hateful or inciting content circulated on Facebook, the distribution of arrests reveals an imbalance in either platform detection or legal enforcement.[85] Thus, from the perspective of Palestinians, Facebook not only imposes a double standard concerning the application of its own Community Standards, but it may also contribute to the disproportionate number of Palestinian arrests related to user activity on the platform.

Democratic societies broadly recognize that unrestricted government surveillance contravenes individual rights and liberties.[86] As such, the unknown arrangement concerning flows of information between Facebook and Israeli authorities proves significant in two ways: (1) by raising suspicions that Facebook is acting as an extension of law enforcement; and (2) by thwarting oversight or procedural safeguards to protect from oppressive forms of government surveillance. Because government surveillance programs often balance national security priorities with individual civil liberties, procedural safeguards and oversight operate as guardrails to check abusive exercise of authority. Statutes like the General Data Protection Regulation ("GDPR"),[87] in the European Union, and the Foreign Intelligence Surveillance Act ("FISA"),[88] in the United States, illustrate different constructions of procedural and legal protections that strive to balance governments national security pursuits with democratic norms of procedural due process. While

[84] *Adalah Fears Facebook's Online Incitement Deal with Israel Will Selectively Target Palestinian Citizens*, ADALAH: LEGAL CTR. ARAB MINORITY RTS. ISR. (Sept. 11, 2016), https://www.adalah.org/en/content/view/8948 [https://perma.cc/CA7Z-MG24].

[85] Eglash, *supra* note 61.

[86] Cohen, *supra* note 41, at 191–99.

[87] Commission Regulation 2016/679, art. 23, 2016 O.J. (L 119) (providing a national security exception in a manner that "respects the essence of the fundamental rights and freedoms and is a necessary and proportionate measure in a democratic society. . .").

[88] 50 U.S.C. §§ 1801–1885(c) (2018).

these statutes are imperfect and subject to robust debate,[89] they provide some threshold requirements and limitations on government surveillance, along with modes of recourse for individuals whose rights are violated under the law. By way of contrast, suspected informal or secret flows of information between Facebook and Israel frustrate notions of democratic legitimacy and accountability.

Due to these enabling factors, governance on Facebook becomes a proxy battle in which disputing narratives around Israeli and Palestinian activity emerge and collide. Palestinians view Israel's efforts to control the content shared on Facebook as an extension of physical occupation to the digital space;[90] meanwhile, Israelis view Facebook as a platform that magnifies the reach of terrorist actors, who encourage violence against their

[89] Karl Manheim & Lyric Kaplan, *Artificial Intelligence: Risks to Privacy and Democracy*, 21 YALE J. L. & TECH. 106, 168 (2019) ("[C]ontrol, transparency, and accountability are running themes throughout GDPR."); Peter Swire & DeBrae Kennedy-Mayo, *How Both the EU and the U.S. Are "Stricter" Than Each Other for the Privacy of Government Requests for Information*, 66 EMORY L.J. 617, 619–20 (2017) ("Based on our study of both the EU and U.S. systems, we believe there are generally effective rule-of-law protections against excessive law enforcement surveillance in both the U.S. and EU Member States. We therefore conclude that these generally effective safeguards provide a promising basis for MLA reform, even where details of the systems differ and specific safeguards on one side do not have precise counterparts on the other."); Laura K. Donohue, *The Case for Reforming Section 702 of U.S. Foreign Intelligence Surveillance Law*, COUNCIL FOREIGN REL. (June 26, 2017), https://www.cfr.org/report/case-reforming-section-702-us-foreign-intelligence-surveillance-law ("Section 702 is an important tool in the intelligence community's arsenal. But the statute should be amended to bring it within constitutional bounds.") [https://perma.cc/5K77-BATX]; L. Rush Atkinson, *The Fourth Amendment's National Security Exception: Its History and Limits*, 66 VAND. L. REV. 1343, 1396–1405 (2013) (providing an account of the evolution of FISA, intelligence warrants, and warrantless surveillance, as well as a discussion of the procedural and substantive limitations of FISA revealed in litigation).

[90] Press release, 7amleh, #Palestine 2017 Report: Palestinian Online Content Targeted Through Mass Surveillance, Digital Occupation and Biased Content Moderation (April 3, 2018), https://7amleh.org/2018/04/03/press-release-palestine-2017-report-palestinian-online-content-targeted-through-mass-surveillance-digital-occupation-and-biased-content-moderation/ ("[T]he report focuses on the 'digital occupation' of Palestinian social media through the use of algorithms, mass surveillance, and tens of military and Secret Service-affiliated Facebook pages utilized to hinder Palestinian online activism and infringe on freedom of speech.") [https://perma.cc/HF6L-2J2N].

citizens.[91] The informal relationships and vague criteria that influence users' experiences on Facebook fuel suspicion and distrust on both sides of these opposing narratives. While Facebook balks at the notion that it exercises greater leniency with Israeli users compared to their Palestinian counterparts,[92] the platform's amorphous relationship with the government, algorithmic backboxes, and secrecy concerning information flows undermine trust in the platform's credibility in their claims of neutrality. Further, the lack of transparency around content moderation leaves users with few tools to contextualize anecdotal experiences, which creates fodder for speculation without any means to dispel of it.[93] These factors together foster the real and perceived double standards and biases that permeate the platform unchecked.

IV. OPPORTUNITIES: DEFINING, DELINEATING, AND CLARIFYING THE ROLE OF CONTENT MODERATION IN GLOBAL AFFAIRS

Today, Facebook's involvement and entanglement with state governments undermines public trust, weakens accountability, and frustrates traditional protections of democratic norms along with human rights. However, the current construction is neither inevitable nor beyond repair. Facebook can act as a better, more responsible steward in preserving the rights of its users by taking steps to grapple with its role in public life, by delineating its actions from those of state actors, and by adopting a human rights orientation toward content moderation decisions.

[91] Associated Press Jerusalem, *Facebook and Israel to Work to Monitor Posts that Incite Violence*, GUARDIAN (September 12, 2016), https://www.theguardian.com/technology/2016/sep/12/facebook-israel-monitor-posts-incite-violence-social-media, ("Israel has argued that a wave of violence with the Palestinians over the past year has been fueled by incitement, much of it spread on social media sites. It has repeatedly said that Facebook should do more to monitor and control the content…") [https://perma.cc/5AVY-VGYW]; Maayan Jaffe-Hoffman, *Israelis Incite Against Arabs on Social Media Every 66 Seconds–Report*, JERUSALEM POST (May 22, 2019), https://www.jpost.com/Israel-News/Israelis-incite-against-Arabs-on-social-media-every-66-seconds-report-590410 ("The Israeli government has in the past accused Arab citizens of Israel and Palestinians of using social networks to incite terrorism against Israelis.") [https://perma.cc/2XHQ-FN7D].

[92] Ylenia Gostoli, *Palestinians Fight Facebook, YouTube Censorship*, AL JAZEERA (Jan. 20, 2018), https://www.aljazeera.com/news/2018/01/palestinians-fight-facebook-youtube-censorship-180119095053943.html [https://perma.cc/79Z7-G6PH].

[93] Speculation or anecdotal evidence could be dispelled with data to contextualize user experiences, but Facebook has declined to explain content moderation determinations in many instances that seem to tilt political power. *See* Olivia Solon, *Facebook Declines to Say Why It Deletes Certain Political Accounts, But Not Others*, GUARDIAN (Jan. 4, 2018), https://www.theguardian.com/us-news/2018/jan/04/facebook-chechnya-ramzan-kadyrov-political-censorship [https://perma.cc/6PVG-Z6BW].

A. Understanding Facebook's Power As Both Distinct From and In Relation to Government.

First, to understand the power dynamics at play, Facebook's agency and power must be disentangled from those of state actors. Legal scholars posit a number of theories used to untangle the complex power dynamics active between private platforms and state actors. The taxonomy developed by this scholarship creates a structure for understanding intermediaries' power in relation to government and law. Within the existing frameworks, there is not consensus around the distribution or balance of power between Facebook and state actors. At times, scholarly work either understates Facebook's agency as too meek or grants it too noble intentions. For instance, Professor Jack Balkin characterizes this dynamic as state co-option and expansion of regulatory reach through the shadow of private companies: in his view, "companies act as a private bureaucracy that implements the state's speech policies."[94] "Jawboning"[95] describes private companies' capitulation to state authority, where, in the absence of concrete law, platforms create opportunities for shadow exercises of government power. However, this framework oversimplifies the complex and competing power dynamics at play because it discounts the amount of leverage Facebook can use to reach consensus with sovereign nations.

Facebook is not a meek organization outmatched by the pressure of the bully pulpit. Kate Klonick's "New Governors" framework recognizes the elastic influence that digital intermediaries exercise in regulating speech independent of state authorities, asserting that, "platforms are both the architecture for publishing new speech and the architects of the institutional design that governs it."[96] While this framework is instructive in identifying the autonomy platforms exercise over the construction and regulation of speech in alignment with their own normative values and interests, it may offer too generous a read on platforms' internalization of, or commitment to, democratic norms.[97]

[94] Jack M. Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011, 2030 (2018).

[95] Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U. CAL. DAVIS L. REV. 1149, 1177 (2018) ("Nation states have a range of different strategies to exert pressure. They can impose fines or criminal penalties. They can threaten prosecution. Or they can engage in jawboning—urging digital infrastructure operators to do the right thing and block, hinder, or take down content.").

[96] See Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1617 (2018). *See generally* Gillespie, *supra* note 12.

[97] Klonick, *supra* note 96, at 1603 ("They are private, self-regulating entities that are economically and normatively motivated to reflect the democratic culture and free speech expectations of their users.").

It is worth noting that Facebook's business interests center on profit maximization, which requires workable, constructive, and efficient relationships with both individual users and the countries in which Facebook operates.[98] Facebook as a corporate enterprise neither anticipated nor internalized the responsibilities inherent to its business operations and has remained reluctant to take ownership over its role in reshaping political and democratic norms.[99] As such, assuming its commitment to democratic norms as applied to individual users may assume too much of Facebook's established track record of accountability as an independent governing agent. It seems that Facebook's approach to regulating speech more accurately reflects uncoordinated, ad hoc decisions to advance its self-interest at a given moment in time, rather than a systematic, comprehensive approach to governance that accounts for the consequences flowing from its decisions.[100] While state regulation traditionally defined the scope of protections and permissibility of public expression, state authority runs up against—rather than above or below—that of platform intermediaries.

Facebook and its peer platforms are best understood as operating with power equivalent to that of state actors. Indeed, these private entities possess tremendous amount of economic and social capital equivalent to or surpassing that of discrete state governments. In fact, private platform companies perceive themselves as "on par with, not subordinate to, governments, including those governments that attempt to regulate them."[101] This is evident in the ways that platform companies like Facebook increasingly posture in a manner reflecting diplomatic relations between sovereign nations, rather than viewing themselves as answerable to any one jurisdiction. Professor Julie Cohen points to how these legal and institutional dynamics have shifted, pointing out that, "Facebook's privacy team travels the world meeting with government officials to determine how best to satisfy their concerns while

[98] *See* Julie E. Cohen, *Information Platforms and the Law*, 2 GEO. L. TECH. REV. 191, 192–193 (2018) ("[P]latform firms are also discrete legal entities with interests and agendas of their own. Platform firms also rely on law and legal institutions to advance their own self-interested goals.").

[99] Sheera Frenkel et al., *Delay, Deny and Deflect: How Facebook's Leaders Fought Through Crisis*, N.Y. TIMES (Nov. 14, 2018), https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html?module=inline [https://perma.cc/768C-C5LN]; Abby Ohlheiser, *Mark Zuckerberg Denies That Fake News on Facebook Influenced the Elections*, WASH. POST (Nov. 11, 2016), https://www.washingtonpost.com/news/the-intersect/wp/2016/11/11/mark-zuckerberg-denies-that-fake-news-on-facebook-influenced-the-elections/ [https://perma.cc/3JP3-NY7B].

[100] *See* Daphne Keller, *Facebook Restricts Free Speech by Popular Demand*, ATLANTIC (September 22, 2019), https://www.theatlantic.com/ideas/archive/2019/09/facebook-restricts-free-speech-popular-demand/598462/ [https://perma.cc/RGQ8-SDY2].

[101] Kristen E. Eichensehr, *Digital Switzerlands*, 167 U. PA. L. REV. 665, 665 (2019).

continuing to advance Facebook's own interests, much as a secretary of state and his or her staff might do."[102] Building on this notion, the agreements and coordinated efforts platforms reach with governments represent a new balance in the exchange of power, which carries significance for those governed by these two regimes.

Michael Birnhack and Niva Elkin-Koren describe these secret flows of information between state and platform intermediaries as "the invisible handshake," characterized by "collaboration between law enforcement agencies and the private sector, beyond the reach of judicial review and away from the critical eye of public opinion."[103] This description explains the way that power has shifted from traditional accountability guardrails to those with authority. Facebook's adoption of policies and practices that appeal to a particular state's sensibilities exports certain value judgments and normative standards across borders. Still, the platform has struggled to reckon with the implications these choices carry for its users in context-specific scenarios, particularly in regard to individuals whose interests are not adequately represented by the state government engaged in coordination. Platform governance represents negotiations between lawmaking entities and private actors to advance some mutually beneficial end. Problems arise where this dynamic excludes consideration of the individual interests at stake—namely, those of the platform's users—and leaves minority groups and marginalized communities particularly at risk.

B. Transparency Can Improve Facebook's Accountability to Users by Creating Opportunities to Clarify and Correct.

Regardless of whether Facebook is justified in its cooperation with the government or in its interpretation of its Community Standards, the platform's lack of transparency leads to confusion and the erosion of user trust. Facebook cannot both posture as a neutral arbiter and simultaneously agree to cooperate with one side of a disputing narrative without considering the broader, significant implications of its actions. Israelis and Palestinians agree that dangerous activity happens on Facebook and that affirmative steps are required to counter the proliferation of social media-incited violence. Facebook must take responsibility for delineating its actions from those of state actors to avoid its furtherance as a tool in geopolitical conflict. This requires Facebook to take dramatic steps in transparency, so that users are empowered to interrogate suspicions of bias and hold the platform accountable when it supports veiled prejudice or disparate treatment of vulnerable groups.

[102] *See* Cohen, *supra* note 50, at 236.

[103] Michael D. Birnhack & Niva Elkin-Koren, *The Invisible Handshake: The Reemergence of the State in the Digital Environment*, 8 VA. J. L. & TECH. 6, ¶ 146 (2003).

Transparency is the first step toward forging pathways for accountability. Discussion around the platform's opacity in its content moderation practices indicates that "the biggest threat this private system of governance poses to democratic culture is the loss of a fair opportunity to participate, which is compounded by the system's lack of direct accountability to its users."[104] When Facebook uses its Community Standards as a means by which to enforce state preferences, and when it supplants procedural requirements to enforce the law, the platform's actions circumvent public accountability and skew public discourse in an anti-democratic fashion. Until recently, Facebook worked diligently to avoid formal regulation by pursuing informal, flexible agreements that facilitated regulatory capture and minimized the legal obligations that could have diminished its agency.[105] The resultant, ongoing lack of transparency may be used to serve the platform's business interests and state interests in accommodating ways at the expense of user accountability and traditional democratic restraints. Moreover, both individual users and state governments now struggle to locate tools by which intermediaries can be held accountable for their influence on the exercise of fundamental rights or denial of personal liberties.

Traditional mechanisms to advance corporate accountability do not stretch well when applied to platform governance and transnational rights. Some individuals have tried to use existing legal avenues available in the United States to seek recourse for harms that occurred in Israeli and Palestinian territories but had tracible connections to conduct on Facebook.[106] These legal claims have not succeeded in court or created the friction necessary to incentivize different conduct on the part of Facebook. In the U.S., these tools conventionally include federal laws like the Alien Tort Claims Act (ATCA) and antiterrorism statutes, along with commonly accepted standards for Corporate Social Responsibility (CSR).[107] While ATCA and antiterrorism statutes provide a hook for corporate liability in some circumstances, both failed to gain traction with courts in the context of social media companies

[104] Klonick, *supra* note 96, at 1603.

[105] *See e.g.*, Laura Kayali, *Facebook Embraces Regulation—Reluctantly*, POLITICO (Jan. 29, 2019), https://www.politico.com/story/2019/01/29/facebook-reluctant-regulation-1130090 [https://perma.cc/J672-8E7N].

[106] *See infra* note 108.

[107] Practical Law Corporate & Securities, *Corporate Governance Standards: Overview*, THOMSON REUTERS PRACTICAL LAW (accessed May 12, 2019); *Expert Q&A on Trends in Corporate Social Responsibility*, THOMSON REUTERS PRACTICAL LAW (Sept. 19, 2013).

providing governments with the tools that facilitated the alleged harms.[108] Importantly, courts have expressly held that Facebook's enforcement of its Community Standards "and implementation of networking algorithms for [the] benefit of all users" are protected from liability under the Communications Decency Act.[109] Finally, CSR norms provide instructive frameworks and guidelines for businesses operating in multinational settings. However, they neither create binding legal obligations nor necessarily speak to the business model of information platforms, which implicate a unique combination of business, government, and individual interests.

Consequently, accountability in the current landscape requires Facebook to take radical and proactive steps towards accountability and to move toward recognizing affirmative obligations and duties to its users. For content moderation practices, this calls for auditing its algorithms, collecting data on its internal practices, and providing content moderators with the tools, training, and support needed to perform their jobs. Facebook should not only collect information, but also make it publicly available. Presently, Facebook's transparency reports cover only a fraction of its content moderation activity.[110] Meaningful accountability to users requires empowering them with the information necessary to discern how their rights are accounted for when using the platform. Facebook has taken some initial, promising steps by publishing civil rights audit reports and releasing more transparency data than previously available.[111] Still, clearer protocols and accountability tools are needed to address concerns related to user experiences, particularly in conflicted

[108] *See generally* Force v. Facebook, 304 F. Supp. 3d 315 (E.D.N.Y. 2018) (dismissing a case that alleged Facebook materially supported terrorist activity by allowing Hamas and its supporters to use the social media platform to further their aims and finding that Facebook's "maintenance of [an] internet platform" and algorithms applied to all users and were protected by the Communications Decency Act ("CDA")); Cohen v. Facebook, 252 F. Supp. 3d 140 (E.D.N.Y. 2017) (finding that fear of future terrorist attacks based on online content was too speculative to satisfy standing requirements and that Facebook content moderation practices were covered by CDA immunity); Corrie v. Caterpillar, 403 F. Supp. 2d 1019 (W.D. Wash. 2005) (holding that company that sold the IDF bulldozers used to demolish Palestinian homes was not liable for aiding or abetting human rights violations).

[109] *Force*, 304 F. Supp. 3d at 331–332.

[110] Facebook reports on content restrictions based on local law by country, with its report noting which removals were mandated by local law, as opposed to being mere violations of the platform's community standards. Facebook does not identify content restrictions and removals by geographic region based on violation of the community standards. *See Content Restrictions,* FACEBOOK (2019), https://transparency.facebook.com/content-restrictions (accessed Oct. 27, 2019) [https://perma.cc/S4UM-9QGV].

[111] Facebook's civil rights audit began in 2018. Up until that point, this information was not publicly available. *See* FACEBOOK, FACEBOOK'S CIVIL RIGHTS AUDIT–PROGRESS REPORT (June 30, 2019), https://fbnewsroomus.files.wordpress.com/2019/06/civilrightaudit_final.pdf [https://perma.cc/C7CK-74QH].

territories. To date, the platform has not provided concrete information as to how its content moderation practices change to account for competing narratives in geopolitically contested territories, nor how it might improve AI to account for these complex and nuanced issues.[112] Providing robust data to users would serve as a promising initial step to address Facebook's content moderation problems. While self-monitoring and self-governance regimes have significant limitations, they are necessary, albeit not sufficient, steps to confront this problem. Transparency without subsequent, thoughtful action cannot ameliorate the challenges of content moderation on global platforms.

C. A Human Rights Approach to Content Moderation Creates Space for Competing Narratives, Experiences, and Truths.

As a threshold matter, Facebook must recognize its responsibility as an influential force in international and geopolitical affairs, and then grapple with what those responsibilities demand in terms of conduct.[113] Its current posture of neutrality undermines Facebook's ability to grapple with these questions. While Facebook vehemently defends its approach to content moderation as a neutral and objective process, this construction is a poor fit to tackle the thorny issues inherent to content moderation decisions.[114] Here, it seems that Facebook confuses neutrality with its underlying normative values.[115] Despite the evidence of bias that has surfaced in recent years, Facebook representatives still deny that Israeli and Palestinian users experience differential treatment on the platform because Facebook does not

[112] *Understanding Social Media and Conflict*, FACEBOOK NEWSROOM (June 20, 2019), https://newsroom.fb.com/news/2019/06/social-media-and-conflict/ [https://perma.cc/HE8R-JTB4].

[113] *See* Frenkel et al., *supra* note 99. In the wake of discovering Facebook's inadvertent facilitation of election meddling, humanitarian crises, and the proliferation of political propaganda, the platform remained reluctant to publicly engage on issues of responsibility and deflected criticism: "When researchers and activists in Myanmar, India, Germany and elsewhere warned that Facebook had become an instrument of government propaganda and ethnic cleansing, the company largely ignored them." *Id.*

[114] *See generally* Anupam Chander & Vivek Krishnamurthy, *The Myth of Platform Neutrality*, 2 GEO. L. TECH. REV. 400 (2018).

[115] In a recent speech at Georgetown University, Mark Zuckerberg outlined the values he believes Facebook stands for, stating, "I want to ensure the values of voice and free expression are enshrined deeply into how this company is governed." These normative values, in the abstract, do not demand a posture of neutrality. However, the instrumentalization of these normative values has translated into a posture of neutrality. Zuckerberg, *supra* note 52.

take a position on the issue.[116] This statement reveals the problematic logic at work, as staking a position and suffusing bias are not the same thing. As Professor Kristen Eichensehr explains, the posture of neutrality "carries with it a risk of undue passivity, tending toward complicity."[117] Here, Facebook seems to use neutrality to shield itself from criticism regarding non-neutral outcomes. Facebook uses its posture of neutrality to deny its active influence on the nature of discourse. In so doing, it fails to acknowledge that "[t]he changes wrought by new technology do not benefit everyone equally."[118] As such, neutrality is not a productive position from which to unpack involvement in unitended outcomes; rather, this posture is self-defeating.

Additionally, Facebook's Community Standards already stake non-neutral positions on matters like hate speech, terrorism, and incitement.[119] At a fundamental level, limitations or rejection of certain kinds of expression reflect a balancing of values like civility, respect, and pluralism with individual rights to expression, participation, and access. While there are many legitimate positions that balance competing interests, grappling with these issues requires baseline recognition of the normative values underlying certain preferences.[120] Thus, Facebook should abandon its posture of neutrality and reorient its content moderation practices toward the protection of human rights.

Facebook should not be expected to resolve the Israeli-Palestinian conflict, but the company should be expected to moderate content in a manner consistent with the depth of trust and reliance it encourages users to place on the platform. This requires Facebook to grapple with competing truths and communities' concerns that their speech, autonomy, and inherent dignity are

[116] Ylenia Gostoli, *Palestinians Fight Facebook, YouTube Censorship*, AL JAZEERA (Jan. 20, 2018), https://www.aljazeera.com/news/2018/01/palestinians-fight-facebook-youtube-censorship-180119095053943.html ("[Facebook] engage[s] all over the world with governments, NGOs, academics. It doesn't mean we take a position.") [https://perma.cc/US73-QELC].

[117] Eichensehr, *supra* note 101, at 36.

[118] Chander & Krishnamurthy*, supra* note 113, at 403; *see also* Zeynep Tufekci, *The Real Bias Built in at Facebook*, N.Y. TIMES (May 19, 2016), https://www.nytimes.com/2016/05/19/opinion/the-real-bias-built-in-at-facebook.html?rref=collection%2Fcolumn%2Fzeynep-tufekci&action=click&contentCollection=opinion&region=stream&module=stream_unit&version=latest&contentPlacement=13&pgtype=collection ("Software giants would like us to believe their algorithms are objective and neutral, so they can avoid responsibility for their enormous power as gatekeepers while maintaining as large an audience as possible.") [https://perma.cc/RH7W-D9LM].

[119] Chander & Krishnamurthy*, supra* note 113, at 405–07.

[120] *Id.* at 405 ("[P]latforms are explicitly non-neutral with respect to certain issues specified in their community guidelines. These guidelines do not simply recapitulate the law, but rather set out a series of normative commitments.").

vulnerable. While Facebook is currently in the process of developing an appeals process through its new Oversight Board,[121] the bandwidth, priorities, agency, composition, and oversight of the Board will influence the degree to which it is able to address context-specific human rights concerns.[122] In particular, Facebook should seek input and participation from human rights stakeholders to ensure that it facilitates fair review for individuals who may be marginalized or disproportionately discriminated against in the existing content moderation model.

As a matter related to but distinct from censorship concerns, Facebook must also grapple with respecting competing truths. Israelis and Palestinians have vastly different narratives about the geopolitical conflict and its motivations, history, and lived experiences. Instead of deferring to individual moderators or private self-governance mechanisms to decide the validity of opposing narratives, Facebook can pull from existing human rights guidelines and resources to inform content moderation decisions. The Global Network Initiative provides an instructive framework for the operationalization of platform governance in a manner that protects individual freedoms of expression and privacy.[123] Professors Deirdre Mulligan and Daniel Griffin explore the relationship between the respect for truth and Google search results, finding that a human rights framework can help platforms "avoid the slippery slope of arbitrating truthfulness" while still engaging in content moderation.[124] They draw from existing human rights reports, finding that, for content intermediaries, respect for the truth involves:

> due diligence to identify, prevent, evaluate, mitigate and account for risks to the freedom of expression and privacy rights that are implicated by the company's products, services, activities and operations to assess actual and potential human rights impacts on individuals, integrating and acting upon the

[121] Kate Cox, *Facebook Plans Launch of its Own "Supreme Court" for Handling Takedown Appeals*, ARS TECHNICA (Sept. 18, 2019), https://arstechnica.com/tech-policy/2019/09/facebook-plans-launch-of-its-own-supreme-court-for-handling-takedown-appeals/ [https://perma.cc/35ZX-25TQ].

[122] Concerns about the establishment of a Facebook Supreme Court are significant, but beyond the scope of this analysis.

[123] *See The GNI Principles*, GLOBAL NETWORK INITIATIVE (2019), https://globalnetworkinitiative.org/gni-principles/ ("The duty of governments to respect, protect, promote and fulfill human rights is the foundation of this human rights framework. That duty includes ensuring that national laws, regulations and policies are consistent with international human rights laws and standards on freedom of expression and privacy.") [https://perma.cc/V9G7-8D34].

[124] Deirdre K. Mulligan & Daniel S. Griffin, *Rescripting Search to Respect the Right to Truth*, 2 GEO. L. TECH. REV. 557, 574 (2018).

> findings, tracking responses, and communicating how impacts are addressed . . . and where due diligence identifies circumstances when freedom of expression and privacy may be jeopardized or advanced[,] . . . employ human rights impact assessments and develop effective risk mitigation strategies as appropriate. . . .[125]

While these steps may leave stakeholders on both sides unhappy with particular outcomes, by adopting an approach grounded in human rights, rather than in neutrality, Facebook can better account for each user's innate dignity and rights while navigating the challenges inherent to managing a platform in this setting.

V. CONCLUSION

As a private enterprise providing a global platform for public expression, Facebook plays a significant but convoluted role in modern political life. Over the past fifteen years, the platform has changed the way political power can be organized and exercised across borders. In its current construction, Facebook's enforcement of Community Standards and business practices function as a vector through which political power can be exerted, consolidated, and restricted in nontransparent ways. Facebook's entanglement in the Israeli-Palestinian conflict demonstrates the significant implications of platform governance along with its potential to reconfigure modes of democratic accountability. In navigating this landscape in a manner consistent with its expressed values, Facebook must grapple with the history and context of the geographic regions in which it operates. By adopting transparency, accountability, and human rights principles in its approach to content moderation, Facebook can disentangle its actions from those of state actors and rebuild trust and credibility with its users.

[125] *Id*. at 575–576 (internal quotations omitted).