ARTICLE

ONLINE RACIALIZATION AND THE MYTH OF COLORBLIND CONTENT POLICY

ÁNGEL DÍAZ[*]

ABSTRACT

This Article presents a critical analysis of social media content moderation, arguing colorblind policies obscure and legitimize systems of white supremacy. Through facially neutral content policies, social media platforms conceal deliberate choices that align racial benefits and burdens with corporate interests. These choices connect the profitability of racism, the regulatory benefit of protecting politicians who trade in bigotry, and the racial biases that inform how platforms conceptualize the harms of online speech. The resulting content policies reinforce a hierarchical structure that upholds the dominant social, political, and economic advantages attendant to whiteness.

As the primary document governing millions of daily decisions regarding online speech, content policy has the unprecedented ability to shape global norms. Although social media companies purport to treat all groups equally, colorblind content policy protects white bigotry while suppressing antiracist and anticolonial resistance. Dominant racial groups are granted extensive latitude of expression, encompassing everything from racial dog whistles to explicitly racist harassment campaigns. Under the guise of protecting humor or political debate, social media companies foster white vigilantism and authoritarian incitement. In contrast, communities of color are policed as dangerous, violent, and uncivilized. Colorblind hate speech rules restrict the ability of marginalized groups to explicitly denounce white racism, while racialized enforcement of violent extremism policy broadly suppresses political debate and sacrifices everything, from satire to journalism, in the name of public safety.

[*] Visiting Assistant Professor, University of Southern California Gould School of Law. I am deeply grateful for thoughtful feedback and comments from LaToya Baldwin Clark, Jordan Barry, Guy-Uriel Charles, David Cruz, Jasmine Harris, James Hicks, Felipe Jiménez, Mugambi Jouet, Emmanuel Mauleón, Erin Miller, Jeesoo Nam, Marcela Prieto, Daria Roithmayr, Daniel Sokol, Aaron Tang, and Franita Tolson. This Article benefitted immensely from workshops at the USC Gould School of Law, University of San Diego School of Law, the LatCrit 2023 Biennial Conference, and the Workshop on Radical Approaches to Platform Governance at University College Dublin. Finally, I thank the *Boston University Law Review* editors, especially Win Gustin, for their terrific work on this Article. All mistakes are mine.

Understanding the asymmetry inherent in content moderation requires an engagement with the history and logics of racism. The illusion of colorblind content policy reinforces racial hierarchies by making them appear natural and inevitable. This Article challenges the discriminatory structure of colorblind content policy and advocates for an alternate approach that incorporates the race-conscious moderation necessary to foster full participation in modern society.

CONTENTS

> *You are hierarchical . . . It's a terrestrial characteristic. When human intelligence served it instead of guiding it, when human intelligence did not even acknowledge it as a problem, but took pride in it or did not notice it at all . . . That was like ignoring cancer.*
>
> —Octavia E. Butler[1]
>
> *We've made the policy decision that we don't think that we should be in the business of assessing which group has been disadvantaged or oppressed.*
>
> —Mark Zuckerberg[2]

INTRODUCTION

Social media platforms' global reach and pervasiveness make them essential avenues for communication. Channeled through words, photos, and videos, these platforms provide valuable insights into the complex multitude of human experience. Joy, fear, desire, insecurity, and outrage all find waiting outlets that can transform these emotions into linguistic expression, political education, and newfound community. At the same time, these platforms present an equal opportunity to misinform, incite violence, and scale efforts to subordinate.

In September 2022, communities around the world took to social media to process the death of Queen Elizabeth II. Where some expressed grief or shared loving tributes, others crafted memes making light of her death or posted in solidarity with the victims of British colonization.[3] Among these voices was Uju Anya, a Carnegie Mellon Associate Professor of Second Language Acquisition. She tweeted,[4] "I heard the chief monarch of a thieving raping genocidal empire is finally dying. May her pain be excruciating."[5]

In a follow-up tweet, Professor Anya said she expressed no sympathy for Queen Elizabeth II, given her supervisory role in a "genocide that massacred and

[1] OCTAVIA BUTLER, DAWN 39-40 (1988).

[2] Casey Newton, *Why You Can't Say 'Men Are Trash' on Facebook*, VERGE (Oct. 3, 2019), https://www.theverge.com/interface/2019/10/3/20895119/facebook-men-are-trash-hate-speech-zuckerberg-leaked-audio [https://perma.cc/C49K-DNJD].

[3] *See* Christopher Rhodes, *Black, Brown, and Irish Twitter Show No Mercy for Queen Elizabeth II*, YAHOO! NEWS (Sept. 9, 2022), https://www.yahoo.com/video/black-brown-irish-twitter-show-195606007.html [https://perma.cc/PFL8-CCS7].

[4] This Article uses the name "Twitter" (and the act of posting to the platform as "tweeting") when referring to the company that has recently rebranded to "X." Because most of my analysis was conducted prior to rebranding, I retain references to Twitter both above and below the line for readability.

[5] *See* Uju Anya (@UjuAnya), TWITTER (Sept. 8, 2022, 1:12 PM), https://web.archive.org/web/20220908160348/https://twitter.com/UjuAnya/status/1567863337991512064; *see also* Sam Biddle, *Twitter Censored Professor's Post for "Abusive Behaviour" Toward the Queen*, INTERCEPT (Sept. 9, 2022, 6:17 PM), https://theintercept.com/2022/09/09/queen-dead-twitter-censor-abuse-uju-anya/ [https://perma.cc/3FNR-5PYL].

displaced half my family."[6] As the post began to spread, it was reposted by Amazon CEO Jeff Bezos, who commented, "This is someone supposedly working to make the world better? I don't think so. Wow."[7] Soon after her initial post, Twitter removed Professor Anya's tweet.[8]

A company spokesperson later told a journalist that the post was removed for violating Twitter's rules against abusive behavior.[9] At the time, Twitter prohibited "targeted harassment of someone" or actions that "incite other people to do so."[10] Specifically, the company purportedly took a zero-tolerance approach for any post that "wishes, hopes, promotes, incites, or expresses a desire for death, serious bodily harm or serious disease against an individual or group of people."[11] Twitter's reasoning behind the policy was that harassment limits people's ability to feel free to participate in public life and can lead to physical and emotional harm.[12] The company's decision elided any analysis as to whether a relatively obscure professor's tweet could harm or silence Queen Elizabeth II, and whether there was a countervailing public interest in protecting harsh criticism of a global political leader.[13]

A few days after the Queen's death, Stephen Miller, former senior advisor to President Trump and founder of America First Legal, tweeted about the threats he perceived to the British monarchy's future legitimacy:

> Key to monarchy is its mystery. Key to its mystery is that monarchs descend from an ancient line of fabled kings & queens. Though it may not be apparent now, a longterm concern for UK monarchy will be if, due to marriages, future monarchs have same family trees as their subjects.[14]

[6] Uju Anya (@UjuAnya), TWITTER (Sept. 8, 2022, 1:51 PM), https://twitter.com/UjuAnya/status/1567933661114429441; *see also* Ariel Zilber, *Carnegie Mellon Silent on Whether It Will Punish Uju Anya over Queen Tweets*, N.Y. POST (Sept. 9, 2022, 2:53 PM), https://nypost.com/2022/09/09/carnegie-mellon-mum-on-discipline-for-uju-anya-over-queen-tweets/ [https://perma.cc/M7YR-NFXM].

[7] Jeff Bezos (@JeffBezos), TWITTER (Sept. 8, 2022, 12:51 PM), https://twitter.com/JeffBezos/status/1567918581614247937.

[8] *See* Anya, *supra* note 5; *see also* Zilber, *supra* note 6.

[9] Biddle, *supra* note 5 (discussing Twitter's response to controversy surrounding Anya's tweet).

[10] *Abusive Behavior*, TWITTER (Sept. 8, 2022), https://web.archive.org/web/20220908155837/https://help.twitter.com/en/rules-and-policies/abusive-behavior [hereinafter *Abusive Behavior*].

[11] *Id.*

[12] *Id.*

[13] By comparison, the outrage directed at Professor Anya following publicity around her tweet led to organized calls for Carnegie Melon to terminate her employment. *See, e.g.*, Emma Folts, *CMU's Answer to Prof's Tweet, Wishing 'Excruciating' Pain for Queen, Stirs Campus Backlash*, PUB. SOURCE (Sept. 12, 2022), https://www.publicsource.org/uju-anya-carnegie-mellon-cmu-professor-queen-elizabeth-tweet-petitions/ [https://perma.cc/QR9S-WHS5].

[14] *See* Stephen Miller (@StephenM), TWITTER (Sept. 12, 2022, 8:48 AM), https://twitter.com/stephenm/status/1569321956725366786.

Long criticized for his white nationalist views,[15] Miller's affinity for white replacement theory provides important context for his argument. In relevant part, white replacement theory views miscegenation and nonwhite immigration as existential threats to white people.[16] Prior to Queen Elizabeth II's death, there was ongoing controversy about her grandson's marriage to a Black woman and the royal family's concern about "how dark" their son would be.[17] In context, it is clear that Miller's tweet was not an abstract musing about what sustains magical bloodlines, but a longstanding racist trope: race-mixing stains whiteness. However, because the tweet never mentioned a race or an individual, it did not trip up Twitter's rules against harassment or hate speech.[18]

The differing responses to Miller and Anya's tweets reflect an enduring double standard in social media content moderation. This Article argues that understanding this asymmetry requires engaging with the logics of racism. Content moderation connects the profitability of white racism, the regulatory benefits of protecting politicians who trade in bigotry, and the racial biases that inform how platforms conceptualize the harms of online speech. The result is a system that guards the social, political, and economic advantages attendant to whiteness. This system is operationalized through content policy,[19] making these

[15] *See, e.g.*, Michael Edison Hayden, *Stephen Miller's Affinity for White Nationalism Revealed in Leaked Emails*, S. POVERTY L. CTR. (Nov. 12, 2019), https://www.splcenter.org/hatewatch/2019/11/12/stephen-millers-affinity-white-nationalism-revealed-leaked-emails [https://perma.cc/33L4-N345]; Abbey Marshall, *Democratic Leaders Call on Stephen Miller To Resign amid Claims He Pushed White Nationalist Beliefs*, POLITICO (Nov. 14, 2019, 11:39 AM), https://www.politico.com/news/2019/11/14/stephen-miller-resign-white-nationalism-070885 [https://perma.cc/HY9E-YLST] (detailing Democratic response to Miller's comments); Amanda Holpuch, *Stephen Miller: The White Nationalist at the Heart of Trump's White House*, GUARDIAN (Nov. 24, 2019, 9:21 AM), https://www.theguardian.com/us-news/2019/nov/24/stephen-miller-white-nationalist-trump-immigration-guru [https://perma.cc/CLA6-2JCB] (noting controversy surrounding Miller's immigration policy). *See generally* JEAN GUERRERO, HATEMONGER: STEPHEN MILLER, DONALD TRUMP, AND THE WHITE NATIONALIST AGENDA (2020) (tracing evolution of Stephen Miller's white nationalist ideas over his life).

[16] *See* Jason Wilson & Aaron Flanagan, *The Racist 'Great Replacement' Conspiracy Theory Explained*, S. POVERTY L. CTR. (May 17, 2022), https://www.splcenter.org/hatewatch/2022/05/17/racist-great-replacement-conspiracy-theory-explained [https://perma.cc/VJ7C-E29A] (discussing history of white replacement fearmongering).

[17] Michelle Tauber, *Meghan Markle Says There Were 'Conversations' About 'How Dark' Archie's Skin Color Would Be*, PEOPLE (Mar. 7, 2021, 9:01 PM), https://people.com/royals/meghan-markle-oprah-interview-conversations-how-dark-archie-skin-color-would-be/ [https://perma.cc/WCY6-3327].

[18] *See Abusive Behavior*, *supra* note 10.

[19] This Article uses the term "content policy" as an umbrella term for the rules and policies platforms deploy to moderate content and behavior on the platform. Twitter refers to them as "Rules and Policies," *Rules and Policies*, TWITTER, https://help.twitter.com/en/rules-and-policies [https://perma.cc/R7F7-NRA9] (last visited Nov. 9, 2023), Meta calls them "Community Standards," *Facebook Community Standards*, META,

policies elemental to the maintenance of white supremacy and an essential site of inquiry for critical scholars.

Once an afterthought for social media companies, content policy is now the main document driving millions of decisions over how people can express themselves online. Content policies are cited in press statements,[20] used to parry regulatory oversight during congressional questioning,[21] and cited in Supreme Court briefs.[22] Social media content policy is also the central focus of discussion and negotiation with governments and civil society, and includes the document Meta's Oversight Board uses to assess the company's removal decisions.[23] Content policy is the closest document to a constitution for the private system of content moderation, but it can be rewritten, ignored, or set aside at will.

More than a mirror for offline bigotry, social media companies shape and foster racial hierarchies through the drafting and enforcement of their policies. These choices determine whose speech will be restricted and whose will be protected, which posts will receive prominent distribution, and which will remain buried in obscurity. Through millions of daily enforcement decisions, social media companies have an unprecedented ability to shape global speech norms. While companies claim their policies apply equally to everyone, this Article argues colorblind content moderation is a racialized system that doles out a measured hand for the powerful and an iron fist for the marginalized.

Under this system, social media companies court, foster, and protect white racism. By requiring explicit racial animus or undeniable calls to violence before company intervention, content policy largely shields the vast arsenal of attacks available to white voices who trade in the language of coded messages and dog whistles. Showcasing racism cloaked as edgy humor or political debate fosters white supremacist ideology, leaving platforms wrong-footed when content boils over into white vigilantism and authoritarian incitement. Conversely, communities of color are policed as violent, suspicious, and uncivilized. This racialized gaze results in policies restricting their ability to organize politically,

https://transparency.fb.com/policies/community-standards [https://perma.cc/T2KM-247D] (last visited Nov. 9, 2023), and YouTube refers to them as "Community Guidelines," *Community Guidelines*, YOUTUBE, https://www.youtube.com/howyoutubeworks/policies/community-guidelines/ [https://perma.cc/6CKP-FS7V] (last visited Nov. 9, 2023).

[20] *See, e.g.*, Biddle, *supra* note 5 (relaying Twitter's statement referencing content policy as justification for deleting Professor Anya's tweet).

[21] *See Breaking the News: Censorship, Suppression, and the 2020 Election: Hearing Before the S. Judiciary Comm.*, 116th Cong. 5 (2020) (statement of Jack Dorsey, Chief Executive Officer, Twitter).

[22] *See, e.g.*, Conditional Petition for a Writ of Certiorari at 8, Twitter, Inc. v. Taamneh, 143 S. Ct. 762 (2022) (No. 21-1496) (citing violation and evasion of content policy to argue Twitter cannot be held liable for posts of ISIS adherents).

[23] *See* FACEBOOK, OVERSIGHT BOARD CHARTER 3 (Sept. 2019), https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf [https://perma.cc/7UR5-KGWD] (granting board authority to interpret Facebook content policy and review individual content moderation decisions). The Oversight Board also has jurisdiction to offer nonbinding policy recommendations to improve the company's moderation systems. *Id.* at 4.

denounce racism, or simply build community with one another. Colorblind hate speech rules restrict the ability of marginalized communities to attack white racism or directly speak about their experiences under white supremacy. Meanwhile, racialized enforcement of violent extremism policy broadly suppresses political debate and sacrifices everything from satire to journalism in the name of public safety.

Colorblind content policies obscure and legitimate a racially hierarchical system by making it appear natural and inevitable. At times, these decisions reflect a desire to foster a more favorable regulatory environment, as many drivers of racial hatred also have political power. At other times, they reveal a limited ability to truly understand racism's nuances and threats, viewing white supremacy as an extreme outlier instead of an organizing principle upon which American society is structured. This approach is further solidified by racism's profitability, as much of what draws people (and attendant advertising revenue) to platforms are popular figures who regularly employ bigotry.

The Article proceeds in four parts. First, this Article provides a theoretical overview of my approach to analyzing online racial stratification. Building on a multidisciplinary approach grounded in critical race theory, I provide foundational definitions for race and racialization, and how whiteness and white supremacy shape social media. These definitions center the cultural, economic, and political dimensions of white supremacy, as well as the role of content policy in "moving whiteness from privileged identity to a vested interest."[24]

Second, this Article traces the evolution of social media content policy, explaining how key actors strategically avoid engaging with the realities of racism. This disengagement has two effects. On the one hand, it explains why platforms are consistently caught wrong-footed in attempts to moderate discourse rife with racial bigotry. On the other hand—and more insidiously—the refusal to address racism is often part of a conscious corporate strategy to appease conservative politicians and to continue leveraging racist content for financial gain.

Third, this Article identifies two main approaches for how platforms consider race: (1) racial targeting and (2) racialized threat assessments. The first approach treats race as a protected category, only prohibiting posts explicitly targeting an individual or group based on their perceived race. This method is typically found in rules against hate speech and harassment. There is no attempt to account for histories of subjugation or how race is reflected in contemporary power dynamics. The second approach rarely makes explicit mention of race at all. Instead, companies use secret blacklists and broad prohibitions to police racialized groups that are viewed as inherently dangerous. This approach is mostly deployed through policies against terrorism and violent extremism.

Building on this typology, I conclude by advancing an alternative model of race-conscious content policy. Interventions include accounting for vertical power arrangements, eliminating prohibitions that overburden political

[24] Cheryl I. Harris, *Whiteness as Property*, 106 HARV. L. REV. 1707, 1725 (1993).

participation, and publishing blacklists that contain banned individuals and organizations. Acknowledging the challenges and dangers of identifying racial groups, I also propose potential starting points that leverage individual design elements specific to individual platforms. Each of these interventions is an invitation to be clear-eyed about the ongoing and mutable nature of racism.

I. SOCIAL MEDIA AND THE STRUCTURAL MAINTENANCE OF WHITE SUPREMACY

This Part develops a theoretical framework for analyzing social media's role in the maintenance of white supremacy. First, this Part defines my approach to analyzing race, racialization, and the structure of white supremacy.[25] Second, it applies analytical insights from critical race theory to examine the role of content policy in legitimating and encoding racial hierarchies.

A. *Race, Racialization, and White Supremacy*

Given the theoretical and policy implications of this Article, it is essential to have a targeted discussion about what race is, how racialization occurs, and what it means for social media companies to advance white supremacy. Otherwise, we risk an analysis that purports to target racial harms without clear language for identifying those harms, let alone providing a meaningful remedy. At the same time, attempting to fasten a singular understanding to these concepts can lead to unproductive and limiting approaches. The adaptive nature of white supremacy requires a flexible approach to understanding the nature of racial subjugation in order to challenge its enduring power.

In this Article, I define "race" as a socially constructed grouping of people based on shared history, political power, and cultural production.[26] These

[25] My analysis of social media extends critical race theory to the corporate system of content moderation. *See generally* CRITICAL RACE THEORY: THE KEY WRITINGS THAT FORMED THE MOVEMENT (Kimberlé Crenshaw, Neil Gotanda, Gary Peller & Kendall Thomas eds., 1995). CRT's decades-long theorization over how law and policy shapes and maintains racial hierarchies is an essential starting point for understanding how content moderation is an important site of study for understanding contemporary racialization. Indeed, many of their central critiques—from their critique of colorblindness, *id.* at 257, to their rejection of First Amendment absolutism, *id.* at 481-83—have analytic insights for contemporary platform governance debates.

[26] This definition is indebted to racial theorizing across decades and disciplines. These include (among others), IAN HANEY LÓPEZ, WHITE BY LAW 7, 9 (2d ed. 2006) ("'[R]ace' is a fluctuating, decentered complex of social meanings that are formed and transformed under the constant pressures of political struggle." (quoting John O. Calmore, *Critical Race Theory, Archie Shepp, and Fire Music: Securing an Authentic Intellectual Life in a Multicultural World*, 65 S. CAL. L. REV. 2129, 2160 (1992))); CHARLES W. MILLS, FROM CLASS TO RACE: ESSAYS IN WHITE MARXISM AND BLACK RADICALISM 181-82 (2003) ("[R]ace is not natural but an artifact of sociopolitical decision making, so that one function of political power is deciding where the crucial boundaries are drawn."); MICHAEL OMI & HOWARD WINANT, RACIAL FORMATION IN THE UNITED STATES FROM THE 1960S TO THE 1990S, 55 (1994) (arguing

categories are not naturally occurring; they are part of an ongoing process that creates social hierarchies to simultaneously distribute racial benefits and burdens. I refer to this process as racialization.[27] The connection between racial categorization and racial hierarchy requires an analysis of white supremacy that is more expansive than typically conceived of by social media companies, which mostly focus on explicit racial animus. Here, I adopt Frances Lee Ansley's definition of white supremacy as "a political, economic and cultural system in which whites overwhelmingly control power and material resources, conscious and unconscious ideas of white superiority and entitlement are widespread, and relations of white dominance and non-white subordination are daily reenacted across a broad array of institutions and social settings."[28]

Dominant groups define and diffuse social hierarchies, making whiteness central to understanding the logics shaping racial hierarchies. Cheryl Harris writes that while whiteness is legally constructed as objective fact, "in reality it is an ideological proposition imposed through subordination."[29] Over time, the acts of subordination become obscured, making the benefits of whiteness appear natural or the product of legitimate achievement.[30] Charles W. Mills argues that this transformation of whiteness from racial category to "neutral baseline" makes any attempt to disrupt this system of racial privilege "sincerely and righteously viewed as an attack on fundamental human rights and freedoms."[31]

A subordinating power of whiteness is the ability to define the universal. The white experience is used as the baseline for neutral objectivity, an organizational starting point. Online, whiteness defines the universal by functioning as the principal logic and animating impulse for how technology gets conceptualized. As André Brock Jr. writes, "whiteness is infrastructural" to the Internet.[32] Social

race "signifies and symbolizes social conflicts and interests by referring to different types of human bodies").

[27] This definition adopts formulations from Kendall Thomas, Nash Professor of Law and Co-Director of the Center for the Study of Law & Culture at Columbia Law School. *See* Kendall Thomas, Comments at Frontiers of Legal Thought Conference, Duke Law School (Jan. 26, 1990) ("We are raced."); D. Marvin Jones, *Darkness Made Visible: Law, Metaphor, and the Racial Self*, 82 GEO. L.J. 437, 440 (1993) ("[R]ace is not so much a category but a practice: *people are raced*."); MILLS, *supra* note 26, at 185 (2003) ("[R]ace as central, political, and primarily a system of oppression—is (at least in broad outline) not at all new; but it has in fact always been present in oppositional African American thought.").

[28] Frances Lee Ansley, *Stirring the Ashes: Race, Class and the Future of Civil Rights Scholarship*, 74 CORNELL L. REV. 993, 1024 n.129 (1989).

[29] Harris, *supra* note 24, at 1730.

[30] *Id.* at 1777.

[31] MILLS, *supra* note 26, at 191.

[32] ANDRÉ BROCK, JR., DISTRIBUTED BLACKNESS: AFRICAN AMERICAN CYBERCULTURES 46 (2020) (describing how "unmarked" Internet resources are commonly understood as white by default). For an analysis of racial benefits and burdens in the context of web traffic, see generally Charlton McIlwain, *Racial Formation, Inequality and the Political Economy of Web Traffic*, 20 INFO., COMMC'N & SOC'Y 1073 (2017) (demonstrating person's race and racialization of websites collectively impacts how individuals navigate Internet).

media's approach to world building, from its user experience to its system of content moderation, is shaped by racialized understandings that center the privileges attendant to whiteness.

For example, despite growing acceptance that race is a social construct,[33] technologies ranging from genetic testing to facial recognition power a contemporary push to reimagine race as a fixed category that can be objectively detected.[34] Social media companies are a part of this movement, typically conceptualizing race as either a category that is approximated through user data for advertising purposes, or as a protected characteristic detectable only by explicit invocation.[35] This resurgence shares a lineage with older attempts to connect racial categorization with "natural, physical divisions among humans that are hereditary."[36] In truth, what is presented as objective or neutral more accurately represents the dominant worldview and power dynamic, with social outcomes that typically recreate and entrench existing racial hierarchies.[37]

B. *Content Moderation's Power To Legitimize White Supremacy*

Through an interconnected set of social practices, race obtains the power to shape knowledge, including broader norms around order and justice.[38] This

[33] *See* EDUARDO BONILLA-SILVA, RACISM WITHOUT RACISTS: COLOR-BLIND RACISM AND THE PERSISTENCE OF RACIAL INEQUALITY IN AMERICA 8 (2018) ("There is very little formal disagreement among social scientists in accepting the idea that race is a socially constructed category."); Ian Haney López, *The Social Construction of Race: Some Observations on Illusion, Fabrication, and Choice*, 29 HARV. C.R.-C.L. REV. 1, 27 (1994) (explaining source of racial categorization is found in "human interaction rather than natural differentiation").

[34] *See generally* DOROTHY ROBERTS, FATAL INVENTION: HOW SCIENCE, POLITICS, AND BIG BUSINESS RE-CREATE RACE IN THE TWENTY-FIRST CENTURY 59 (2011) (discussing interaction of contemporary science and technology with race); Alex Najibi, *Racial Discrimination in Face Recognition Technology*, SCI. NEWS, HARV. UNIV. (Oct. 24, 2020), https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/ [https://perma.cc/V7KC-3KH9] (noting study analyzing accuracy of facial recognition programs by IBM and Microsoft finding greater inaccuracy for women of color).

[35] *See Meta Privacy Policy*, META (June 15, 2023), https://www.facebook.com/privacy/policy [https://perma.cc/R9XL-HCPT] (noting Meta collects data about an individual's racial and ethnic background).

[36] *See* López, *supra* note 33, at 27 (1994) (describing and rejecting notions of "biological race").

[37] *See* Jessica Eaglin, *When Critical Race Theory Enters the Law & Technology Frame*, 25 MICH. J. RACE & L. 151, 162 (2021) (discussing race as sociohistorical structure shaping how humans interact). Additionally, critical race theory scholar Margaret Chon predicted in the nascent days of the Internet that the digital sphere would "reinscribe racial fault lines and reinforce racism." Margaret Chon, *Erasing Race? A Critical Race Feminist View of Internet Identity-Shifting*, 3 J. GENDER, RACE & JUST. 439, 442-43 (2000) (discussing problems with treating Internet as equal racial playing field).

[38] *See* Kimberlé Williams Crenshaw, *Twenty Years of Critical Race Theory: Looking Back To Move Forward*, 43 CONN. L. REV. 1253, 1349 (2011) ("Race is not natural, yet race is embedded in social relations, many of which are naturalized by the knowledge-making disciplines that we have inherited and participate in reproducing.").

racialized knowledge embeds white understandings of the world and imposes them on others, formulating specific problems and limiting the scope of potential solutions. The vested interest in racial hierarchies and their attendant privileges can let dominant approaches to knowledge fail to fully grapple with group privilege and rule. Charles W. Mills describes this pattern of nonknowing as a structure of "white ignorance, motivated inattention, self-deception, historical amnesia, and moral rationalization."[39] The system of social media content moderation reflects this pattern of nonknowing, deploying a system of speech regulation that accepts existing racial hierarchies and structures policies in a manner that prevents meaningful redress for the harms imposed on marginalized communities. The profit-driven choices informing these racial burdens and benefits reflect unique powers of scale and regulatory freedom.

This Section advances a two-part framework for understanding how content policy contributes to the cultural, economic, and political project of white supremacy. The first is through colorblind policies that obscure racial hierarchies and make them appear natural. The second is through a racialized understanding of speech harms that valorize white bigotry while broadly removing posts from communities of color under the guise of hate speech, harassment, or terrorism.

1. Equating Universal Policies with Equal Opportunity Obscures Racial Subjugation

Social media content policy typically prohibits attacks based on a person's race, regardless of the purpose or differing histories at play.[40] This colorblind approach accepts existing racial hierarchies and makes them appear natural and inevitable. The decision to equate universal rules with equal access obscures the subordination that colorblind content policy facilitates. By erasing the history and effects of white domination,[41] race becomes reduced to merely "'skin color' or country of ancestral origin."[42] Cheryl Harris analyzes this apolitical understanding of race as recasting the "privileges attendant to whiteness as legitimate race identity."[43]

Treating all uses of racist speech equally erases a foundational purpose behind racialization as a tool for legitimating one group's privilege over others.[44] As

[39] MILLS, *supra* note 26, at 190.

[40] *See* sources cited *infra* note 91.

[41] Neil Gotanda, *A Critique of "Our Constitution Is Colorblind,"* 44 STAN. L. REV. 1, 1-2 (1991) ("[C]olorblindness is a form of race subordination in that it denies the historical context of white domination and Black subordination.").

[42] *Id.* at 4. Gotanda advances four understandings of race that the Supreme Court deploys throughout its opinions: status-race, formal-race, historical-race, and culture-race. *Id.* at 3-4.

[43] Harris, *supra* note 24, at 1768-69.

[44] *See* Kimberlé Williams Crenshaw, *Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law*, 101 HARV. L. REV. 1331, 1379-81 (discussing "white norm" as "statement of the positive social norm, legitimating the continuing domination of those who do not meet it").

further discussed *infra* Parts II-III, colorblind content policy provides several economic and political advantages for social media companies. Economically, colorblind policies enable companies to enter new markets without adequate attention to local dynamics or social hierarchies that pose context-specific dangers. Moreover, universal colorblind policy is typically narrow, prohibiting any explicit attack based on race.[45] This limited restriction allows companies to maximize content posted to their platforms, including racist content that has proven allure to people.[46] Finally, reducing analyses of racism to a narrow formula fosters faster and increasingly automated moderation, sacrificing accuracy and effectiveness. By limiting moderator discretion and imposing a narrow set of prohibited content, platforms proliferate a restricted understanding of racism and train their automated systems to replicate these choices at scale. This simplified logic, distributed worldwide through thousands of workers and automated systems, ensures the digital understanding of racism is encoded and perpetuated into the future.

On the political front, colorblind content policy upholds a discriminatory status quo to avoid negatively impacting groups with regulatory power. Colorblind policies that purport to treat everyone equally take a nominally consistent stance but do little to prohibit the type of racism deployed by dominant groups. Dominant-group racism largely operates via dog whistles or other less explicit forms of racism unlikely to trip emerging social norms that shun open bigotry.[47] On the other hand, colorblind content policy is more likely to impact the ways marginalized groups discuss the operation and impact of structural racism. Living under a system of racial subjugation, marginalized groups are more likely to explicitly mention the subordinating role of whiteness and white supremacy. At times, these critiques manifest in direct and explicit attacks directed toward the powerful. To effectively fight the subordinating role of hate propaganda, policy must account for historical oppression[48] and differentiate dissent directed to the powerful from hate speech directed to the

[45] *See* sources cited *infra* note 91 (comparing Twitter, Meta, and YouTube hate speech policies).

[46] *See, e.g.*, REBECCA LEWIS, DATA & SOC'Y, ALTERNATIVE INFLUENCE: BROADCASTING THE REACTIONARY RIGHT ON YOUTUBE 43 (2018), https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf [https://perma.cc/U64Z-4Q3G] (using tweet shared by British far-right conspiracy theorist Paul Joseph Watson receiving plaque YouTube sent him for surpassing one million subscribers as example of potential intervention point for YouTube to assess content of his channel).

[47] However, when colorblindness ceases to benefit the powerful, policies are changed to restore the benefits attendant to whiteness. *See* discussion *infra* Section III.B (discussing use of racialized threat assessments).

[48] *See* MARI MATSUDA, CHARLES R. LAWRENCE III, RICHARD DELGADO & KIMBERLÈ WILLIAMS CRENSHAW, WORDS THAT WOUND: CRITICAL RACE THEORY, ASSAULTIVE SPEECH, AND THE FIRST AMENDMENT 36 (1993) (proposing identification of "worst, paradigm[atic] example of racist hate messages" should consider, in part, whether "the messages is directed against a historically oppressed group").

marginalized.[49] Colorblind content policy suppresses this possibility under the guise of neutrality and efficiency.

2. Content Policy Legitimizes Racist Cultural Norms

White supremacy depends on a cultural system that justifies a racial hierarchy with whiteness at its zenith. In practice, this system valorizes the contributions of whiteness while protecting the violence, exploitation, and erasure that facilitate its hegemony. Online, conscious and unconscious understandings about people are shaped, refined, and advanced though content policy. Hate speech, harassment, and terrorism policies define who is a threat and who is merely being provocative. By establishing policies around acceptable speech that reflect racist assumptions about marginalized communities, social media companies allow their online platforms to reflect their conceptions of racialized others.

Social media facilitates cultural hegemony through dehumanization, cultural exploitation, and erasure. Communities of color are policed in a manner that reinforces subordinating perceptions of their cultural contributions as lewd, violent, and uncivilized. André Brock, Jr. argues "unmarked" spaces are typically coded as white and "belonging to whiteness," and thus "civilized until a nonwhite actor or group is seen utilizing them."[50] Content policy shapes what constitutes uncivilized conduct and aligns it with nonwhiteness. By broadly labeling content from subordinated communities as incitement to violence, terrorism, misinformation, or hate speech, platforms generate signals and systems for what it means to challenge the status quo. Ignoring the differences between hate speech and dissent creates an operational efficiency that coincides with racial hierarchies.[51]

A separate (but no less essential) aspect of racial domination is the erasure and extraction of nonwhite contributions. The same principle that treats communities of color, especially blackness, as uncivilized is also an essential avenue for creating knowledge, art, and cultural advancement. Often, these advancements are shaped by modes of expression fostered under the boot of censorship. Online slang, facilitated by conversation that might otherwise be removed, eventually gets repurposed and used by dominant groups.[52] Similarly,

[49] *Id.* at 10 (discussing key difference between disparities in power wielded by targets of communication).

[50] BROCK, *supra* note 32, at 46 (internal quotation marks and emphasis omitted).

[51] When analyzing hate speech, Mari Matsuda highlighted the importance of accounting for historical oppression and differentiating between dissent directed to the powerful and hate speech directed to the marginalized. *See* MATSUDA ET AL., *supra* note 48, at 10.

[52] *See generally* Kendra Calhoun & Alexia Fawcett, "They Edited Out Her Nip Nops": Linguistic Innovation as Textual Censorship Avoidance on TikTok 11 (unpublished manuscript) (on file with author) (discussing use of self-censorship employed by marginalized groups to avoid content violations); Taylor Lorenz, *Internet 'Algospeak' Is Changing Our Language in Real Time, from 'Nip Nops' to 'Le Dollar Bean,'* WASH. POST (Apr. 8, 2022,

the "uncivilized" digital practices of nonwhiteness nonetheless shape new product features and later highlight the political opportunities that an open Internet facilitates.[53]

In sum, analyzing racial logics is an essential starting point to understanding the enduring double standard in social media content moderation. Critical race theory provides a framework for analyzing how content policy legitimizes a discriminatory status quo. Economic efficiency, regulatory concerns, and cultural prejudices push content policy toward an understanding of race that is shallow at best and intentionally subordinating at worst. The next Part discusses the development of content policy at social media companies, analyzing the concerns, worldviews, and external pressures that produced a racialized understanding of speech harms.

II. THE DEVELOPMENT OF COLORBLIND CONTENT POLICY

This Part provides an overview of the evolution of social media content policy, with a particular focus on how companies developed an approach that avoided grappling with structural racism. This approach, which I refer to as "colorblind content policy," not only ignores the vertical hierarchy that exists between racial groups but also protects the power attendant to whiteness. At times, this explains why platforms are consistently caught wrong-footed in their attempts to moderate racist discourse that trades in everything from slurs to coded messages to violent appeals. At other times, the refusal to address the role of race is part of company strategy to appease conservative politicians and their desire to freely leverage racism for political and economic gain.

To understand how social media companies arrived at colorblind content policy, we must analyze the goals and worldviews of the individuals tasked with drafting content policy: early moderators, content policy managers, and company executives. Three interconnected themes emerge.

First, whether through gut intuition or ad hoc policy administration, content moderation advances a racialized understanding of what is "acceptable" bigotry and what goes too far. For white speakers, racist speech is simply pushing the envelope or protecting manifestations of ethnic pride. For communities of color, hateful language, regardless of the target, requires immediate action to prevent real-world harm.

Second, the administration of scalable and "value-neutral" content policy eschews race in favor of speed and falsely apolitical moderation. This approach eliminates the ability to develop policy that addresses the fluid nature of racism

7:00 AM), https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/ (describing use of algorithmic-avoidant language usually created originally by communities of marginalized people).

[53] *See, e.g.*, Khiara M. Bridges, *Language on the Move: "Cancel Culture," "Critical Race Theory," and the Digital Public Sphere*, 131 YALE L.J.F. 767, 776-77 (2021-2022) (discussing practice of online cancel culture and its power as "digital weapon of the weak that allows 'coalitions of the Othered' to commune, and perhaps heal, through acts of public condemnation" (footnotes omitted)).

while facilitating its stratifying power. It also derives from a distrust of decentralized moderators from around the world, a significant departure from the largely white, upper-class moderators that led early moderation efforts.[54] Uniformity strips the nuance and nimbleness necessary to meaningfully combat the fluid nature of racism.

Finally, the strategic focus on risk management and political appeasement has resulted in a two-tier system that doles out a measured hand for the powerful and an iron fist for the marginalized. This approach reflects an attempt to ward off regulation, but also coincides with executive worldviews that do not view white supremacist speakers as serious threats until their actions spill over into offline violence.

A. *How Much Racism Is Too Much?*

Content moderation was often an afterthought for platforms. At YouTube, Heather Gillette—hired as an office manager—also took on the role of handling content screening, driven in part to avoid negative press coverage.[55] Meta initially had a small group of people tasked with moderating content through water cooler advice, email chains, and a list of examples.[56] Three years into Twitter's operation, when it had around five million people on its platform, that company's support team was composed of three employees, and a "significant part" of their support queries were regarding "rules of engagement on Twitter."[57]

The guiding principle of early content policy was moderator intuition. Early moderators removed things that made them uncomfortable, which was largely "Hitler and naked people."[58] Among Meta's early moderators, the informal guidance was "[i]f something makes you feel bad in your gut, take it down."[59] Twitter's content moderation efforts were extremely limited, earning it an

[54] *See, e.g.*, Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1634 (2018) (discussing changes in demographic makeup of Facebook moderators from "homogenous college students" to outsourced moderation in places like Hyderabad, India (internal citations omitted)).

[55] *See* MARK BERGEN, LIKE, COMMENT, SUBSCRIBE: INSIDE YOUTUBE'S CHAOTIC RISE TO WORLD DOMINATION 31-33 (2022) (discussing Gillette's early efforts to lead YouTube's "Safety, Quality, and User Advocacy" team, including instance where she pushed everyone to scrub YouTube of genitalia to avoid potentially bad story from journalist about salacious material on site).

[56] *See* SHEERA FRENKEL & CECILIA KANG, AN UGLY TRUTH: INSIDE FACEBOOK'S BATTLE FOR DOMINATION 35-36 (2021) (discussing dissemination of moderation policymaking through informal channels without context or explanation).

[57] *See* Biz Stone, *The Zen of Twitter Support*, TWITTER BLOG (Jan. 15, 2009), https://blog.twitter.com/official/en_us/a/2009/the-zen-of-twitter-support.html [https://perma.cc/27AS-T5NC].

[58] Simon Van Zuylen-Wood, *"Men Are Scum": Inside Facebook's War on Hate Speech*, VANITY FAIR (Feb. 26, 2019), https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech [https://perma.cc/C87Y-6UZL] (discussing Meta's early content moderation policy).

[59] FRENKEL & KANG, *supra* note 56, at 36.

enduring reputation as the "free speech wing of the free speech party."[60] At YouTube, removals were primarily limited to pornography, spam, and incitement to violence; moderators were instructed to use their judgment.[61]

Platforms have always known racism existed on their services. For them, the question was: What is a tolerable amount of racism? David Willner, one of Meta's first content moderators, described one of their earliest objectives as "Don't become a radio for future Hitler."[62] While platforms may have been on the lookout for Nazism, they were ill-equipped to see that modern fascism might not exactly replicate the original. By treating Nazism as an extremity, instead of white supremacy as common and widespread, companies set themselves up to constantly be blindsided by the real world. Even as awareness grew of the hatred spread by racist conspiracy theorists like Alex Jones, forcing platforms to act, companies' inability to understand the nature of racism led to half measures. At Meta, Zuckerberg did not view Alex Jones as a hate figure, so he allegedly overrode his team's enforcement recommendations and settled on narrower enforcement against Jones.[63] Similarly, YouTube initially tolerated openly racist comments from popular creators ranging from Felix ("PewDiePie") Kjellberg to white supremacist Stefan Molyneux, justifying their videos calling for the death of Jews or advocating for a resurgence of eugenics as "pushing the envelope" or simply participation in the marketplace of ideas.[64]

B. *Operationalizing Content Moderation: Math, Speed, and Scale*

Content policy focusing on scale, speed, and economics necessarily involves choices that avoid grappling with the context-specific operation of racial stratification. The role of formalized content policy is important for understanding two components: (1) the shift from standards to rules, and (2) the desire to transform complex and enduring social questions into engineering challenges. In each instance, this decision requires a flattening of experience while imposing a system ensuring that current social stratification continues.

[60] *See* Josh Halliday, *Twitter's Tony Wang: 'We Are the Free Speech Wing of the Free Speech Party,'* GUARDIAN (Mar. 22, 2012, 11:57 AM), https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech [https://perma.cc/T7JY-UQSN].

[61] *See* BERGEN, *supra* note 55, at 85 (recounting early content moderation at YouTube).

[62] FRENKEL & KANG, *supra* note 56, at 92.

[63] *See* Ryan Mac & Craig Silverman, *"Mark Changed the Rules": How Facebook Went Easy on Alex Jones and Other Right-Wing Figures*, BUZZFEED NEWS (Feb. 22, 2021, 1:14 PM), https://www.buzzfeednews.com/article/ryanmac/mark-zuckerberg-joel-kaplan-facebook-alex-jones [https://perma.cc/479K-M2FV].

[64] *See* BERGEN, *supra* note 55, at 226, 279 (describing how YouTube initially defended platforming of PewDiePie and several alt-right YouTubers as simply pushing envelope before eventually ending plans to release series with PewDiePie, removing him from their premium ad tier, and deplatforming other alt-right YouTubers after similar social media companies removed their accounts).

As companies grew, the improvised list of standards and exceptions proved unwieldy. Moderation teams grew from a handful of repurposed employees to small but dedicated teams, with some moderation efforts beginning to move offshore to places like Hyderabad.[65] The identities, worldviews, and biases of the individuals shifted.[66] Moderation was no longer contained within the homogeneity of mostly white and college-educated Americans.[67] The difficult and traumatizing work of reviewing hundreds of racist and violent images began receding deeper into obscurity as part of the broader resource extraction that exploits the Global South.[68]

At Meta, individuals like Dave Willner and Jud Hoffman began "formalizing and consolidating" the original rules.[69] According to Willner, this move sought to eliminate "standards that evoked nonobservable values, feelings, or other subjective reactions" and replace those standards with "objective" rules.[70] The goal was to arrive at a system where moderator decisions did not vary based on the person's viewpoint and biases.[71] Set aside as too "vague, capricious, fact dependent, and costly to enforce," a rule-based system instead sets out firm lines that sacrifice nuance for efficiency.[72] The ongoing pursuit of speed and scale reduces complex human tendencies into a set of formulas. Content policy is the most visible manifestation of this approach, but it is only one part of a broader project that extends to the increasing reliance on algorithms applying limited logics to automate the interpretation of content policy. This approach to quickly disposing of a question via invisible decisions grows out of the similarly dismissive approach that tasked a group of people with moderating, which later developed into massive systems where thousands of people view swaths of objectionable content without adequate pay or protection.[73] The movement away

[65] *See* Klonick, *supra* note 54, at 1633-34 (discussing outsourcing of content moderation by Facebook and subsequent shift to low-wage labor performed by individuals in places like Philippines and India).

[66] *See id.*; *see also* SARAH ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA 38-40 (2019) (discussing how labor of social media content moderation "is fractured organizationally and geographically"; located in places like India, Ireland, and Philippines; and broken up into four types: "in-house, boutiques, call centers, and microlabor websites").

[67] *See* Klonick, *supra* note 54, at 1634 (chronicling shift of moderation teams from "homogenous college students" to international teams).

[68] A discussion of racial capitalism and its manifestation in commercial content moderation is beyond the scope of this Article but will be taken up in future work. For analyses of the origins and development of commercial content moderation, see generally SARAH ROBERTS, *supra* note 66.

[69] Klonick, *supra* note 54, at 1634.

[70] *Id.* at 1633-34.

[71] *Id.* (describing intent of approach as ensuring consistency and uniformity across moderators).

[72] *Id.* at 1632 (comparing rule-based and standard-based methods for content moderation).

[73] *See generally* ROBERTS, *supra* note 66 (chronicling work of moderators managing content on large online platforms).

from standards eliminated the possibility to deploy race-conscious and context-specific content moderation standards tailored to the varied dynamics across the spaces where platforms offer their services.

In an attempt to avoid dealing with the messiness of historic oppression, attorneys and policy managers perpetuated the myth that everyone was equal online. In leaked audio from an employee Q&A, Mark Zuckerberg explained the company's ethos for thinking about social stratification: "We've made the policy decision that we don't think that we should be in the business of assessing which group has been disadvantaged or oppressed."[74] This position crystalized an approach to content moderation in which companies are uninterested in addressing the complexities of the global communities they host. This narrow vision of racism keeps content policy either blind or neutral to the multifaceted and complex operation of racism.

C. *Regulatory Fears and Ideological Alignment*

Founders and executives retain ultimate control over the drafting and administration of content policy. These individuals typically step in for high-profile decisions like the suspension of Donald Trump,[75] but their visions also carry over into the broader content-moderation landscape. It is essential to analyze the outsized role of these individuals because they provide insight into the products they are building and for whom they are intended.

In 2016, Meta met with a set of conservative figures like Glenn Beck, Arthur Brooks, and Jenny Beth Martin (among others) to assure them of Facebook's neutrality.[76] According to a former Facebook employee, only Republican employees were allowed in the room.[77] This meeting continued an ongoing claim of bias against conservatives that, while unproven, remains loud and shakes platforms. Meeting with conservative pundits "marked a pivotal moment for [Sheryl] Sandberg, whose decisions would become increasingly dictated by fears that a conservative backlash would harm the company's reputation and invite greater government scrutiny."[78] At YouTube, when asked by an employee what her number one fear was, Susan Wojcicki immediately replied "regulation."[79] This may explain why the company was slow to react to the

[74] Newton, *supra* note 2.

[75] Mark Zuckerberg, FACEBOOK (Jan. 7, 2021, 5:47 AM), https://www.facebook.com/zuck/posts/10112681480907401 [https://perma.cc/L744-LGFC] (announcing decision to extend block of Donald Trump's Facebook and Instagram accounts after his actions on January 6, 2021).

[76] *See* FRENKEL & KANG, *supra* note 56, at 81.

[77] *Id.*

[78] *Id.* at 82-83.

[79] BERGEN, *supra* note 55, at 366.

immense popularity of alt-right YouTubers who actively harassed other creators, and pulled back after political blowback cooled.[80]

Largely uninterested in the maintenance of an online community, many founders either left their companies or handed off content policy to deputies. In their place rose a set of individuals who operated with a more political mindset, seeking to foster favorable regulatory environments.[81] Connecting the policymaking and lobbying efforts as part of political strategy is essential for understanding the drafting and enforcement of colorblind content-moderation policy. Frequent and ongoing cries of anticonservative bias facilitate an approach to content policy that avoids line drawing in a manner that could have a disparate impact on the media and politicians. For example, Joel Kaplan's involvement in Meta's efforts to combat hate speech and misinformation reflects a concern that it would be seen as politically biased.[82] At YouTube, similar concerns facilitated the protection of white nationalists like Richard Spencer and Stefan Molyneux for years.[83]

Efforts to quiet concerns around so-called censorship of conservative voices demonstrate the limits of principled content policy. But while this is doubtless a political calculation, it is important to also consider the extent to which political appeasement coincides with executives' own racial discomfort. While there may be debates and individual deviation at the moderator level, executives are called upon to make the most consequential decisions.[84] Their worldviews inform not only individual enforcement, but the operation of policy overall. Largely white and upper class, these individuals were raised in de facto segregated schools and neighborhoods that limited their exposure to the lived reality of marginalized

[80] *Id.* (asserting YouTube began trying to internally regulate through updated hate speech and harassment rules in attempt to evade government regulation). After negative news coverage regarding PewDiePie's racist content cooled, the company reestablished ties with him. *See* Mark Bergen, *How YouTube Broke Up with PewDiePie (Then Got Back Together Again)*, VERGE (Sept. 6, 2022, 1:00 PM), https://www.theverge.com/23339163/pewdiepie-like-comment-subscribe-mark-bergen-book-excerpt-youtube-adpocalypse [https://perma.cc/D5FB-37JA].

[81] *See* ROBERTS, *supra* note 66, at 93.

[82] Benjamin Wofford, *The Infinite Reach of Joel Kaplan, Facebook's Man in Washington*, WIRED, https://www.wired.com/story/facebook-joel-kaplan-washington-political-influence/ (Mar. 14, 2022, 3:30 PM) (describing how Joe Kaplan, Facebook's Vice President of Global Policy, manages balance of political speech on Facebook).

[83] *See* BERGEN, *supra* note 55, at 260-61, 379. While this Article focuses on the United States, similar approaches occur in other countries. In India, for example, a similar approach avoids enforcement of hate speech policies that could harm the government in power. *See* Newley Purnell & Jeffrey Horwitz, *Facebook's Hate-Speech Rules Collide with Indian Politics*, WALL ST. J. (Aug. 14, 2020, 12:47 PM), https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346 (describing Facebook's actions regarding Indian politicians with posts violating Facebook's hate speech rules).

[84] *See, e.g.*, Zuckerberg, *supra* note 75.

communities.[85] We see the enduring consequences of these experiential deficiencies over the years. Executives enforce rules against white nationalists and conspiracy theorists in a limited capacity—failing to see the dangers they pose—only to later have to renew enforcement efforts after they are connected to offline violence.

Despite ongoing violence and bigotry connected to white supremacist groups, executives' reactions are to apply limited policies against explicit bigotry. Even belated enforcement actions against neo-Nazis can face reversal once a chief executive changes. For example, after Elon Musk purchased Twitter in October 2022, he led efforts to reinstate the accounts of numerous white supremacists and conspiracy theorists.[86] Similarly, when YouTuber PewDiePie posted explicitly anti-Semitic posts calling for the death of Jews, executives believed his actions were driven by awkwardness, not hatred.[87] Time and time again, these individuals show they cannot "identify the systemic biases of the world," let alone lead efforts to combat them.[88]

III. COLORBLIND CONTENT POLICY IN PRACTICE

To understand how content policy upholds racial stratification, we must understand how content policy addresses questions of race. This Article identifies two main approaches: (1) racial targeting, and (2) racialized threat assessments. The first approach treats race as a protected category, prohibiting direct targeting of an individual or group based on their perceived race.[89] This method is typically found in rules against hate speech and harassment. There is no attempt to account for histories of subjugation or the ways in which race is reflected in contemporary power dynamics. Instead, colorblind hate speech and harassment rules foster white supremacy while suppressing the voices of communities of color speaking out against it or simply speaking with one another using provocative or unsterilized language. This choice preserves status quo discrimination under the banner of equality.

The second approach rarely makes explicit mention of race at all. Instead, secret blacklists and broad prohibitions are used to police racialized groups that are perceived as a threat. This approach is deployed through policies against terrorism and violent extremism ("TVE"). For racialized minorities, their speech

[85] FRENKEL & KANG, *supra* note 56, at 65 ("He couldn't recognize, perhaps, that his life experience—his safe and stable upbringing, his Ivy League pedigree, his ability to attract investors—was not the same as everyone else's.").

[86] Taylor Hatmaker, *Elon Musk Just Brought an Infamous Neo-Nazi Back to Twitter*, TECHCRUNCH (Dec. 2, 2022, 4:22 PM), https://techcrunch.com/2022/12/02/elon-musk-nazis-kanye-twitter-andrew-anglin/ [https://perma.cc/3PWV-25JA].

[87] BERGEN, *supra* note 55, at 280 (noting after PewDiePie made anti-Semitic statements, YouTube executives did not discuss issue publicly).

[88] FRENKEL & KANG, *supra* note 56, at 65 (discussing limited life experiences and empathy of Mark Zuckerberg).

[89] Content policy can expand beyond race to account for certain related concepts, such as ethnicity, nationality, and immigration status. *See infra* note 91.

is subjected to policing that severely limits their ability to speak about simple topics, let alone participate in nuanced political discussions.[90] On the other hand, dominant groups and their leaders are permitted wide latitude to praise and support the use of violence.

A. *Racial Targeting*

Race is a protected characteristic across all platforms, but some expand the list to account for attacks based on national origin, immigration status, ethnicity, and religious affiliation.[91] This protected-characteristic approach is what Neil Gotanda defines as "formal-race": where race is treated as a "neutral, apolitical description[]" disconnected from any reality of social status or historical oppression.[92] This approach feeds into platforms' desires to deploy scalable solutions that minimize moderator discretion and enable quick decisions.[93] Treating all racial classifications equally provides a consistent line for moderators to act, but it also makes "a vertical hierarchy appear horizontal" and hides the subjugation inherent to racial stratification.[94]

Platforms moderate race as a target mostly through their policies against hate speech and harassment. In each instance, the stated concerns are the potential

[90] *See, e.g.*, *Case Decision, Case Decision 2021-009-FB-UA*, OVERSIGHT BD., https://www.oversightboard.com/decision/FB-P93JPX02 [https://perma.cc/AK4L-Z5ZK] (last visited Nov. 9, 2023) (discussing vagueness in Meta's Dangerous Individuals and Organizations Community Standard). The DOI policy, which appear to primarily regulate content in the Middle East and North Africa, primarily impacts the speech of communities of color. *See* Public Comment Appendix at 28, *2021-009-FB-UA*, https://oversightboard.com/attachment/1192744021229855/ [https://perma.cc/Q3R2-UADU] (last visited Nov. 9, 2023) (drawing FaceBook Board's attention to disproportionate effect standards have on journalists and activists in the middle east, where "[i]n 2020, over the span of a day, Facebook deactivated the accounts of 52 Palestinian journalists and activists. It also deleted at least 35 accounts of Syrian journalists and activists documenting human rights abuses.").

[91] *Hate Speech*, META TRANSPARENCY CTR. [hereinafter *Meta Hate Speech Policy*], https://transparency.fb.com/policies/community-standards/hate-speech/ [perma.cc/TB9J-KMPD] (last visited Nov. 9, 2023). Twitter and YouTube's protected characteristics significantly overlap, but Twitter excludes immigration status. *Compare Hateful Conduct*, TWITTER HELP CTR. [hereinafter *Twitter Hate Speech Policy*] (Apr. 2023), https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [https://perma.cc/DGE2-H8BK], *with Featured Policies*, GOOGLE TRANSPARENCY REP. [hereinafter *YouTube Hate Speech Policy*], https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en (last visited Nov. 9, 2023). Both Twitter's and YouTube's protected characteristics lists overlap significantly with Meta's, although Twitter does not protect immigration status, and YouTube protects victims of a violent event and veterans. *See Meta Hate Speech Policy*, *supra*; *Twitter Hate Speech Policy*, *supra*; *YouTube Hate Speech Policy*, *supra*.

[92] Gotanda, *supra* note 41, at 4.

[93] *See* Klonick, *supra* note 54, at 1634 (discussing Facebook's development of content moderation guidelines).

[94] STEVE MARTINOT, THE MACHINERY OF WHITENESS: STUDIES IN THE STRUCTURE OF RACIALIZATION 86 (2010).

for inciting offline violence and the impact on people's ability to communicate online.[95] Over time, company framings began incorporating statements that signaled an understanding of how race and power can impact the harms of hate speech and harassment, but the underlying rules have remained the same.[96]

While broad in theory, protections against racist attacks are narrow in practice. Content policies typically prohibit only explicitly *direct* attacks explicitly *based* on a protected characteristic.[97] While these policies expanded over the years to account for specific manifestations of racism,[98] in each instance direct targeting based on a protected characteristic is necessary, and all groups are subject to the same protection. Platforms' approaches to moderating race as a target can be broken down into two categories: (1) the use of direct attacks such as slurs, stereotypes, and harassment; and (2) calls for racial violence and claims of racial inferiority, exclusion, or supremacy. This Part discusses each approach in turn.

1. Direct Attacks

The dominant approach to addressing racism leverages prohibitions against the use of slurs and stereotypes, as well as rules against racist harassment. Despite an understanding of the harms stemming from this content, the policy approach still demands explicit connections between the racial target and the prohibited comparison. This formality creates ineffective rigidity by targeting a force powered by flexibility. By removing any understanding about the histories of subjugation, this approach makes moderation easier—but ultimately less effective. Content policy that does not account for racism's benefits and burdens creates new ways to perpetuate the unequal distribution. It can also function as a silencing force against voices pushing back against social injustices.

Slurs, stereotypes, and dehumanization trade in the power of shared understanding, a shortcut to beliefs shaped by racist tropes. The maintenance of these comparisons plays an important role in normalizing racism, upholding racial hierarchies, and laying a path toward incitement to violence.[99] While this makes moderation essential, platform approaches are limited by requiring explicit targeting, even though the cultural strength and general understanding of these racist norms means that no explicit mention of a target race is necessary

[95] *See, e.g.*, *Meta Hate Speech Policy*, *supra* note 91 ("[Hate speech] creates an environment of intimidation and exclusion, and in some cases may promote offline violence.").

[96] *See, e.g.*, *Twitter Hate Speech Policy*, *supra* note 91 ("Research has shown that some groups of people are disproportionately targeted with abuse online. For those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature, and more harmful.").

[97] *See, e.g.*, *id.*; *Meta Hate Speech Policy*, *supra* note 91 ("We define hate speech as a direct attack against people . . . ."); *YouTube Hate Speech Policy*, *supra* note 91.

[98] *See infra* note 181 and accompanying text.

[99] *See* MATSUDA ET AL., *supra* note 48, at 23-24 (stressing racist hate messages and propaganda easily lead to violence).

to cause harm. Similarly, the ongoing struggle against racism causes groups to constantly attempt to defang language or speak out against their oppression using racially charged language.[100] Platform policies equating all use of racially charged language fail to recognize that what makes racist speech particularly harmful is its use as "a mechanism of subordination, reinforcing a historical vertical relationship."[101]

Commitment to universal content policy leads to situations where platforms tie themselves into knots trying to justify disparate outcomes. For example, Meta grappled with how to moderate posts stating "men are trash," a statement which may violate rules against dehumanization, but barring such statements would fail to consider the purpose behind the policy.[102] There is no call to violence, no meaningful exclusion from public participation, and no historical subjugation to draw from. Instead, the moderation decisions for these posts use self-imposed restrictions as a justification for removals that do little to advance the purpose and values of the rules. While Meta's own moderation team favored a hate-speech policy that accounted for power imbalances, head of global content policy management Monika Bickert's instinct was to treat each group equally.[103] She was fearful that accounting for historical discrimination would make it so that the hate speech policy would apply "to everybody—except for men," and that it would be poorly received.[104] In the end, the policy remained the same: with all forms of dehumanization treated equally.[105]

Colorblind moderation of racist comparisons also misses an essential component of modern bigotry: denial and deflection. The racist is the person who made something explicitly about race, not the person who understood the message and needed no explicit reference. Implementing a content-moderation system that adopts this defensive and deflective understanding of racism protects racists. Platforms' mechanical approach to understanding this type of attack also creates an easy escape valve to protect even explicit racism when it comes from a dominant group or a speaker with power.

Take a platform-provided sample stereotype: "All [protected characteristic or quasi-protected characteristic] are 'criminals.'"[106] This sample lists both the protected characteristic and the stereotype (criminals). Requiring an explicit connection between a stereotype and the protected group systematizes a narrow understanding of how attacks materialize. Instilling an order of operations that is easily understood by a moderator or a machine flattens the experience of racism to one that requires no doubt about a person's intent.

[100] *Id.* at 40 (arguing hate speech arising from experiences of oppression should be tolerated).

[101] *Id.* at 36.

[102] Van Zuylen-Wood, *supra* note 58 (recounting discussions about such posts).

[103] *See id.* (sharing Bickert's reasoning for treating genders equally in hate speech).

[104] *Id.*

[105] *See id.*

[106] *Meta Hate Speech Policy*, *supra* note 91 (alteration in original).

The absurdity of this approach is reflected in the protections for Congressman Clay Higgins's direct attacks against "radicalized" Islamic suspects.[107] In a Facebook post, he wrote "Hunt them, identify them, and kill them. Kill them all. For the sake of all that is good and righteous. Kill them all."[108] The post, which relied both on incitement to violence and the use of Islamic stereotypes, nonetheless stayed up because the post targeted a subcategory of Muslims: radicalized ones.[109] This undisclosed ability to simply add a qualifier to avoid enforcement results in a rule that takes a forgiving look at even naked bigotry. The radicalized, the violent, the illegal, the underperforming, the unpatriotic—with or without a racialized target, this qualifiers further narrowed enforcement.

Operationalizing colorblind content policy, a 2017 Meta internal training document informed moderators that when considering female drivers, Black children, and white men, only white men would be protected by the hate speech policy.[110] The rationale was that both race (white) and sex (male) are protected characteristics, whereas the other examples included what Meta considers quasi- or nonprotected characteristics, namely age (in the Black children example) and driver status (in the female drivers example).[111] In response to media fallout, Meta announced it would change its enforcement systems to deprioritize comments about "whites," "men," and "Americans," but it appears that these changes were only reflected in internal guidance (if at all) because no change was visible in the company's public-facing policies.[112] While a rule update allegedly eliminated the protection for subgroups, Meta still did not take down Congressman Higgins's post.[113] According to Bickert, this is because the government official's post is inherently newsworthy.[114] Tied to a simple on-off switch, this approach does not attempt to limit its distribution, make an assessment of the potential for violence, or use any other context-specific determinations that would account for the connection between powerful figures making dehumanizing stereotypes and offline violence. Thus, this formal

[107] Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children*, PROPUBLICA (June 28, 2017, 5:00 AM), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms [https://perma.cc/5UUP-6YYT].

[108] Congressman Clay Higgins (@captclayhiggins), FACEBOOK (June 4, 2017, 5:57 PM), https://archive.is/95FO1.

[109] Angwin & Grassegger, *supra* note 107.

[110] *Id.*

[111] *See id.*

[112] Adam Smith, *Facebook Comments Like 'White Men Are Stupid' Were Algorithmically Rated as Bad as Antisemitic or Racist Slurs, According to Internal Documents*, INDEPENDENT (Dec. 4, 2020, 11:20), https://www.independent.co.uk/life-style/gadgets-and-tech/facebook-comments-algorithm-racism-b1766209.html [https://perma.cc/W57F-MJ9W].

[113] Van Zuylen-Wood, *supra* note 58 (noting post still did not violate company hate speech rules).

[114] *Id.*

approach to hate speech captures little but the most explicit forms of racism, giving even naked bigotry a deferential reading when used by those with power.

a. *Silencing Resistance While Protecting Coded Racism*

While it would seem like a list of banned words could be easily operationalized, in practice, such an approach would fail to account for instances where groups attempt to reclaim slurs or where context eliminates their use as hate speech. These complications are not easily reflected in policy, though platforms like YouTube sometimes try to account for it with policies that prohibit the use of slurs with the intent to "incite or promote hatred."[115] Meta says its rules leave space for individuals to share content that contains slurs or hate speech to condemn it or spread awareness, and that it permits individuals to sometimes use certain violating words, such as slurs, if they are used "self-referentially or in an empowering way."[116] However, this breathing room is limited in practice, as it requires individuals to "clearly indicate their intent," which imposes a rigid manner of speaking disconnected from casual forms of online discourse.[117] If there is ambiguity, the company warns the content may be removed.[118]

Colorblind content policy limits the ability of racialized people to speak openly about the experience of oppression. This type of content can take many forms. Some people do the work of educating a broad audience, some speak internally to communities of color, while others speak out angrily against their lived reality using inflammatory language.[119] At other times, individuals simply share hate speech they receive from other people to raise awareness or to condemn it.[120] From education to exasperation, whiteness emerges as an essential framing for understanding racism. But due to its use of explicit racial language, reference to whiteness are likely to run afoul of colorblind content policies.

No matter the approach, the use of racially explicit language differs from comparable language to discuss historically marginalized groups. Breathing room for "[e]xpressions of hated, revulsion, and anger directed against members of historically dominant groups by subordinated-group members" as a natural outgrowth of discrimination makes the harm substantially different from similar

[115] *YouTube Hate Speech Policy*, *supra* note 91.

[116] *Meta Hate Speech Policy*, *supra* note 91.

[117] *Id.*

[118] *Id.*

[119] *See* Monica Anderson, *Social Media Conversations About Race*, PEW RSCH. CTR. (Aug. 15, 2016), https://www.pewresearch.org/internet/2016/08/15/social-media-conversations-about-race/ [https://perma.cc/PGD6-UX73].

[120] *See, e.g.*, Renata Avila, *Fighting Racism and Hate Speech with Community Solutions*, DW AKADEMIE (Nov. 6, 2020), https://akademie.dw.com/en/fighting-racism-and-hate-speech-with-community-solutions/a-55522082 [https://perma.cc/53UN-82DC] (stating tactics for combatting racial hate speech by communities include "de-amplification, de-monetization, education, counter-speech, reporting and training").

language used against racially marginalized groups.[121] Nonetheless, the use of explicit terms like "whiteness" or "white people" can lead to content removal or disfavored treatment by ranking algorithms based on colorblind policies. Communities of color, at the vanguard of online language creation, often use variations like "yt," "pale," and "saltine" for comic effect as well as to avoid content moderation practices.[122]

While some instances of in-group speech become widespread Internet slang, others are co-opted for repressive ends. For decades, dog whistles have allowed politicians to use coded language that avoids explicit bigotry but taps into implicit racial anxieties.[123] Recently, "stay woke"—an in-group reminder for Black people to remain vigilant against the pervasive nature of American racism—has increasingly become the latest conservative dog whistle.[124] Governor Ron DeSantis proclaimed that "Florida is where woke goes to die."[125] On social media, people associate wokeness with "entitlement" and "anti-normal."[126] Executives at social media companies have also joined the fray, with Twitter owner Elon Musk saying, "The woke mind virus is either defeated or nothing else matters."[127]

Replacing the word "woke" with "Black" provides insight into the dual nature of this co-opted term, and how it functions as a covert slur. Decrying "wokeness"

[121] MATSUDA ET AL., *supra* note 48, at 38.

[122] *See, e.g.*, Lorenz, *supra* note 52 (algorithms leading people to say "saltines" when "literally talking about crackers"); @caileneasely, TIKTOK (Apr. 1, 2022), https://www.tiktok.com/@caileneasely/video/7081836877002575146.

[123] *See* IAN HANEY LÓPEZ, DOG WHISTLE POLITICS: HOW CODED RACIAL APPEALS HAVE REINVENTED RACISM AND WRECKED THE MIDDLE CLASS 130 (2014); Adam R. Shapiro, *The Racist Roots of the Dog Whistle*, WASH. POST (Aug. 21, 2020, 6:00 AM), https://www.washingtonpost.com/outlook/2020/08/21/racist-roots-dog-whistle/.

[124] "Stay woke" is an expression with roots dating back to 1930s Black nationalism through the civil rights era. The modern iteration incorporates elements of hip-hop and Black online culture, focusing on Black consciousness. *See* Bijan C. Bayne, Opinion, *How "Woke" Became the Least Woke Word in U.S. English*, WASH. POST (Feb. 2, 2022, 4:42 PM), https://www.washingtonpost.com/opinions/2022/02/02/black-history-woke-appropriation-misuse/.

[125] Emily Mae Czachor, *"Florida Is Where Woke Goes To Die," Gov. Ron DeSantis Says After Reelection Victory*, CBS NEWS (Nov. 9, 2022, 2:33 PM), https://www.cbsnews.com/news/ron-desantis-florida-where-woke-goes-to-die-midterm-election-win/ [https://perma.cc/XUS7-5Y73].

[126] *See, e.g.*, Shane Kidwell, *Wokeness and Cancel Culture Have No Place in Schools*, LINKEDIN (July 16, 2021), https://www.linkedin.com/pulse/wokeness-cancel-culture-have-place-schools-shane-kidwell/ [https://perma.cc/4DWQ-7EW6] (discussing beliefs about wokeness and entitlement); Paul Cameron, *Are We Being 'Polluted' By Gay = Normal?*, RENEWAMERICA (May 7, 2023), https://www.renewamerica.com/columns/cameron/230507 [https://perma.cc/2GYN-P2CB] (stating "[*w*]*okeness* . . . is a *vigorously* pro-homosexual/anti-heterosexual, pro-black/anti-white, pro-trans/anti-normal, pro-feeling/anti-rational, pro-equality/anti-achievement, pro-female/anti-male philosophy").

[127] Elon Musk (@elonmusk), TWITTER (Dec. 12, 2022, 7:25 AM), https://twitter.com/elonmusk/status/1602278477234728960 [https://perma.cc/FW7N-2DVE].

allows people to express (or leverage) fears, anxieties, pleasures,[128] and prejudices without being labeled a racist. In effect, "woke" functions as a stand-in for racialized understandings of otherness and disruption. For example, "stay woke" breaks from traditional English grammar rules, evoking incongruity with the "natural" order and allowing individuals to use the term with mocking derision that perpetuates longstanding stereotypes of Black ignorance. At the same time, the term represents in-group coordination among Black people, tapping into longstanding fears of communication that dominant groups cannot understand, undermining their authority.[129] Finally, "woke" is a term that leverages racial bigotry to demonize progressive demands for civil rights and wealth redistribution.[130] Much like the phrase "welfare queen" was used to attack public benefits, "wokeness" evokes racism to attack several progressive demands even if the beneficiaries extend well beyond communities of color.[131] The effectiveness of woke-as-slur resides in its ability to avoid tapping into explicit bigotry—the only bigotry that might drive social media companies to act. Free from moderation, attacking wokeness is embraced by everyone, from conservative politicians to militias and white nationalists, who leverage this anxiety into protests at school boards, libraries, and drag brunches.[132] The inability for an active and ongoing accounting of how race and bigotry operate allows this ongoing animosity to escalate unchecked.

b. *Racialized Harassment*

Harassment is an enduring reality of online life.[133] According to the Pew Research Center, 41% of U.S. adults have experienced online harassment, with women being more than twice as likely to report that their most recent

[128] *See* Brock, *supra* note 32, at 31-34 (analyzing role of libidinal economy in digital communication, and pleasure certain communities derive from anti-Black racism). I argue part of the allure of "woke" is the pleasure certain people derive from using a word that carries the derision of a slur without the attendant public backlash that would follow more publicly derided terms.

[129] Aja Romano, *A History of "Wokeness"*, VOX (Oct. 9, 2020, 10:00 AM), https://www.vox.com/culture/21437879/stay-woke-wokeness-history-origin-evolution-controversy [https://perma.cc/UTH9-MBCL].

[130] Samuel L. Perry & Eric L. McDaniel, *Why "Woke" Is a Convenient Republican Dog Whistle*, TIME (Jan. 26, 2023, 8:00 AM), https://time.com/6250153/woke-convenient-republican-dog-whistle/ [https://perma.cc/4NNN-XMUM].

[131] *Id.*

[132] *See* Conor Murray, *Definitive Guide to the Anti-'Woke' Protests: From Bud Light to Target to the U.S. Navy—and Everyone Else*, FORBES (May 24, 2023, 3:37 PM), https://www.forbes.com/sites/conormurray/2023/05/20/far-right-pundits-are-slamming-companies-including-nike-adidas-and-ford-for-lgbtq-outreach-as-pride-month-nears.

[133] For an analysis of online harassment, see generally DANIELLE KEATES CITRON, HATE CRIMES IN CYBERSPACE (2014); EMILY A. VOGELS, THE STATE OF ONLINE HARASSMENT (Jan. 13, 2021), https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2021/01/PI_2021.01.13_Online-Harassment_FINAL-1.pdf [https://perma.cc/KS6A-DD75].

experience was very or extremely upsetting.[134] Roughly half of Black and Hispanic individuals subjected to online harassment reported it being tied to their race or ethnicity, compared with 17% of white targets.[135] Amnesty International has tracked Twitter's efforts to address online abuse against women since 2017, reporting time and again that the platform has failed to adequately protect women—particularly those with intersectional identities—noting LGBTQ+ women, women with disabilities, and women from ethnic or religious minorities are disproportionately harmed by abuse.[136]

Online harassment can cause offline harms ranging from doxxing (publicly revealing private information about an individual with malicious intent) to violence, but it can also lead to online harms, such as causing people to withdraw from social media or to self-censor around certain topics.[137] As a report from PEN America notes, individual harms stemming from harassment "have systemic consequences: undermining the advancement of equity and inclusion, constraining press freedom, and chilling free expression."[138]

Platform rules against harassment acknowledge the impact of online harassment but take divergent and vague approaches to balancing it against freedom of expression. For example, Meta says harassment "prevents people from feeling safe," but that it also wants to ensure people can share "critical commentary of people who are featured in the news or who have a large public audience," positing the two values as oppositional.[139] YouTube's rules do not provide a policy rationale, but its exception for "[d]ebates related to high-profile officials or leaders" suggests that it too attempts to balance open debate and individual safety.[140] Twitter, on the other hand, says people "should feel safe expressing [their] unique point[s] of view" and prohibits harassment "to facilitate healthy dialogue on the platform"; it weighs "freedom of

[134] VOGELS, *supra* note 133, at 4.

[135] *Id.* at 9.

[136] *See Twitter Still Failing Women over Online Violence and Abuse*, AMNESTY INT'L (Sept. 22, 2020), https://www.amnesty.org/en/latest/news/2020/09/twitter-failing-women-over-online-violence-and-abuse [https://perma.cc/9QYG-RLKZ].

[137] *See* JILLIAN C. YORK, SILICON VALUES: THE FUTURE OF FREE SPEECH UNDER SURVEILLANCE CAPITALISM 196 (2021).

[138] Viktorya Vilk, Elodie Vialle & Matt Bailey, *No Excuse for Abuse: What Social Media Companies Can Do Now To Combat Online Harassment and Empower Users*, PEN AM., https://pen.org/report/no-excuse-for-abuse [https://perma.cc/8JCA-ETN8] (last visited Nov. 9, 2023).

[139] *Bullying and Harassment*, META TRANSPARENCY CTR. [hereinafter *Meta Harassment Policy*] https://transparency.fb.com/policies/community-standards/bullying-harassment/ [https://perma.cc/RU6C-HBZ3] (last visited Nov. 9, 2023).

[140] *Harassment & Cyberbullying Policies*, YOUTUBE HELP [hereinafter *YouTube Harassment Policy*], https://support.google.com/youtube/answer/2802268 [https://perma.cc/AS7E-UUTB] (last visited Nov. 9, 2023).

expression"[141] against the platform's interest in "direct access to powerful figures, and maintaining a robust public record [that] provides benefits to accountability."[142] According to Twitter, "insults may be present in tweets related to heated debate over matters of public policy," but the company is more likely to remove a tweet targeting "a private individual without . . . relevant political context."[143] Platforms rightfully provide more breathing room for criticism of public figures, but only Twitter specifies who qualifies as a public figure and addresses how these same individuals can have an outsized ability to prompt their followers to harass individuals.[144]

While platform harassment policies incorporate a wide range of prohibitions, they provide considerable discretion for platforms to choose when and how they will act—a departure from typical platform attempts to move toward easily administrable rules. For example, YouTube prohibits not only direct threats and incitement but also "repeatedly" engaging in abusive behavior or "inciting hostility between creators," and Meta prohibits repeatedly making unwanted advances and bullying.[145] Similarly, Twitter prohibits continually posting unreciprocated posts "with malicious content, to target an individual."[146] These terms are not well defined, and none of the platforms specify what level of repeated or aggressive attacks merit action. Similarly, while platforms acknowledge the importance of context and assessing patterns of behavior, it remains unclear how platform moderators acquire the necessary context to make accurate determinations. Instead, Twitter and Meta's policies say in "certain" circumstances, they may need to hear from the person being targeted to understand "context and intent" or to "understand that the person targeted feels bullied or harassed."[147] This flexibility, if combined with clear standards that account for power dynamics, could chart an alternate path forward.

Despite doing a more thorough job incorporating flexibility to account for the myriad harms that attach to harassment, content policy makes little attempt to connect harassment policy with race, gender, and sexuality. By failing to connect harassment with existing systems of social stratification, platforms leave themselves ill-equipped to react to attacks that are initially developed against

[141] *Abuse and Harassment*, TWITTER [hereinafter *Twitter Harassment Policy*] (June 2023), https://help.twitter.com/en/rules-and-policies/abusive-behavior [https://perma.cc/PK44-SMDZ].

[142] *About Public-Interest Exceptions on Twitter*, TWITTER, https://perma.cc/BF6A-DF UB (last visited Nov. 9, 2023).

[143] *About Public-Interest Exceptions on Twitter*, *supra* note 142.

[144] *See id.*

[145] *YouTube Harassment Policy*, *supra* note 140; *see Meta Harassment Policy*, *supra* note 139.

[146] *Twitter Harassment Policy*, *supra* note 142.

[147] *Meta Harassment Policy*, *supra* note 139; *see Twitter Harassment Policy*, *supra* note 142 ("To help our teams understand the context, we sometimes need to hear directly from the person being targeted to ensure that we have the necessary information prior to taking any enforcement action.").

marginalized groups (women of color) before being expanded to more general audiences.

For example, YouTube is the only platform protecting people targeted with "prolonged" or "malicious insults" based on their protected characteristic.[148] However, the company ignored the realities of racist harassment, despite an awareness of how many of the platform's most popular creators profited from racism and often made videos indirectly encouraging their followers to harass people of color, including other creators.[149]

In late May 2019, Carlos Maza, a former *Vox* journalist, detailed ongoing racist and homophobic harassment he faced from popular far-right YouTuber Steven Crowder and his followers.[150] Crowder repeatedly harassed Maza with language that moved between explicit and implicit racism popular on the platform.[151] Crowder called Maza a "lispy sprite," a "little queer," "Mr. Gay Vox," and "gay Mexican" in videos.[152] Maza reported that these attacks led to ongoing harassment from Crowder's followers, including doxxing, death threats, and harassment via a torrent of phone calls and texts.[153] In response, Crowder posted a video where he denounced the harassment carried out by his followers and claimed his comments about Maza were mere political humor.[154] YouTube agreed and posted a response on Twitter saying that while Crowder's language "was clearly hurtful, the videos as posted don't violate our policies."[155] YouTube again justified its protection of Steven Crowder's racism and

[148] *YouTube Harassment Policy*, *supra* note 140.

[149] *See* BERGEN, *supra* note 55, at 9, 294 (detailing anti-Semitic elements in videos by creator PewDiePie and harassment of Black creators by other creators).

[150] Julia Alexander, *YouTube Investigating Right-Wing Pundit Steven Crowder for Harassing Vox.com Host*, VERGE (May 31, 2019, 9:23 PM), https://www.theverge.com/2019/5/31/18647621/youtube-steven-crowder-bullying-harassment-twitter-vox-carlos-maza [https://perma.cc/8ZLM-4H65] (linking Maza's thread on Twitter describing harassment).

[151] As of August 2023, Crowder has nearly six million subscribers to his YouTube channel. *See* Steven Crowder (@StevenCrowder), YOUTUBE, https://www.youtube.com/@StevenCrowder [https://perma.cc/6H3W-ZHC4] (last visited Nov. 9, 2023).

[152] Eli Rosenberg, *A Right-Wing YouTuber Hurled Racist, Homophobic Taunts at a Gay Reporter. The Company Did Nothing.*, WASH. POST (June 5, 2019, 3:57 PM), https://www.washingtonpost.com/technology/2019/06/05/right-wing-youtuber-hurled-racist-homophobic-taunts-gay-reporter-company-did-nothing/.

[153] Alexander, *supra* note 150; Kevin Roose, *A Thorn in YouTube's Side Digs Even Deeper*, N.Y. TIMES (Feb. 12, 2020), https://www.nytimes.com/2020/02/12/technology/carlos-maza-youtube-vox.html.

[154] Alexander, *supra* note 150.

[155] Nick Statt, *YouTube Decides That Homophobic Harassment Does Not Violate Its Policies*, VERGE (June 4, 2019, 8:55 PM), https://www.theverge.com/2019/6/4/18653088/youtube-steven-crowder-carlos-maza-harassment-bullying-enforcement-verdict [https://perma.cc/F5GQ-HZHK].

incitement because he did not *explicitly* call on his followers to harass Maza.[156] It was only after a critical mass of civil society, company employees, politicians, and industry executives added public pressure that YouTube revised its position and temporarily demonetized Crowder's YouTube channel.[157] But this was not an explicit walk back of the decision; instead, the company acted because of Crowder's sale of "Socialism is for F*gs" T-shirts on his channel.[158] The company said that "further investigation" of Crowder's channel revealed a "pattern of egregious actions" that harmed an undefined "broader community."[159] The public fallout from Maza's experience prompted the company to revise their harassment policy, which the company unveiled a few months later in December 2019.[160] This expanded policy contained a new creator-on-creator harassment policy prohibiting "demeaning language that goes too far."[161]

The post-hoc policy change to address behavior already falling within the scope of the platform's existing content standards exemplifies the ways new policies are part of a corporate strategy to justify belated enforcement. YouTube's policy and enforcement mechanism gave the company wide latitude to intervene and protect people from racist attacks. Of the major platforms, YouTube engages in lucrative revenue sharing with content creators and has a process for keeping popular creators abreast of policy changes.[162] This relationship with some of the most popular creators in practice means the company is more aware than most about the accounts trading in everything from casual racism to explicit white supremacy.[163] But the popularity of the content, combined with conservative regulatory pressure and platform leadership's

[156] Benjamin Goggin, *YouTube's Week from Hell: How the Debate over Free Speech Online Exploded After a Conservative Star with Millions of Subscribers Was Accused of Homophobic Harassment*, BUS. INSIDER (June 9, 2019, 1:31 PM), https://www.businessinsider.com/steven-crowder-youtube-speech-carlos-maza-explained-youtube-2019-6.

[157] Julia Alexander, *YouTube Revokes Ads from Steven Crowder Until He Stops Linking to His Homophobic T-shirts*, VERGE (June 5, 2019, 3:07 PM), https://www.theverge.com/2019/6/5/18654196/steven-crowder-demonetized-carlos-maza-youtube-homophobic-language-ads [https://perma.cc/G3MM-WUXB].

[158] *Id.*

[159] *Id.*

[160] *See* Matt Halprin, *An Update to Our Harassment Policy*, YOUTUBE OFF. BLOG (Dec. 11, 2019), https://blog.youtube/news-and-events/an-update-to-our-harassment-policy [https://perma.cc/4KDR-Y42S].

[161] *Id.*

[162] *YouTube Partner Program Overview & Eligibility*, YOUTUBE HELP, https://support.google.com/youtube/answer/72851 [https://perma.cc/68DN-W6EJ] (last visited Nov. 9, 2023) (describing channels are continuously checked for compliance with policies on eligibility for revenue sharing and advertising over time).

[163] *See* BERGEN, *supra* note 55, at 380 ("A colleague on [the violent extremism] team once confessed . . . they were so swamped with material that they rarely touched videos marked as white supremacist.").

willingness to excuse racism as heated humor, results in platform inaction in the face of rising attacks against communities of color.[164]

In this way, financial interest, regulatory fears, and gut intuition collectively power platform decisions to ensure their content policies reinforce existing racial hierarchies. Severing the ways in which harassment is powered by bigotry sets up a foreseeably underdeveloped approach. Not only does it fail to appropriately weigh the dangers of malicious attacks cloaked in humor, but it sets up a context-neutral enforcement system that ends up shielding powerful figures from marginalized communities seeking accountability.

2. Violent Hate Speech and Calls for Racial Inferiority, Exclusion, or Supremacy

The second approach to policing hate speech is through prohibitions against violent incitement and claims of racial superiority. Prohibitions against violent hate speech are universal, reflecting platform commitments to preventing affiliation with offline violence.[165] But the narrow application creates an overly restrictive order of operations. There must be an explicit call for violence as well as a reference to a specific racial group. For example, "death to all (racial group)."

Similarly, prohibiting claims that one race is superior to another, or that members of a particular race are less intelligent, capable, or should be excluded, would seem to cut to the core of white supremacy. Attacking the cultural myths that justify racial hierarchies is an important intervention point. Combatting these tropes is an ongoing struggle for communities of color, as white supremacy is a mutable monster that finds new manifestations with every generation.[166] But this rule is narrower than it appears. Content policies typically require both an explicit mention of race and explicit claims regarding decreased intelligence, capabilities, or specific calls for segregation.[167]

If the goal is to prevent offline violence, rules that fail to consider the role of race and racism are set up to fail. Ignoring the history and context through which

[164] *Id.* at 379-81 (describing how political risk and prioritization prevented revisions to hate speech policy).

[165] *See, e.g.*, *Meta Hate Speech Policy*, *supra* note 91 (prohibiting direct attacks against people on basis of protected characteristics); *YouTube Hate Speech Policy*, *supra* note 91 (prohibiting content promoting violence against people based on protected attributes); *Twitter Hate Speech Policy*, *supra* note 91 (prohibiting inciting behavior targeting people belonging to protected categories).

[166] From phrenology to poll taxes to risk assessments, there is no shortage of permutations used to justify racially discriminatory treatment. *See generally* DOROTHY ROBERTS, FATAL INVENTION: HOW SCIENCE, POLITICS, AND BIG BUSINESS RE-CREATE RACE IN THE TWENTY-FIRST CENTURY (2011); Jessica Eaglin, *When Critical Race Theory Enters the Law & Technology Frame*, 25 MICH. J. RACE & L. 151 (2021); MICHELLE ALEXANDER, THE NEW JIM CROW: MASS INCARCERATION IN THE AGE OF COLORBLINDNESS 37 (3d ed. 2020); RICHARD ROTHSTEIN, THE COLOR OF LAW: A FORGOTTEN HISTORY OF HOW OUR GOVERNMENT SEGREGATED AMERICA (2017).

[167] *See, e.g.*, sources cited *supra* note 91.

race is socially created prevents these rules from being meaningful checks against racist attacks. A failure to understand the modern operation of racism sets the intervention point at a level of naked racism that is disconnected from its current manifestation. While this may be a principled choice in favor of freedom of expression, it more closely represents a refusal to see how racism operates because it reveals an inconvenient truth with how entwined it is with political speech. At best, it leaves platforms in their existing reactive stance. This is reflected through the reliance on "designate[d] . . . violent events" such as white supremacist mass shootings or the January 6th insurrection.[168] The designation does not operate within hate speech policies, instead leveraging broad removal tools saved for moderation based on violent extremism policy.

Social media content moderation during the 2020 uprisings are a manifestation of this approach in practice. The murders of George Floyd and Breonna Taylor, the culmination of more than a decade's worth of documented and unaccountable police killings, contributed to the largest outgrowth of solidarity in years.[169] The Movement for Black Lives and other activists' digital organization, along with the distance required by a global pandemic, made social media's role as the fabled public square[170] of central importance.

Aside from one of the central cries—"Black Lives Matter"—demands such as "defund the police" did not make explicit reference to race. Similarly, the opposition used language that was not racially explicit. Indeed, the race-neutral language used was not even particularly new. Calls against "thugs," "looters," and "rioters" are long-established dog whistles to attack Black people without being considered racist.[171] But as this coalition of social justice protests swelled beyond Black neighborhoods, the anxiety it gave white communities found new purchase. For example, the Kenosha Guard, a militia organization, organized

[168] *Dangerous Organizations and Individuals*, META TRANSPARENCY CTR. [hereinafter *Meta DOI Policy*], https://transparency.fb.com/policies/community-standards/dangerous-individuals-organizations/ [https://perma.cc/Q4PK-KLD9] (last visited Nov. 9, 2023); *see* Guy Rosen, *Our Response to the Violence in Washington*, META NEWSROOM (Jan. 7, 2021, 11:05 AM), https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/ [https://perma.cc/BPR2-95NW] (removing posts supporting January 6th insurrection through reactionary "emergency" measures instead of ongoing and preventative approach understanding nature of racial hate speech and implicit calls for violence).

[169] *See* Amna A. Akbar, *An Abolitionist Horizon for (Police) Reform*, 108 CAL. L. REV. 1781, 1783 ("Then came the 2020 uprisings following the police murder of George Floyd in Minneapolis, among the largest social movement mobilizations in U.S. history."); Larry Buchanan, Quoctrung Bui & Jugal K. Patel, *Black Lives Matter May Be the Largest Movement in U.S. History*, N.Y. TIMES (July 3, 2020), https://www.nytimes.com/interactive/2020/07/03/us/george-floyd-protests-crowd-size.html.

[170] *See* Packingham v. North Carolina, 582 U.S. 98, 104 (2017) (comparing social media to streets and parks as essential venues for public gatherings and speech).

[171] *See* LÓPEZ, *supra* note 123, at 130.

calls to bring weapons to meet social justice protesters.[172] In one post, the group wrote "Any patriots willing to take up arms and defend our city tonight from the evil thugs? No doubt they are currently planning on the next part of the city to burn tonight."[173] The post was reported by people on the platform before any escalation in violence, but Meta chose not to act.[174]

One of the most infamous individuals to heed these calls was Kyle Rittenhouse, who traveled from his home in Illinois to Kenosha, Wisconsin to partake in the defense.[175] Over the course of one night, Rittenhouse shot three people, killing two and injured another.[176] The following day, Meta designated the event "a mass murder," removed Rittenhouse's accounts from Facebook and Instagram, blocked his name from appearing in search, and claimed it would also remove posts praising and supporting Rittenhouse.[177] The company also announced a "new" policy addressing militia organizations, which captured the Kenosha Guard Group and Event Page.[178] In a subsequent call with company employees, Zuckerberg faced questions regarding the company's failure to grapple with the proliferation of hate speech, anti-Semitism, and white supremacy on the platform.[179] Company leadership blamed the slowness to act on content moderators' unawareness of "how certain militias" operate.[180]

[172] *See* Russell Brandom, *Facebook Takes Down "Call to Arms" Event After Two Shot Dead in Kenosha*, VERGE (Aug. 26, 2020, 11:49 AM), https://www.theverge.com/2020/8/26/21402571/kenosha-guard-shooting-facebook-deplatforming-militia-violence [https://perma.cc/AGG7-BX9J] (reporting disappearance of event page organized by Kenosha Guard).

[173] *Id.*

[174] Russell Brandom, *Facebook Chose Not To Act on Militia Complaints Before Kenosha Shooting*, VERGE (Aug. 26, 2020, 5:15 PM), https://www.theverge.com/2020/8/26/21403004/facebook-kenosha-militia-groups-shooting-blm-protest [https://perma.cc/KZ5J-5VSE].

[175] *See* Haley Willis et al., *Tracking the Suspect in the Fatal Kenosha Shootings*, N.Y. TIMES (Nov. 22, 2021), https://www.nytimes.com/2020/08/27/us/kyle-rittenhouse-kenosha-shooting-video.html (detailing Rittenhouse's movements before and during protests).

[176] Vanessa Romo & Sharon Pruitt-Young, *What We Know About the 3 Men Who Were Shot by Kyle Rittenhouse*, NPR (Nov. 20, 2021, 8:56 PM), https://www.npr.org/2021/11/20/1057571558/what-we-know-3-men-kyle-rittenhouse-victims-rosenbaum-huber-grosskreutz [https://perma.cc/7GY5-95U3].

[177] Brian Fishman (@brianfishman), TWITTER (Aug. 27, 2020, 4:40 PM), https://twitter.com/brianfishman/status/1299084287686434816 [https://perma.cc/ED8R-Q34H%5D] (describing actions taken by Facebook's former head of Dangerous Organizations and Individuals).

[178] Brian Fishman (@brianfishman), TWITTER (Aug. 27, 2020, 4:42 PM), https://twitter.com/brianfishman/status/1299084786112307200 [https://perma.cc/X3NQ-V6AZ%5D] (reporting Kenosha Guard's violation of new policy).

[179] Ryan Mac, *Facebook Employees Are Outraged at Mark Zuckerberg's Explanations of How It Handled the Kenosha Violence*, BUZZFEED NEWS (Aug. 28, 2020, 3:10 PM), https://www.buzzfeednews.com/article/ryanmac/facebook-employees-slam-zuckerberg-kenosha-militia-shooting [https://perma.cc/R5E3-S67R].

[180] *Id.*

Contemporaneous reports found numerous instances of individuals expressing solidarity with Rittenhouse, calling him a "patriot."[181]

The decision to use a new policy disconnected from racism explains the delay, but it reflects the company's ongoing refusal to acknowledge the role of racism in the attack, the rise of militias, or the racist response to social justice protests. It also suggests this explicit prohibition against incitement to violence is of little practical effect so long as explicit targeting is not made. Moderation of Rittenhouse's posts based on Meta's Dangerous Individuals and Organizations ("DIO") policy may have allowed them to justify reversing their decision after Rittenhouse was later found not guilty by a jury.[182] The swift reversal of the ban is likely connected to the mass right-wing support for Rittenhouse, and their criticism that Meta's actions were premature.[183] Similarly swift reversals are rare, and there is no policy specifying that acquittal by a jury triggers removal from the potential for eliciting violence. There are numerous individuals no longer on state Foreign Terrorist Organization blacklists who have not yet been restored by platforms.[184]

B. *Racialized Threat Assessments*

Terrorism and violent extremism ("TVE") policies represent a different approach to content moderation. Whereas hate speech and harassment policies use a narrow approach to protect freedom of expression, TVE policies rely on broad enforcement that flattens nuance in the name of eradication. This tactic notes no difference among speakers and severely restricts the scope of acceptable discourse.

For the most part, TVE policies do not explicitly consider race. Instead, race becomes most visible through the groups and individuals that are targeted, based on platform assessments of dangerousness. This dangerousness is largely informed by Western government threat assessments and attendant compliance

[181] Katie Paul, *Praise for Wisconsin Shooter Shared Widely on Facebook Despite Ban*, REUTERS (Sept. 2, 2020, 10:43 PM), https://www.reuters.com/article/global-race-facebook-idINKBN25U09A [https://perma.cc/HRG9-Z8JL] (reporting posts supporting Rittenhouse were "racking up thousands of shares").

[182] James Clayton, *Facebook Reverses Kyle Rittenhouse Policy*, BBC (Dec. 1, 2021), https://www.bbc.com/news/technology-59486397 [https://perma.cc/Z4W4-UTVW] (noting Meta will "still remove content that celebrates the death of the individuals killed in Kenosha," but "will no longer remove content containing praise or support of Rittenhouse").

[183] *Id.*

[184] Faiza Patel & Mary Pat Dwyer, *So, What Does Facebook Take Down? The Secret List of 'Dangerous' Individuals and Organizations*, BRENNAN CTR. FOR JUST. (Nov. 8, 2021), https://www.brennancenter.org/our-work/analysis-opinion/so-what-does-facebook-take-down-secret-list-dangerous-individuals-and [https://perma.cc/4YMK-GAZ6] (reporting Facebook DIO blacklists include Houthis and many affiliates "despite the Biden Administration's removal of the group's [Foreign Terrorist Organization] designation in February 2021").

issues, but it is also shaped by the internalized biases of platforms themselves.[185] This system's racialized harms are no less discriminatory because they occur without traditional markers of racial animus. In fact, steadfast commitment to this discriminatory system despite known harms inflicts a separate badge of inferiority that has gone largely unacknowledged.[186] At the same time, deferential policies for policing militant far-right content demonstrate how platforms view overbroad enforcement as unacceptable for some groups but a necessary tradeoff for others. Within this racialized framework, no value holds supreme, even bans against explicit calls for violence.

This Part proceeds in two sections. First, this Part analyzes platforms' TVE policies. The selective deployment of vague and narrow terms constructs a vision of race and protects the fostering of racism. Whereas marginalized groups face censorship of all but the most unambiguous attempts to decry certain political figures or neutrally document violence, white supremacy is protected in instances aside from *explicit* praise or representation, protecting a wide range of cultural and political expression that upholds and perpetuates the same ideology. Second, this Part analyzes the secret list of individuals and organizations covered by Meta's DIO policy to explore not only how it treats groups, but also how this vision upholds Western notions of national security. These lists and rules appear designed with a unilateral focus on Muslim extremism and an expanding interest in combating gangs and cartels.

1. Policing Racialized Threat Assessments

After years of resisting calls to remove "terrorist" speech, the major social media platforms conceded to pressure from the U.S. and European governments to aggressively remove content deemed as supporting terrorism.[187] Largely shaped by Western government calls to launch an "offensive" against Islamic State of Iraq and Syria ("ISIS") propaganda, TVE policies disproportionately target speech from Muslim and Arabic-speaking communities.[188]

[185] *See infra* Section III.B.1 (analyzing TVE policies' disproportionate focus on Middle East and North Africa).

[186] Several decades ago, critical race theorist Charles Lawrence III conceptualized *Brown v. Board of Education*, 347 U.S. 483 (1954) as a case about the harms of hate speech and the Court's ruling as a rejection of the "defamatory symbolism of segregation." *See* MATSUDA ET AL., *supra* note 48, at 9.

[187] KATRIEN LUYTEN, EURO. PARLIAMENTARY RSCH. SERV., ADDRESSING THE DISSEMINATION OF TERRORIST CONTENT ONLINE 2 (2021), https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/649326/EPRS_BRI(2020)649326_EN.pdf [https://perma.cc/MGU7-PQ5X].

[188] Amar Toor, *France Wants Facebook and Twitter To Launch an "Offensive" Against ISIS Propaganda*, VERGE (Dec. 3, 2015, 11:38 AM), https://www.theverge.com/2015/12/3/9842258/paris-attacks-facebook-twitter-google-isis-propaganda [https://perma.cc/GCJ6-RUBT].

TVE policies typically remove content that "affiliate[s] with or promote[s] the activities of violent and hateful entities."[189] Twitter and YouTube do not define their prohibitions, instead providing a handful of examples for the types of content that might constitute glorification, promotion, affiliation, and support.[190] Meta is the only platform that attempts to define these terms, acquiescing to years of pressure from civil society and a policy recommendation from its self-regulatory Oversight Board.[191]

Meta defines "praise" as speaking positively, providing a sense of achievement, legitimizing a cause by making claims it is somehow legally, morally, or otherwise justified, or aligning oneself ideologically.[192] This definition is informative to the extent its capaciousness is laid bare. Not only are each of these definitions open to interpretation, but it also captures a wide range of opinions.

In practice, prohibitions against "praise" or "glorification" carry the load of discretionary enforcement, while "material support" and "representation" provide easier but less frequent removals. Platforms do not break down TVE removals by type. However, the Global Internet Forum to Counter Terrorism ("GIFCT"), of which each platform is a founding member, does.[193] The GIFCT reports, which date back to 2019, consistently show that glorification accounts for most of the hashes in the database.[194] As with the platforms' policies, GIFCT's approach is vague: glorification is defined as content that "glorifies, praises, condones, or celebrates attacks after the fact."[195] However, unlike the

[189] *Violent and Hateful Entities Policy*, TWITTER HELP CTR. [hereinafter *Twitter TVE Policy*] (Apr. 2023), https://help.twitter.com/en/rules-and-policies/violent-entities [https://perma.cc/9869-LH6S].

[190] *See id.* (describing "[e]xamples of the types of content that violate this policy"); *Violent Extremist or Criminal Organizations Policy*, YOUTUBE HELP, https://support.google.com/youtube/answer/9229472 [https://perma.cc/TGM6-2UVB] (last visited Nov. 9, 2023) [hereinafter *YouTube TVE Policy*] (describing "examples of content that's not allowed on YouTube").

[191] *See Meta DOI Policy*, *supra* note 168; *Case Decision 2020-005-FB-UA*, OVERSIGHT BD. (June 12, 2023), https://transparency.fb.com/oversight/oversight-board-cases/nazi-quote [https://perma.cc/VCS5-SRTK].

[192] *Meta DOI Policy*, *supra* note 168.

[193] GLOB. INTERNET F. TO COUNTER TERRORISM, TRANSPARENCY REPORT JULY 2021, at 9-10 (2021), https://gifct.org/wp-content/uploads/2021/07/GIFCT-Transparency Report2021.pdf [https://perma.cc/MUR6-K4J6] (explaining categories of terrorist content in hash-sharing database).

[194] *Id.* at 10 (listing glorification of terrorist attacks at 77.2% of total hashes); GLOB. INTERNET F. TO COUNTER TERRORISM, TRANSPARENCY REPORT JULY 2020, at 4 (2020), https://gifct.org/wp-content/uploads/2020/10/GIFCT-Transparency-Report-July-2020-Final.pdf [https://perma.cc/3NJ7-8SVJ] (listing at 72%); GLOB. INTERNET F. TO COUNTER TERRORISM, TRANSPARENCY REPORT 2019, at 3 (2019), https://gifct.org/wp-content/uploads/2020/10/GIFCT-Transparency-Report-July-2019-Final.pdf [https://perma.cc/2X4Z-RZZK] (listing at 85.5%).

[195] GLOB. INTERNET F. TO COUNTER TERRORISM, *supra* note 193.

secrecy that shrouds platform designations, GIFCT claims it primarily uses the UN Security Council's Consolidated Sanctions List.[196]

The definitions of "praise" or "glorification" are imprecise, setting up a system of immense uncertainty. TVE content tends to be very political in nature, reflecting tensions that have boiled over into violence. While platforms attempt to provide some breathing room for this broad prohibition, it is limited to news reporting or posts that "report on, condemn, or neutrally discuss" prohibited persons and entities.[197] This protection is likely limited in practice, since the policy still requires individuals to "clearly indicate their intent" or face removal.[198] This typically results in a narrower vision of acceptable discourse, encompassing only the denouncement or documentation of human rights abuses. This exception is ill-suited to accommodate the messy and unclear reality of political debate. In communications with the Oversight Board, it appears that Meta operationalizes this exception through internal guidance provided to its moderators.[199] But in one instance, Meta claimed it lost the moderator guidance and failed to provide it to moderators.[200] This suggests that this attempt to operationalize a policy might be more window dressing than a substantive protection.

The focus on violence in the Middle East and North Africa sets up a situation where decades of Western occupation have created feelings of general anger at Western forces, potentially fostering sympathy with designated groups and figures. The flattened discourse levels impose a stratified experience that limits debate, education, solidarity, mourning, and joy—the variety of experience that is essential for a vibrant space. For broader diasporas, their ability to similarly contribute to discourse, for which they may have unique insights if not personal connections, also faces intense restriction.

By contrast, platform definitions of substantive support and representation are narrower. Substantive support refers to fundraising, providing material aid, calls to action, recruiting, or "[c]hanneling information" on behalf of a designated group.[201] Nonetheless, this designation still provides ample space for misapplication. What is the difference between news reporting and "channeling information"? Can someone fundraise for the nonillicit aspects of an organization? Twitter's and YouTube's definitions similarly seek to restrict financial support to designated groups.[202] Failure to differentiate between

[196] *Id.* However, the GIFCT did make ad hoc designations for attacks in Christchurch, New Zealand and Halle, Germany, in 2019 and Glendale, Arizona in 2020. *Id.*

[197] *Meta DOI Policy*, *supra* note 168.

[198] *Id.*

[199] *Case Decision 2021-006-IG-UA*, OVERSIGHT BD. (July 8, 2021), https://www.oversightboard.com/decision/IG-I9DP23IB/ [https://perma.cc/B4LR-LRQL].

[200] *Id.* (noting Facebook lost internal guidance on meaning of "support" for three years).

[201] *Meta DOI Policy*, *supra* note 168.

[202] *YouTube TVE Policy*, *supra* note 190; *Twitter TVE Policy*, *supra* note 189.

militant and potentially altruistic arms of groups can interfere with humanitarian or nonviolent aspects of their work.[203]

Meta's definition of "representation," which likely overlaps with other platform definitions of "affiliation," includes claiming membership in a designated entity or operating an account, page, event, or group that purports to represent a designated entity.[204] This designation appears to mostly restrict the ability of TVE entities to maintain a presence on platforms. This limitation, properly scoped, makes sense. But it typically goes beyond removing accounts and attempts to evade enforcement; it also imposes a blackout that can capture broader representations. Representations can capture pictures of an ISIS flag or the face of a designated entity, like a billboard photo of former Iranian general Qasem Soleimani, regardless of their purpose.[205] The indiscriminate removal of representation can cause several harms. First, it can capture content unrelated to the prohibited entity.[206] Requiring that clear denouncement attach to any representation imposes an unworkable standard where the entity may be wholly unrelated to the content. Second, the documentation of human rights abuses may also fail to contain the disclaimers or neutrality necessary to avoid removal.[207] Where is the line between reporting versus representation? The broad ban does not answer these questions or attempt to explain them; the discretion to interpret broadly or narrowly remains with platforms. For example, in multiple reported instances, Meta has erroneously deleted news articles and suspended accounts of journalists and human rights activists, including the accounts of at least thirty-

[203] Twitter's limited restriction that seeks to curb only a designated entity's "illicit" actions, if meaningfully applied, could serve as a model. *See Twitter TVE Policy*, *supra* note 189.

[204] *Meta DOI Policy*, *supra* note 168.

[205] Emily McPherson, *Why Hackers Are Using ISIS Flags To Disable People's Facebook Accounts*, 9NEWS (July 1, 2020, 11:18 AM), https://www.9news.com.au/national/facebook-and-instagram-hackers-why-hackers-are-using-isis-flags-to-disable-peoples-accounts/ca23b836-f6ca-4fd8-b8ad-09e5c3a39120 [https://perma.cc/AK3P-UVFE]; Isobel Cockerell, *Instagram Shuts Down Iranian Accounts After Soleimani's Death*, CODA (Jan. 10, 2020), https://www.codastory.com/authoritarian-tech/instagram-iran-soleimani/ [https://perma.cc/D7JW-SUPS] (noting after Soleimani's death, Instagram removed all Soleimani-related content, including posts critical of Iran).

[206] *See* Jon Porter, *Instagram Blames 'Enforcement Error' for Removal of Posts About Al-Aqsa Mosque*, VERGE (May 13, 2021, 7:54 AM), https://www.theverge.com/2021/5/13/22433861/instagram-al-aqsa-mosque-posts-takedown-error-facebook-moderation [https://perma.cc/4FBP-BYM8] (noting Facebook took down posts concerning protests at Al-Aqsa Mosque because Al-Aqsa Martyrs' Brigades is designated violent entity).

[207] Olivia Solon, *'Facebook Doesn't Care': Activists Say Accounts Removed Despite Zuckerberg's Free-Speech Stance*, NBC NEWS (June 15, 2020, 4:54 PM), https://www.nbcnews.com/tech/tech-news/facebook-doesn-t-care-activists-say-accounts-removed-despite-zuckerberg-n1231110 [https://perma.cc/4MXU-RB2L] (noting Meta deleted posts and accounts of journalists and activists documenting human rights violations in Syrian Civil War).

five Syrian journalists in the spring of 2020 and fifty-two Palestinian journalists and activists in a single day in May 2020.[208]

By choosing to rely on prohibitions of expansive categories like "support" and "glorification," platforms have established a regime in which a wide range of political speech and human rights documentation is inevitably swept up in a removal dragnet. Overall, platform policy regarding terroristic content pays little heed to nuance and context, willingly accepting errors that affect communities with little political power. Hindering the ability of groups to freely express themselves blocks essential parts of lived experience.

Meta's moderation practices regarding conflicts between Israel and Palestine provide an illustrative example of this approach. In May 2021, tensions between Israelis and Palestinians escalated in the face of forced evictions from the Sheikh Jarrah neighborhood of East Jerusalem.[209] Jewish settler groups sought to override Palestinian property interests in the area and make the land available.[210] As the decision made its way through the Israeli court system, protests sprang up throughout the city.[211] The police response fueled clashes, including outside the Al-Aqsa mosque, where Israeli police shot at protesters and worshipers emerging from services concluding the holy month of Ramadan.[212] In response, the militant wing of Hamas launched indiscriminate rocket attacks.[213] Israel responded with missiles of its own.[214]

These volatile weeks had online corollaries for Palestinians in the Gaza Strip, for the broader diaspora, and for the outside world. Within the Gaza Strip, Palestinians shared posts commensurate with the moment. Some documented human rights violations, published statements or infographics, or otherwise engaged in frontline digital activism.[215] Platform responses resulted in the

[208] *Id.*

[209] Reality Check Team, *Sheikh Jarrah: Why Could Palestinians Lose Their Homes in Jerusalem?*, BBC (Aug. 3, 2021), https://www.bbc.com/news/57239690 [https://perma.cc/FY3X-XAG7] (reporting "weeks of protests and clashes" surrounding Sheik Jarrah evictions).

[210] *See id.* (noting Israeli law compels return of Palestinian-occupied land in East Jerusalem to largely Jewish "original owners before 1948," but Palestinians do not have "equal power" to recover lost land in Israel proper).

[211] *Id.*

[212] Yolande Knell, *Al-Aqsa Mosque: Dozens Hurt in Jerusalem Clashes*, BBC (May 8, 2021), https://www.bbc.com/news/world-middle-east-57034237 [https://perma.cc/V78L-TDFR] (stating Israeli police "fired rubber bullets and stun grenades").

[213] Omar Shakir, *Jerusalem to Gaza, Israeli Authorities Reassert Domination*, HUM. RTS. WATCH (May 11, 2021, 6:00 PM), https://www.hrw.org/news/2021/05/11/jerusalem-gaza-israeli-authorities-reassert-domination [https://perma.cc/9QY7-Q8KL].

[214] *Id.*

[215] THE ARAB CTR. FOR SOC. MEDIA ADVANCEMENT, THE ATTACKS ON PALESTINIAN DIGITAL RIGHTS 2 (May 21, 2021), https://7amleh.org//storage/The%20Attacks%20on%20Palestinian%20Digital%20Rights.pdf [https://perma.cc/7HK8-AX6R] [hereinafter THE ATTACKS ON PALESTINIAN DIGITAL RIGHTS].

disproportionate censorship of Palestinian voices.[216] Civil society groups documented hundreds of content removals, account suspensions, hashtag blocks, and other enforcement actions.[217] In a two-week period, 7amleh, the Arab Center for Social Media Advancement, documented five hundred cases of content moderation impacting Palestinians, with Meta accounting for 85% of those restrictions.[218] Often, there was little-to-no notice provided to people.[219] This response was not unprecedented. Civil society groups have long documented this second-tier of digital personhood imposed on Palestinians.[220] Through coordination and advocacy, these groups managed to leverage media attention sustained enough to prompt platform responses. Platforms replied with a familiar refrain: that some of the instances were because of technical glitches.[221]

Meta's DIO Policy played a central role in how the platform moderated the incident. In one instance, employees discovered the DIO list had an entry for Al-Aqsa, which was meant to cover the Al-Aqsa Martyrs' Brigade, an entity on the American Foreign Terrorist Organizations list.[222] But separate from the organization, Al-Aqsa is a common expression and was also the site of the mosque where several clashes occurred.[223] Even when the system operated as intended, the appearance of groups such as al-Qassam Brigades (the military wing of the Palestinian group Hamas) stunted the way the clashes could be discussed.[224] The DIO's ban on praising designated groups threatened to capture all but the most explicit denouncements. This included a post from the media outlet Al Jazeera reporting on the conflict.[225] On May 10, 2021 a person shared a news article reporting on a threat by al-Qassam Brigades to fire rockets if Israeli forces did not withdraw from the Al-Aqsa mosque and Sheikh Jarrah.[226] Meta removed and then later republished the person's post after the Oversight

[216] *Id.* at 2-3 (noting this occurred while extremist Israeli groups' content was not removed).

[217] *Id.* at 3.

[218] *Id.*

[219] *Id.* at 5 (stating this occurred with almost half of content removals on Instagram).

[220] *See* THE ARAB CTR. FOR SOC. MEDIA ADVANCEMENT, FACEBOOK AND PALESTINIANS: BIASED OR NEUTRAL CONTENT MODERATION POLICIES? 6, 14-15 (2018), https://7amleh.org/wp-content/uploads/2018/10/booklet-final2-1.pdf [https://perma.cc/2G3N-HQXU] (noting Palestinians face "unprecedented" censorship by social media platforms while anti-Palestinian sentiment is overlooked).

[221] THE ATTACKS ON PALESTINIAN DIGITAL RIGHTS, *supra* note 215, at 3.

[222] Ryan Mac, *Instagram Censored Posts About One of Islam's Holiest Mosques, Drawing Employee Ire*, BUZZFEED NEWS (May 12, 2021, 5:10 PM), https://www.buzzfeednews.com/article/ryanmac/instagram-facebook-censored-al-aqsa-mosque [https://perma.cc/A2DW-YMYR].

[223] *Id.*

[224] *See Case Decision 2021-009-FB-UA*, OVERSIGHT BD. (Sept. 14, 2021), https://www.oversightboard.com/decision/FB-P93JPX02 [https://perma.cc/KYG5-UP26].

[225] *See id.*

[226] *Id.*

Board accepted the case for review.[227] The Meta Oversight Board affirmed the decision to reverse the original removal, critiqued the vague policy, and called for an independent assessment of potential bias in Meta's moderation practices in Arabic and Hebrew.[228] A year later, an external auditor found that Meta's policy and enforcement "had an adverse human rights impact . . . on the rights of Palestinians . . . freedom of expression, freedom of assembly, political participation, and non-discrimination, and therefore on the ability of Palestinians to share information and insights about their experiences as they occurred."[229]

Detailing potential reasons for the overenforcement against Palestinian voices, the auditor noted (1) error-prone algorithms; (2) Meta's interpretation of its legal obligations regarding U.S.-designated terrorist organizations; and (3) content moderators that were unable to adequately understand the Palestinian dialect of Arabic.[230] Additionally, Meta's automated tools included a "hostile speech classifier" that attempted to detect content with a "high likelihood of violating Meta's policies."[231] This classifier was deployed to assess Arabic content, but not content in Hebrew.[232] Not only was the tool unilaterally deployed, it was also less accurate for Palestinian Arabic because "the dialect is less common, and because the training data—which is based on the assessments of human reviewers—likely reproduces the errors of human reviewers due to lack of linguistic and cultural competence."[233] The auditor also found that erroneous content removals had subsequent issues that were often left uncorrected. These included "'false' strikes that impacted visibility and engagement" after erroneous removals.[234]

Assessing whether the disparate impact on Palestinian voices reflected bias on Meta's part, the auditor found no intentional bias but various instances of "unintentional bias."[235] But Meta chose to draft a broad policy and enforce it using error-prone methods that would have a disparate impact on Palestinian voices. Moreover, even after documentation of these effects was confirmed, the

[227] *Id.* (agreeing with Facebook's decision to reinstate post regarding threat of violence from al-Qassam Brigades).

[228] *Id.* (recommending Facebook add criteria and examples to DIO policy and engage independent party to determine if Facebook's content moderation has been applied without bias).

[229] BUS. FOR SOC. RESP., HUMAN RIGHTS DUE DILIGENCE OF META'S IMPACTS IN ISRAEL AND PALESTINE IN MAY 2021, at 4 (Sept. 2022), https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf [https://perma.cc/R39V-H9BD].

[230] *Id.* at 5 (noting these as "possible root causes for over-enforcement that Meta should investigate further").

[231] *See id.* at 5 (describing Meta's use of classifiers to identify hostile speech).

[232] *Id.* at 5.

[233] *Id.* at 8.

[234] *Id.* at 5.

[235] *Id.* at 7-8. The auditor defined intentional bias as "where some people are deliberately treated differently than others," and unintentional bias as "where policies and processes may be neutral on their face, or in place for reasons of legal compliance, but impact some people differently than others." *Id.* at 7.

policies and practices largely stayed the same.[236] In short, the DIO policy and its enforcement became a central component for policing how the world can communicate, critique, and document the impact of Israeli occupation on the lives of Palenstinians.[237] As noted by impacted stakeholders, Meta has become "another powerful entity repressing their voice that they are helpless to change."[238]

On the other hand, when Russian forces commenced their invasion of Ukraine on February 24, 2022, Meta promptly made adjustments to ensure its DIO and violence policies aligned with Ukrainians defending their homeland.[239] In March 2022, Meta announced a series of changes, including allowing individuals "in some countries to call for violence against Russians and Russian soldiers in the context of the Ukraine invasion," call for death to Russian President Vladimir Putin and Belarusian President Alexander Lukashenko, and even call for the explicit removal of Russians from Ukraine and Belarus.[240] As part of these revisions, the company even overturned restrictions placed on the Azov Battalion, a Ukrainian Neo-Nazi organization that wears an array of Nazi symbols, including the Totenkopf and Sonnenrad.[241]

The differing approaches to Palestinian and Ukrainian violence not only reflect a double standard, but they also reflect platform alignment with a stratified vision of race that upholds global white supremacy. It instantiates a vision in which only some communities are entitled to see violence as a component of autonomy, only some communities have broad protections for freedom of expression, and only some communities are accorded global support. The DIO policy is meant to restrict the freedoms of people that Western

[236] *See id.*

[237] *See generally id.* (describing effect of Meta content moderation on Palestinian freedom of expression).

[238] *Id.* at 6.

[239] Ryan Mac, Mike Isaac & Sheera Frenkel, *How War in Ukraine Roiled Facebook and Instagram*, N.Y. TIMES (Mar. 31, 2022), https://www.nytimes.com/2022/03/30/technology/ukraine-russia-facebook-instagram.html (explaining Meta's suspension of "some of the quality controls that ensure that posts from people in Russia, Ukraine and other Eastern European countries meet its rules").

[240] Munsif Vengattil & Elizabeth Culliford, *Facebook Allows War Posts Urging Violence Against Russian Invaders*, REUTERS (Mar. 11, 2022, 12:04 AM), https://www.reuters.com/world/europe/exclusive-facebook-instagram-temporarily-allow-calls-violence-against-russians-2022-03-10/ [https://perma.cc/3YEU-L3MD] (describing changes in Meta hate speech policy in light of Russian invasion of Ukraine).

[241] Sam Biddle, *Facebook Allows Praise of Neo-Nazi Ukrainian Battalion If It Fights Russian Invasion*, INTERCEPT (Feb. 24, 2022, 12:44 PM), https://theintercept.com/2022/02/24/ukraine-facebook-azov-battalion-russia/; Christopher Miller, *Ukraine's Far-Right Forces See an Opportunity in Russia's Invasion Threat To Grow Their Violent Movement*, BUZZFEED NEWS (Jan. 31, 2022, 5:59 PM), https://www.buzzfeednews.com/article/christopherm51/ukraine-russia-invasion-far-right-training [https://perma.cc/9EHL-YKN5] (describing members of Azov Battalion's military uniforms as being adorned with Nazi symbols).

governments view as a threat. As a result, even company blacklists are set aside when they threaten privileged groups or conflict with Western foreign policy. It also perpetuates a broad national security framework that otherizes communities of color, treating them as second-class citizens no matter where they find themselves in the global diaspora.

2. Tiered Enforcement: Protecting Whiteness

After numerous incidents involving white supremacists and conspiracy theorists, platforms unveiled new approaches that simultaneously excused delayed action and limited the scope of their actual enforcement. In each instance, they elected to treat the threats of white supremacy through ad hoc policies and measured approaches traditionally withheld from groups with less political power.

For example, after a white supremacist livestreamed his attacks on mosques in Christchurch, New Zealand, Meta announced it was enacting additional measures to combat white supremacy.[242] However, this rollout struck a different tone than was applied for combating groups like ISIS and al-Qaeda.[243] The company said it was not attempting to ban "American pride" or limit people's ability to "demonstrate pride in their ethnic heritage."[244] Instead, the company said it was banning the "praise, support and representation of white nationalism and white separatism."[245] While the company's initial announcement said that it was banning more than two hundred organizations under the policy, later posts revealed the number ended up somewhere closer to twelve.[246]

As part of its announcement, Meta noted that its hate speech policies had "long" prohibited white supremacy.[247] But the announcement revealed yet again how narrow its vision of white supremacy is in practice. Even at the time, the company's own civil rights auditor urged the platform to move away from its

[242] *Standing Against Hate*, META NEWSROOM (Mar. 27, 2019), https://about.fb.com/news/2019/03/standing-against-hate/ [https://perma.cc/SP68-CJNT] (announcing ban on praise of white nationalism on Facebook and Instagram).

[243] *See* Nick Clegg, Vice President of Glob. Affs. & Commc'n, *Facebook Does Not Benefit from Hate*, META NEWSROOM (July 1, 2020), https://about.fb.com/news/2020/07/facebook-does-not-benefit-from-hate/ [https://perma.cc/KCN2-QG34] ("99% of the ISIS and Al Qaeda content we remove is taken down before anyone reports it to us.").

[244] *Standing Against Hate*, *supra* note 242.

[245] *Id.*

[246] *Compare Combating Hate and Extremism*, META NEWSROOM (Sept. 17, 2019), https://about.fb.com/news/2019/09/combating-hate-and-extremism [https://perma.cc/FG9Q-AA8P] ("We've banned more than 200 white supremacist organizations from our platform."), *with* Adam Mosseri, *An Update on Our Equity Work*, META NEWSROOM (Sept. 9, 2020), https://about.fb.com/news/2020/09/an-update-on-our-equity-work [https://perma.cc/JD2W-3P7E] ("This includes removing 23 different banned organizations, over half of which supported white supremacy.").

[247] *Standing Against Hate*, *supra* note 242.

limited approach requiring the use of specific words.[248] The company's blog post said it was banning "praise, support, and representation," but in reality the policy only bans "explicit" posts, which Meta defined to mean posts containing the terms "white nationalism" and "white separatism"—leaving untouched any post espousing the same ideology but avoiding using those explicit terms.[249] This insight, confirmed by a civil rights audit,[250] provides another example of the gaps that exist between blog posts, public policies, and internal enforcement.

In 2020, responding to further acts of white supremacist violence connected to groups like the Kenosha Guard and QAnon, Meta announced a sudden removal of militias and QAnon groups through an updated DIO policy deploying a new tier system.[251] Most relevant, a new Tier 3 created a special enforcement category for "Militarized Social Movements, Violence-Inducing Conspiracy Networks, and individuals and groups banned for promoting hatred."[252] This Tier largely functions as a catch-all for content associated with white supremacy. Bans against praise and support are not applied to this group—only ones against their presence or coordination.[253] This suggests an even lower standard than "representation."[254]

Despite numerous acts of violence, including an insurrection and several mass shootings, platforms made a choice to deploy selective efforts.[255] There are many potential reasons for this. First, there has been political blowback for wide-scale enforcement against content that has connections to American Republican elected officials.[256] From QAnon, to the Proud Boys, to the January 6th

[248] LAURA MURPHY, FACEBOOK'S CIVIL RIGHTS AUDIT – PROGRESS REPORT 9 (June 30, 2019), https://about.fb.com/wp-content/uploads/2019/06/civilrightaudit_final.pdf [https://perma.cc/RB99-2VSE] ("The Audit Team recommends that Facebook expand the white nationalism policy to prohibit content which expressly praises, supports, or represents white nationalist ideology even if it does not explicitly use the terms 'white nationalism' or 'white separatism.'").

[249] *Id.* (describing narrowness of Meta's policy).

[250] LAURA MURPHY, FACEBOOK'S CIVIL RIGHTS AUDIT FINAL REPORT 8 (July 8, 2020), https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf [https://perma.cc/D2NM-UAWQ] (advocating for Facebook policy going beyond prohibiting explicit references to white nationalism).

[251] *An Update to How We Address Movements and Organizations Tied to Violence*, META NEWSROOM (Oct. 17, 2022, 8:00 AM), https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/ [https://perma.cc/S5EH-DRS2]; *Meta DOI Policy*, *supra* note 168.

[252] *Meta DOI Policy*, *supra* note 168.

[253] *An Update to How We Address Movements and Organizations Tied to Violence*, *supra* note 251.

[254] *Meta DOI Policy*, *supra* note 168 (banning "representation" of Tier 1 and Tier 2 entities on Facebook).

[255] *Compare Standing Against Hate*, *supra* note 242 (announcing ban on praise of white nationalism), *with* Clegg, *supra* note 243 (discussing removal of ISIS and al-Qaeda content).

[256] *See* Mac & Silverman, *supra* note 63 (discussing Mark Zuckerberg's lenient response to posts by right-wing figures to "avoid political backlash").

insurrection, individuals and groups tied to this offline violence also have connections to figures like Donald Trump, Josh Hawley, and Marjorie Taylor-Greene (among others).[257] Drafting policies that would have political consequences and further build on unproved claims of bias against conservatives could be too consequential.[258] Second, policymakers and leadership may simply fail to understand the nuance or fail to consider supporters of these groups to be truly dangerous.[259] The introduction of tiers and qualifiers appear to function as a protection against subjecting traditionally powerful communities and their political connections to overenforcement.

3. Secret Blacklists Conceal Racialized Threat Assessments

TVE policy is operationalized through the specific people, organizations, and events that platforms add to undisclosed blacklists. As a result, understanding the disparate impact of these policies requires an understanding of the covered entities.

In statements, the platforms have all indicated that they rely on national and international terrorism designations, such as the U.S. Treasury Department's list of foreign terrorist organizations or the United Nations Security Council's consolidated sanctions list, but that these are supplemented with their own designations.[260] Reliance on these lists is a key factor in predicting disparate

[257] *See, e.g.*, Philip Bump, *Ted Cruz's Electoral Vote Speech Will Live in Infamy*, WASH. POST (Jan. 6, 2021, 3:46 PM), https://www.washingtonpost.com/politics/2021/01/06/ted-cruzs-electoral-vote-speech-will-live-infamy (discussing ways Ted Cruz's Senate speech contributed to January 6th insurrection); Danny Hakim & Elaina Plott, *Josh Hawley, Vilified for Exhorting Jan. 6 Protesters, Is Not Backing Down*, N.Y. TIMES (Mar. 8, 2021), https://www.nytimes.com/2021/03/08/us/politics/josh-hawley-vilified-for-exhorting-jan-6-protesters-is-not-backing-down.html (recounting Senator Hawley's statement he is "not backing down" after encouraging January 6th insurrectionaries).

[258] *See* Mac & Silverman, *supra* note 63.

[259] *See* Sam Biddle, *Revealed: Facebook's Secret Blacklist of "Dangerous Individuals and Organizations*,*"* INTERCEPT (Oct. 12, 2021, 1:16 PM), https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/ [https://perma.cc/3VLB-MLM7] (noting how Facebook's tiered DIO policy indicates they view Muslim organizations as most dangerous, while predominantly white groups are generally in more lenient Tier 3).

[260] *See* Monika Bickert & Erin Saltman, *An Update on Our Efforts To Combat Terrorism Online*, META NEWSROOM (Dec. 20, 2019), https://about.fb.com/news/2019/12/counterterrorism-efforts-update/ [https://perma.cc/LU94-MXVP] (describing Facebook, Microsoft, Twitter, and YouTube's collaboration with "experts in government, civil society and academia" to combat terrorism on their platforms); Biddle, *supra* note 259 (explaining how "Facebook takes most of the names in [its] terrorism category directly from" U.S. government's Specially Designated Global terrorists sanctions list); *European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech*, EURO. COMM'N (May 31, 2016), https://ec.europa.eu/commission/presscorner/detail/en/IP_16_1937 [https://perma.cc/BP7A-AE5E]; *Sanctions List Search*, OFF. OF FOREIGN ASSETS CONTROL, https://sanctionssearch.ofac.treas.gov [https://perma.cc/CFC8-VMYC] (last visited Nov. 9, 2023) (compiling U.S. Treasury Department's sanctions

impact on certain communities. Many U.S. and international sanctions lists target al-Qaeda, the Taliban, and ISIS, making it likely that over-removals will disproportionately affect Muslim communities.[261] It was only recently that platforms began providing definitions for the types of organizations covered, but those descriptions provide limited insight. For example, Twitter now provides a general description of what a "violent extremist group" and "violent organization" is, mostly describing an entity that "deliberately target[s] humans or essential infrastructure with physical violence and/or violent rhetoric as a means to further [its] cause."[262]

Despite paltry descriptions, no platform publishes a list of the entities subjected to their TVE policy. The principal method of determining which entities are subjected is usually through trial and error—attempting to trip the automated enforcement algorithms that deploy broad enforcement. However, in 2021, *The Intercept* published a leaked copy of Meta's Dangerous Individuals and Organizations list.[263] The document confirmed what many advocates have long suspected, from the heavy-handed approach to voices from the Middle East and North Africa to the more surgical approach to white supremacist groups.[264] The white supremacists banned were largely dead, like Adolf Hitler and Joseph Goebbels, along with contemporary figures that rose to an untenable public outcry such as Alex Jones, Richard Spencer, David Duke of the KKK, and Gavin McInnes from the Proud Boys.[265] The list of groups banned as Hate Banned Entities were largely European rock bands.[266] Finally, the list of Militarized Social Movements was the most expansive of the company's removal efforts pertaining to white supremacy, containing several hundred groups varying from antigovernment to pro-Trump groups.[267]

By comparison, the individuals and groups listed in the crime and terror categories were almost exclusively people of color.[268] In addition to efforts to target Middle Eastern groups like ISIS and Hamas, the list also targeted drug cartels and gangs operating in Latin America.[269] This appears to set up the infrastructure for expanding national security concerns, such as the U.S.

list) [hereinafter *Sanctions List Search*]; *United Nations Security Council Consolidated List*, U.N. SEC. COUNCIL (Aug. 30, 2023), [https://perma.cc/8LN2-WW8B]; *Twitter TVE Policy*, *supra* note 189.

[261] *See Sanctions List Search*, *supra* note 260; *United Nations Security Council Consolidated List*, *supra* note 260.

[262] *Twitter TVE Policy*, *supra* note 189.

[263] Biddle, *supra* note 259.

[264] *See id.* (describing "differences in demographic composition between Tiers 1 and 3," with harsher penalties for Muslim and other predominantly nonwhite organizations).

[265] *Id.*

[266] *Id.*

[267] *Id.*

[268] *See id.*

[269] *Id.*

government's interest in combatting "transnational organized crime."[270] Finally, within the United States, Meta's DIO list of designated criminal organizations appeared unilaterally focused on several Black and Latino gangs with Chicago origins,[271] raising questions not only about disparate impact, but whether government involvement is behind this targeted enforcement.

TVE lists, through reliance on government designations and closed-door pressure, operate as an enforcement mechanism for Western threat assessments. Building on decades-long occupations, broad swaths of online communities are forced to endure a stifled online experience. Unwittingly or not, TVE policy functions as an extension of Western colonialism or occupation, adapted to cover an essential method of communication. This system carries penalties like account suspensions, content removals, hashtag blocks, and search filters.[272] By imposing denouncement and neutrality as the only acceptable sources of discourse, it treats entire communities as incapable of nuanced discussion of world events. This is one manifestation of how TVE policy carries a separate harm: treating communities of color as second-class citizens of the digital sphere. This stratified operation treats communities as inherently suspicious, necessary to police aggressively, and dangerous. By comparison, even in the wake of numerous attacks and ongoing harassment campaigns, the attendant controls for white supremacists are narrow and temporary. The importance of unencumbered discourse is only permitted for communities that platforms do not view as an ongoing threat.

IV. TOWARD RACE-CONSCIOUS REALISM IN CONTENT POLICY

This Part proposes steps for incorporating race-conscious content moderation at the policy level. While there is unlikely to be consensus as to what constitutes "good" content moderation practice, these interventions reject a value-neutral position and seek to advance pragmatic solutions for the victims of racial subjugation. Whether looking at individual decisions or the structures that inform the broader system, the study of content moderation requires a clear articulation of what is right and what is just.

First, this Part outlines methods for platforms to "see race" in its multifaceted operation. These include studying online racial formation and deploying context-specific policies based on local dynamics and social hierarchies. Second, this Part proposes steps for more effective antiracist content moderation. These include accounting for indirect incitement to hatred and violence, lowering the demonetization threshold for racism, and generating a public list of high-reach public figures. During particularly volatile moments, such as elections or violent uprisings, platforms should develop transparent and limited preclearance measures for some high-reach individuals to limit the reach

[270] *See, e.g.*, Exec. Order No. 14060, 86 Fed. Reg. 71,593 (Dec. 15, 2021) (establishing council on organized crime).

[271] Biddle, *supra* note 259.

[272] *See, e.g.*, *The Attacks on Palestinian Digital Rights*, *supra* note 215, at 3.

of incitement. Finally, this Part proposes interventions to protect dissent by marginalized communities. These include publishing public lists of individuals and organizations covered under terrorism and violent extremism policies, ending prohibitions against "praise" and "glorification," and establishing greater transparency regarding content removal requests from platforms and government actors.

A. *Seeing Race*

There is understandable trepidation to consider race, as it is a tool that can easily be co-opted for harmful ends. These could range from authoritarian to capitalistic ends. The reality is that both dangers already exist and are actively causing harm. For example, race and racial proxies are a lasting feature in online advertising. Meta settled multiple lawsuits and plugged avenues for racial targeting only to find new proxies deployed.[273] Similarly, government involvement in social media surveillance,[274] data sharing,[275] and content moderation[276] are already essential features of platform governance. Data-driven policing is increasingly built on relationships, networks, and profile information obtained from social media, even if many of the assumptions drawn are bigoted and inaccurate.[277] A race-conscious approach acknowledges race's persistent role in American life while seeking to put it toward restorative ends. The ability to differentiate between insidious and protective forms of racial classification is not only a possible task, it is an essential one. Documenting racial harms enables more effective government redress and reinvigorates civil rights protections for the digital age.

[273] Julia Angwin, Ariana Tobin & Madeleine Varner, *Facebook (Still) Letting Housing Advertisers Exclude Users by Race*, PROPUBLICA (Nov. 21, 2017, 1:23 PM), https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin [https://perma.cc/B3GA-HBUH] (explaining how advertisements violated Fair Housing Act).

[274] *See, e.g.*, Hannah Bloch-Wehba, *Content Moderation as Surveillance*, 35 BERKELEY TECH. L.J. 1297, 1299 (2021); *LAPD Social Media Monitoring Documents*, BRENNAN CTR. FOR JUST. (Dec. 15, 2021), https://www.brennancenter.org/our-work/research-reports/lapd-social-media-monitoring-documents [https://perma.cc/46RE-UL7K] [hereinafter *LAPD Social Media Monitoring Documents*] (describing Center's request for documents about LAPD's use of social media monitoring).

[275] *See, e.g.*, Michael Edison Hayden, *"Antifa Civil War" Fake News Story Treated as "Threat" by DHS Officials, Emails Reveal*, NEWSWEEK (Apr. 14, 2018, 10:13 AM), https://www.newsweek.com/antifa-civil-war-fake-news-story-treated-threat-department-homeland-security-885251 [https://perma.cc/MKC6-B6QW].

[276] *See, e.g.*, Rabea Eghbariah & Amre Metwally, *Informal Governance: Internet Referral Units and the Rise of State Interpretation of Terms of Service*, 23 YALE J.L. & TECH. 542, 545 (2021).

[277] *See, e.g.*, Forrest Stuart, *Code of the Tweet: Urban Gang Violence in the Social Media Age*, 67 SOC. PROBLEMS 191, 192 (2019) (discussing how overstating effects of social media violence can reinforce myth of Black criminality).

Turning this into practice is, of course, challenging. There is a need to thoroughly examine platform dynamics and adjust based on how they work. Instagram is not YouTube, and Twitter is not TikTok. Understanding race within a platform using anonymous accounts is different from a video-first usage where a corporeal body exists. In each instance, individuals may be using a fictitious persona. Even policies like Facebook's real-name policy[278] can be gamed, a practice commonly undertaken by police officers engaged in undercover surveillance operations.

Experimentation based on specific platform structures could incorporate everything from voluntary self-reporting to analysis based on context clues derived from analyses of account behavior. For example, some Black communities are well-versed in discerning individuals engaged in digital blackface, employing strategic hashtags like "#YourSlipIsShowing" to combat a coordinated harassment campaign that used fake profiles and attempted to use African American Vernacular English to pass off as Black people.[279] Learning from the lived experience of impacted communities[280] is an essential starting point for race-conscious content policy. Another approach might build on survey methods Meta is already deploying through a partnership with YouGov to assess outward perceptions of fairness and equity.[281]

B. *Fighting Racism*

As a preliminary matter, social media companies should adopt a broader scope of actionable racism[282] beyond the moment when it is explicitly expressed

[278] *Names Allowed on Facebook*, FACEBOOK HELP CTR., https://www.facebook.com/help/229715077154790/ [https://perma.cc/5WRM-VJH3] (last visited Nov. 9, 2023) (outlining rules for names on Facebook).

[279] *See* Rachelle Hampton, *The Black Feminists Who Saw the Alt-Right Threat Coming*, SLATE (Apr. 23, 2019, 5:45 AM), https://slate.com/technology/2019/04/black-feminists-alt-right-twitter-gamergate.html [https://perma.cc/8LYV-SBEU] (describing instances where hashtags were utilized to call out fake accounts).

[280] *See* Mari Matsuda, *Looking to the Bottom: Critical Legal Studies and Reparations*, 22 HARV. C.R.-C.L L. REV. 323, 324 (1987).

[281] Aisha Malik, *Instagram Will Survey US Users About Race To Assess If It Is 'Fair and Equitable'*, TECHCRUNCH (July 28, 2022, 2:01 PM), https://techcrunch.com/2022/07/28/instagram-race-survey/ [https://perma.cc/KDM3-9Y3C].

[282] MATSUDA ET AL., *supra* note 48, at 23 (discussing various forms of racism). In this Article, I adopt the definition of racism articulated by Mari Matsuda:

> Racism, as used here, comprises the ideology of racial supremacy and the mechanisms for keeping selected victim groups in subordinated positions. The implements of racism include:
>
> 1. Violence and genocide;
> 2. Racial hate messages, disparagement, and threats;
> 3. Overt disparate treatment; and
> 4. Covert disparate treatment and sanitized racist comments.

Mari Matsuda, *Public Response to Racist Speech: Considering the Victim's Story*, 87 MICH. L. REV. 2320, 2332 (1989) [hereinafter *Public Response to Racist Speech*].

or when animus crosses over into violence. Mari Matsuda reminds us that violence "is the final solution, as fascists know, barely held at bay while the tactical weapons of segregation, disparagement, and hate propaganda do their work."[283]

So long as hate speech and harassment policies remain unmoored from the power dynamics that create differentiation in risk and harm, the costs will remain borne by the communities "least able to pay."[284] Moving forward, platforms should acknowledge and document the unique ways in which minority communities are most susceptible to harassment and hate speech and the ways in which such content can result in both offline and online harms. Too often, platforms delay action until real world violence, such as mass shootings, make the enforcement of these policies ineffective, forcing them to use blunt enforcement tools that inevitably swallow large swaths of ordinary speech. A race-conscious approach to hate speech and harassment should allow for more sophisticated and gradual enforcement mechanisms, as well as tools for people to protect themselves.

To properly address the harms of racism, platform policies must abandon colorblind policies that treat all hate speech and harassment equally. Instead, these policies should incorporate "the connection of racism to power and subordination."[285] To achieve this, platforms should experiment with different approaches, such as building out their existing descriptions of protected categories or deploying a tiered system that imposes gradual enforcement penalties. Regardless of the approach, platforms must account for historical subjugation and the horizontal relationship imposed between racial groups.

Race-conscious hate speech and harassment policies should also account for lessons about the tactics of white supremacy online. These iterations will vary depending on the platform and its affordances, but the focus should be on how individuals and entities modernize messages of inferiority, justify oppression, and call for persecution. Similarly, content policies should account for indirect coordination of racial harassment—for example, accounting for indirect calls for followers to harass members of marginalized communities. This must also account for historical oppression, otherwise efforts to hold politicians accountable could get swallowed up in expanded enforcement. Finally, platforms should undertake the process of publishing a public list of prominent individuals and subjecting relevant subsets of accounts to preclearance moderation during volatile moments, such as protests and elections. Incorporating these lessons will not always translate into expanded content removals, but it could provide alternative points for demonetization, interstitials, or other intermediate controls to deter racist harassment. Communication regarding content policy enforcement should deploy clear and detailed warnings, specific references to policies, and opportunities to adjust.

[283] *Public Response to Racist Speech*, *supra* note 282, at 2335.

[284] MATSUDA ET AL., *supra* note 48, at 48.

[285] *Id.* at 36.

C. *Protecting Dissent*

Race-conscious TVE policy requires a clear-eyed assessment of how this policy largely advances racialized threat assessments. These assessments manifest in country-specific sanctions lists, but platform lists also contain prejudiced assumptions. As a preliminary matter, platforms should publish a public list of the individuals and organizations covered by their TVE policies. To the extent that they simply rely on sanctions lists from the United States or United Nations, the designation should be public. These disclosures will help assess whether policy rules, such as those addressing white supremacy, are written in a manner that does not miss the organizations that are driving violence, and whether these policies remain predominantly focused on ISIS and al-Qaeda.

Second, policies that broadly target content based on "praise" or "glorification" should not be used, regardless of the type of violent extremism being targeted.[286] These imprecise terms will inevitably capture expressions of general sympathy for, or understanding of, certain viewpoints, not to mention news reporting. Relying on vague labels makes it more likely that content will be misinterpreted or inaccurately flagged by automated tools. These terms also introduce opportunities for policy misuse, as praise or glorification provide catchall categorizations that become easy and opaque ways to justify removal.[287] Instead, enforcement should be limited to entity specific accounts, subjecting the moderation of posts that express praise, support, or sympathy to regular enforcement under hate speech policies. Meta's decision to allow praise of certain conspiracy networks like QAnon and hateful individuals like Alex Jones reflects a decision to avoid overburdening speech from people with more powerful political support; the same calculations should be extended to people from marginalized communities.[288]

Finally, TVE policies should also incorporate historical subjugation and contemporary power dynamics into their analysis for specific designated "violent events."[289] Marginalized communities need social media for organizing, political education, and documenting and exposing human rights abuses. Additionally, relationships between government actors and social media companies require greater public transparency. Instances where the government asks for content removal based on community standards instead of local law should be publicly disclosed as part of transparency efforts. These disclosures should note the government agency and the specific rule violation. In instances of hate speech and harassment, platforms should track the targeted person or group; in instances of violent extremism policy, platforms should track the designated person or individual that triggered the policy.

[286] *See, e.g.*, *Standing Against Hate*, *supra* note 241 (explaining policies against "praise" of white nationalism).

[287] *See id.*

[288] *See* Mac & Silverman, *supra* note 63 (outlining exceptions to Meta's rules on combating misinformation and hate speech).

[289] *Meta DOI Policy*, *supra* note 168.

For a moment during social media's infancy, political movements could seize on the delay repressive governments had in understanding the democratizing threats of networked communication.[290] That moment is gone. Instead, repressive governments around the world surveil,[291] hack,[292] discredit,[293] and disappear[294]; the toolbox of repression is adapted to the digital present. If platforms wish to take credit for their role in facilitating protest movements ranging from the Arab Spring to Black Lives Matter, their TVE policies must be narrowed to avoid being an essential part in authoritarian crackdowns.

CONCLUSION

Social media is more than a mirror for offline bigotry; it is an active developer of the ways racial stratification is conceived, protected, and advanced. The status quo approach to drafting and interpreting content policy protects the cultural, political, and economic advantages attendant to whiteness. In other words, the standard approach for understanding and redressing racism leaves communities of color trapped in another person's imagination.[295] Whether it is our past, present, or future, racial subjugation is understood as natural and inevitable. Challenging this discriminatory system requires not only mapping the specific ways that content policy advances white supremacy, but also proposing an alternative vision that seeks to provide protections for the victims of racial subjugation. To be sure, this task is not without peril. At its core, content moderation is a censorship regime, one that largely operates outside of democratic transparency or accountability. But the dangers of misuse must not prevent us from attending to the rise in racial hatred and authoritarianism that floods our online and offline communities.[296] Race-conscious content policy

[290] *See generally* ZEYNEP TUFEKCI, TWITTER AND TEAR GAS: THE POWER AND FRAGILITY OF NETWORKED PROTEST (2017).

[291] *LAPD Social Media Monitoring Documents*, *supra* note 274 (noting social media has been used by governments to surveil events like protests).

[292] *See, e.g.*, Stephan Shankland, *Pegasus Spyware and Citizen Surveillance: Here's What You Should Know*, CNET (July 19, 2022, 8:49 AM), https://www.cnet.com/tech/mobile/pegasus-spyware-and-citizen-surveillance-what-you-need-to-know/ [https://perma.cc/6VZY-XJUR] (discussing how software like Pegasus spied on protesters).

[293] *See, e.g.*, Hayden, *supra* note 275.

[294] *Iranian Activist Disappears After Criticizing Internet Bill*, ASSOCIATED PRESS (Feb. 26, 2022, 5:57 AM), https://apnews.com/article/technology-iran-media-social-media-dubai-685dd71ae6299f9411c703cbbec0cc7a [https://perma.cc/5M8M-32GW].

[295] Thanks to Ruha Benjamin for first exposing me to this idea, and for encouraging me to engage in an imagination battle.

[296] Addressing First Amendment absolutism that prevents redress for the victims of racism, Mari Matsuda writes:

> There is, in every constitutional doctrine we devise, the danger of misuse. For fear of falling, we are warned against taking a first step. Frozen at the first amendment bulkhead we watch the rising tide of racial hatred wash over our schools and workplaces. Students victimized by racist speech turn to university administrators for redress, and are told that

faces the world as it is so that we can redirect it toward what it must become: a place of dignity and equal opportunity.

> the first amendment forecloses institutional action. We owe those students a more thoughtful analysis than absolutism.

MATSUDA ET AL., *supra* note 48, at 50.