Allard K. Lowenstein International Human Rights Clinic • Yale Law School

MANAGING AND MITIGATING INDIRECT HATE SPEECH ON META SOCIAL MEDIA PLATFORMS

This Report was researched and authored by students, faculty and staff of the Allard K. Lowenstein International Human Rights Clinic at Yale Law School. This Report represents the views and perspectives of the authors. It does not represent an institutional position of Yale Law School.

ISBN #979-8-218-27299-9
Cover design by Shannon Sommers

TABLE OF CONTENTS

INTRODUCTION

This white paper proposes a human rights approach to content moderation of a particular kind of hate speech in the context of conflict or crisis situations: "proxy" or indirect hate speech, defined in this paper as hate speech that is likely to contribute to violence but does not explicitly name a protected characteristic. The Lowenstein International Human Rights Clinic at Yale Law School ("Lowenstein") presented this research to the Facebook Oversight Board (the "Oversight Board") at the Board's invitation.[1] It should be noted that the views and opinions expressed in this paper are those of its authors and do not necessarily reflect the views or positions of the Oversight Board or any other entity.

The term "proxy hate speech" is currently largely, if not completely, absent from the literature. However, the concepts of "indirect hate speech" or "code words," which address similar content, are widely recognized as constituting an important and complex challenge across technological, legal, and policy dimensions. Notably, these types of hate speech are inherently more difficult to identify than more direct forms of hate speech, and determining whether something is or is not hate speech can have immediate and drastic consequences in the context of areas of emerging or present conflict. As the Board recognized in its *Alleged Crimes in Raya Kobo* opinion, for instance, "true reports on atrocities can save lives in conflict zones, while unsubstantiated claims regarding civilian perpetrators are likely to heighten risks of near-term violence."[2]

At the outset, it is important to note that civil society and experts consulted by Lowenstein working on moderation of hate speech on social media noted again and again that the speech on which this white paper focuses represents only a part of a much larger problem in Meta's moderation of speech that violates human rights norms and its own community standards. As is widely recognized, Meta has done an extremely poor job removing hate speech from its social media platforms.[3] Whether this is due to technological challenges inherent in content moderation, an underinvestment in the necessary resources, or a combination of both, the broader problem is not merely that difficult-to-discern hate speech is slipping through the virtual cracks; rather, there is a widespread failure to remove even the most direct and obvious examples of hate speech—including speech that is clearly defined as such by Meta's own community standards. A focus solely on "proxy" hate speech without appropriately emphasizing the ways in which Meta has repeatedly fallen short on direct hate speech would be problematic and misleading.

Last spring and summer, Global Witness, an international non-governmental organization,[4] tested Meta's ability to detect hate speech in three countries characterized by volatility or outright violence: Myanmar, where hate speech has incited violence and genocide against the Rohingya;[5] Ethiopia, which has for over two years been embroiled in a brutal civil war[6] characterized by widespread war crimes;[7] and Kenya, weeks before a hotly contested election in an unstable political landscape.[8] In each context, the result was the same: an evident failure to detect clear instances of hate speech.[9]

For each country, Global Witness identified some of the worst examples of real-life hate speech that had been posted on Facebook and submitted them for approval as advertisements, removing them after they had been evaluated but before they could be published.[10] In Myanmar, eight "highly offensive and disturbing" posts in Burmese, "filled with dehumani[z]ing language and direct calls for killings," were each approved.[11] Similarly, in Ethiopia, twelve examples of Amharic-language hate speech comparing people to animals and calling for individuals to be killed, starved, or "cleansed"—several of them amounting to calls for genocide and none of them difficult to interpret—were nevertheless accepted for publication, notwithstanding the fact that the majority of these examples had *previously been removed from Facebook* for violating its community standards on hate speech.[12] The English-language posts submitted as advertisements in Kenya, which "compar[ed] specific tribal groups to animals and call[ed] for rape, slaughter[,] and

beheading," were initially rejected only for failing to comply with the Grammar and Profanity policy, but after minor corrections, they, along with other posts written in Swahili, were accepted.[13]

These findings are particularly concerning for three reasons. First, Meta claims that it "hold[s] advertisers to even stricter [] policies" than ordinary posts with regards to its Community Standards, including the one pertaining to hate speech.[14] It is therefore reasonable to assume that if Meta is failing to detect hate speech in advertisements, it is even less likely to be able to do so in organic posts.[15] Second, Meta has special reason to prioritize removing potential hate speech in these particular countries given the high risks that hate speech will result in substantial harm there.[16] International human rights law places particular emphasis on hate speech that has the potential to incite violence,[17] and relatedly, Meta has acknowledged and emphasized the particular steps it has taken in Myanmar, Ethiopia, and Kenya intended to reduce hate speech.[18] If Meta is failing to identify obvious hate speech in high-risk contexts where it is expending significant resources, its practices are likely similar, if not worse, in contexts where fewer resources are employed. Finally, the hate speech in these investigations should have been "relatively easy tests" for Meta as they involved the main languages spoken in countries where many languages are spoken.[19] Given the results of these investigations, there should be significant concern that Meta is even worse at identifying hate speech written in less commonly spoken languages.

While the Global Witness investigations concededly involve a small amount of hate speech, internal Meta documents suggest that the hate speech problem is much more widespread. Private internal communications obtained by former Meta-employee-turned-whistleblower Frances Haugen revealed that Meta removes only three to five percent of the hate speech on its platform.[20]

Thus, this white paper will propose a lens under which content moderation of hate speech at Meta should be viewed if content moderation practices are to meaningfully improve in a manner that protects and respects human rights. In other words, this white paper aims to both retain a focus on indirect hate speech in high-risk contexts, while recognizing the need to effectively address the harmful phenomenon of online hate speech as a whole. It proceeds in three parts.

As background, Part I discusses and defines hate speech generally as well as what the paper will refer to from this point forward as "indirect hate speech," the term that is more commonly used in the literature (as compared to "proxy hate speech") and understood by practitioners in this space. As this part explains, a broader definition of hate speech is a necessary starting point to properly capture the indirect hate speech that is the topic of this white paper. Closing the gaps in Meta's existing definition of hate speech will not only cover a broader swathe of online hate speech as conceptualized by human rights law, but it may improve existing enforcement of hate speech that is already covered by the current policies by clarifying and elucidating their scope. This part further assesses indirect hate speech in the context of hate speech at large. It relies on desk research of secondary literature as well as interviews of human rights advocates who specialize in the area of online hate speech prevention. Interviewees largely work at organizations focusing on the area more generally, but some are also based in organizations focused on reducing online hate speech in particular country contexts, and they were selected to ensure a diverse range of views. The views expressed by the interviewees were considered in conjunction with other research, but afforded significant weight due to their expertise on these issues. For awareness, some interviewees requested to remain anonymous; others consented to having their identities revealed.

Part II outlines and presents a "signals" framework for determining whether something is indirect hate speech and, just as importantly, which content to prioritize in the context of large-scale enforcement. This proposal includes online signals, which relate to the content of the post itself and the ways in which users interact with it, and offline signals, which relate to the real-world social and political context in which the post exists. Online signals include: (i) proxy language; (ii) account history; (iii) reach and engagement; and

(iv) explicit disclaimers. Offline signals include: (i) local risk of conflict; (ii) identity of the target; and (iii) identity of the poster. This paper uses country case studies to illuminate why these proposed signals are useful. The countries are almost all characterized by emerging or active conflict; the remainder were chosen on the basis of having significant research available on hate speech. Notably, the list of signals is not exhaustive, and whether any given signal should be dispositive or whether a more holistic consideration should be embraced is not contemplated by this paper. Instead, the precise mechanics of how this signals framework would interact with Meta's enforcement process should be determined by Meta's own technical content moderation experts.

Part III briefly concludes with an overview of the signals framework and a brief treatment of operations and infrastructure in relation to a potential institutionalization of such framework.

1 CONCEPTUALIZING HATE SPEECH

1.1 META'S DEFINITION OF HATE SPEECH

Meta defines hate speech narrowly in the Facebook Community Standards, which "outline what is and is not allowed on Facebook."[21] Because "Meta wants people to be able to talk openly about the issues that matter to them, whether through written comments, photos, music, or other artistic mediums, even if some may disagree or find them objectionable," it states that it limits expression only when it serves one or more of the following values: authenticity, safety, privacy, and dignity.[22] Hate speech is one of the categories of speech that Meta identifies as objectionable, and it is defined as follows:

> We define hate speech as a direct attack against people—rather than concepts or institutions—on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.[23]

The policy further defines attacks as "violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing[,] and calls for exclusion or segregation."[24] The definition also requires that the attack be "on the basis of" the protected characteristic. No additional information is provided in the public-facing community standards, but how Meta determines whether an attack is "on the basis" of a protected characteristic may significantly impact the reach and effectiveness of this policy. This definition allows considerable room for discretion by content moderators. Meta identifies particular words that are likely to constitute "attacks" within its market-specific "slur lists."[25] For example, Meta has a list that covers the "Southern Cone" (Chile, Argentina, Uruguay, and Paraguay)[26] and another that covers sub-Saharan Africa.[27] Additionally, the company's internal guidelines for content moderators identify certain categories of individuals that may be targets of hate speech despite not falling squarely into one of its protected characteristics. In its *Russian Poem* decision, the Oversight Board noted that attacks on certain professions may constitute hate speech when referenced alongside a protected characteristic (*e.g.*, "Russian soldier").[28] In short, Meta enforces its hate speech definition in part by pre-identifying certain categories of victims and attacks that may cause speech to rise to the level of hate.

Even so, as illustrated below, the current definition is overly narrow when compared with international human right standards. Not all hate speech takes the form of *direct* attacks that explicitly name a protected characteristic and make explicit that the attack is "on the basis" of the protected characteristic. Concepts and institutions often serve as stand-ins for protected characteristics. For example, calls for action against the Tigrayan People's Liberation Front ("TPLF"), a political organization, can contribute to violence against ethnic Tigrayans generally—and indeed be intended to do so—without ever issuing a direct attack

against Tigrayans.[29] Physical characteristics, geographic descriptors, and historical references—among many other concepts—also frequently serve as proxies for protected characteristics.

In order to capture more implicit or indirect forms of hate speech, Meta must first adopt a more capacious definition, as well as a clear method for assessing whether a given post falls within or without that definition. The remaining sections in this part provide a framework as to how to build a more nuanced definition of hate speech. Section B relies on desk research on international human rights law, focusing on both human rights conventions that implicate hate speech as well as guidelines for social media companies specifically. Section C relies on a series of interviews with experts from civil society organizations focused on countering online hate speech, both generally and in particular country contexts. Section D, relying on the analysis from Sections B and C, proposes a broader definition of hate speech for Meta to use and explains how it improves upon the current one.

1.2 INTERNATIONAL HUMAN RIGHTS STANDARDS

1.2.1 Human Rights Conventions

As a threshold matter, it is worth noting that international human rights standards are largely directed towards states. Nevertheless, it is useful to examine these standards to illuminate what Meta should define as hate speech while acknowledging that Meta differs from state actors in various ways.

There is no single definition of hate speech under international human rights law. However, several treaties identify hate speech[30] as an unlawful category of expression and a clear exception to freedom of speech.[31] Article 20 of the International Covenant on Civil and Political Rights ("ICCPR") requires parties to ban "propaganda for war" and "[a]ny advocacy of national, racial[,] or religious hatred that constitutes incitement to discrimination, hostility, or violence."[32] The American Convention on Human Rights more broadly prohibits "incitement[] to lawless violence."[33] The International Convention on the Elimination of all Forms of Racial Discrimination ("ICERD") is even more restrictive, forbidding "all dissemination of ideas based on racial superiority or hatred," in addition to speech that incites discrimination or violence.[34] In addition to these limitations, states may place domestic restrictions on hate speech in ways that are necessary for respecting the rights and reputation of others or for the protection of national security or public order.[35]

Notably, none of these notions of hate speech limit themselves to direct references to vulnerable groups. The ICCPR prohibits *any* advocacy of national, racial, or religious hatred that constitutes incitement. The ICERD similarly forbids *all* dissemination of ideas based on racial superiority or hatred, regardless of whether the racial group is explicitly named. Human rights conceptions of hate speech, and whether a direct reference to a vulnerable group is required, are therefore more capacious than the Facebook Community Standards definition.

On at least one occasion, the European Court of Human Rights acknowledged that speech that only indirectly refers to a protected characteristic can amount to hate speech. In *Sanchez v. France*, the Court held that the European Convention did not give the applicant a right to denigrate Muslim immigrants on Facebook. The comments at issue did not name any ethnicity or religion directly but mentioned concepts closely associated with Muslims. Specifically, the commenter wrote that a certain politician had: "transformed Nîmes into Algiers, there is not a street without a kebab shop and mosque; drug dealers and prostitutes reign supreme, no surprise he chose Brussels, capital of the new world order of sharia . . . ."[36] The Court determined that these implicit attacks "clearly encouraged incitement to hatred and violence against a person because of their belonging to a religion."[37] These statements would not have amounted to hate speech under Meta's current definition.

International human rights law is divided on the question of whether certain speech can be inherently wrongful or whether it must contribute to tangible harms to constitute hate speech. The ICERD prohibits dissemination of ideas based on racial hatred or superiority, regardless of their effects. In interpreting the ICCPR, the Human Rights Committee has found that statements may be so discriminatory as to meet the standard for incitement without an inquiry into their effects. For example, in *Faurisson v. France*, the impugned statements "were of a nature as to raise or strengthen antisemitic feelings."[38]

At other times, human rights authorities have determined that speech must cause harm to contravene international law. According to the United Nations Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, speech only constitutes "incitement" if there is a "real and imminent danger of violence resulting from the expression."[39] In practice, international bodies have accepted evidence of an increase in prejudicial views among the audience of a particular statement as proof of incitement. In *Ross v. Canada*, the Human Rights Committee determined that a teacher's antisemitic comments were unlawful under Article 20(2), because they caused students to hold Jewish people in contempt.[40] Assuming that implicit forms of hate speech—*e.g.*, hate speech by proxy—do not meet the *Faurisson* standard for inherent wrongfulness, one would need to examine the statement's impacts on attitudes and actions toward a protected group to determine its lawfulness.

The Rabat Plan of Action argues for the severity of hatred as the key consideration for determining whether speech constitutes "incitement" under Article 20(2) of the ICCPR. The United Nations General Assembly proposed a six-factor test for assessing severity of hatred, which includes: (1) the social and political context; (2) the speaker's status within society and among her audience; (3) the speaker's intent; (4) the content and form of the speech; (5) the size of the audience and means of dissemination; and (6) the action advocated and the likelihood of said action being taken.[41]

Of course, international law does not formally bind the conduct of social media companies. Nevertheless, Meta should carefully consider these standards in crafting its own content moderation policies. The ICCPR and other relevant treaties apply to states, which in turn must ban hate speech under domestic law, which may include, where needed, regulating the private sector.[42] However, many governments are unable to effectively regulate online hate speech, while others have used such regulation to shrink civic spaces and target activists and human rights defenders.[43] As such, international human rights bodies have encouraged social media companies to police hate speech on their platforms in ways that align with international human rights law. The United Nations Guiding Principles on Business and Human Rights establish that private companies must "respect human rights" and provide "appropriate and effective remedies."[44] More specifically, the United Nations Secretary General committed in 2019 "to establish and strengthen partnerships with new and traditional media to address hate speech."[45] In short, Meta should define hate speech with international law in mind.

1.2.2 Guidelines for Social Media Companies[46]

Several international human rights bodies have developed suggestions for how social media companies should define and respond to online hate speech. Meta can look to these documents—in addition to human rights treaties and jurisprudence—as it evaluates its community standards. The Draft Effective Guidelines on Hate Speech—published by the United Nations Special Rapporteur on Minority Issues—urges social media companies to offer protection to minorities in the area of incitement to hatred and hate speech at least to the extent required under international human rights standards aimed at states.[47] The Special Rapporteur goes on to provide a concrete working definition of "online hate speech":

> Social media companies are responsible for effectively prohibiting and removing content in the shortest time possible that is discriminatory, hostile[,] or violent towards those (community members) with protected characteristics on the basis of any identity factor and

> especially those belonging to national or ethnic, religious[,] and linguistic minorities on the basis of their minority identity.[48]

The definition acknowledges the power imbalances involved in how hate speech is regularly deployed on social media. Irrespective of the definition that a social media company chooses, the Guidelines encourage social media companies to clearly define "broad, subjective, or ambiguous terms such as 'direct' and 'attack.'"[49] Although the Facebook Community Standards define "attack," the meaning of "direct" remains unclear and perhaps is overly limiting.

Finally, the Special Rapporteur suggests that content moderation decisions should follow a "necessity and proportionality" analysis.[50] In other words, any moderation decision taken should be the "least restrictive means" available for protecting the rights of the targeted person or group (*i.e.*, necessary) and should be considered alongside other enforcement options on a spectrum of severity (*i.e.*, proportional).[51] Although conducting a thorough necessity and proportionality analysis for every Facebook or Instagram post is not operationally feasible, Meta should aim to design its policies and enforcement processes to only limit speech when one or more person's rights are at stake and should do so in the least restrictive manner possible.

Regional human rights bodies have also published guidance on the application of human rights principles to social media platforms.[52] Further discussion, however, is beyond the scope of this paper.

1.3 FEEDBACK FROM CIVIL SOCIETY AND HATE SPEECH ORGANIZATIONS

Interviews with civil society have revealed two central concerns: (1) the need for Meta to develop a more contextualized and nuanced approach to understanding hate speech; and (2) the need to promote better enforcement by Meta of its existing policies on hate speech. As to the first point, civil society actors have expressed concerns about orienting a content-moderation policy for indirect hate speech around predefined proxy categories. Their feedback points to the value of centering the analysis on contextualized factors and avoiding overly broad generalizations. This aligns with existing recommendations from experts/researchers on the need to 'institutionalize impermanence' when moderating hate speech. As to the second point, civil society actors have directed attention to deficiencies in Meta's current approach to enforcement. Their feedback expresses concern about a lack of willingness and engagement on Meta's part to take down hateful content. These are two interrelated, but distinct issues, and this white paper addresses them both.

1.3.1 Contextualized and Nuanced Approaches

Interviews with civil society have been informative in understanding how to structure a more contextualized and nuanced framework for countering online hate speech. Specifically, the Online Hate Prevention Institute[53] ("Institute") and the University of California, Berkeley's D-Lab[54] ("D-Lab") have developed different contextual approaches to addressing online hate speech. For the Institute, a framework oriented around narratives, and for the D-Lab, a framework that conceives of hate speech as a spectrum, inform their understandings of hate speech. Their approaches reveal the importance of incorporating context in evaluations of online hate speech.

Context and narrative largely shape the Institute's framework. Andre Oboler, the Chief Executive Officer of the Institute, believes that a context-driven approach is necessary in tackling hate speech, emphasizing the need for local political knowledge to understand when a given piece of content constitutes indirect hate speech. With an eye towards context, the Institute has developed its work around "hate narratives," which Oboler believes provide a more helpful framework for understanding hate speech. For example, the Institute has created a narrative schema for understanding antisemitism, as well as one for understanding anti-Asian hate, relying on context-specific factors to develop these narratives.

The narrative schema for addressing antisemitism identifies major categories—and within those categories, subcategories—for the types of content of which a content moderator should take notice. The major categories include the following: (1) Holocaust-related content; (2) incitement to violence; (3) classic antisemitism; and (4) antisemitism related to Israel. Each major category encompasses multiple subcategories that delve into further detail on the types of information that may reflect antisemitic content. According to Oboler, the categories are derived from the text of the International Holocaust Remembrance Alliance's working definitions of antisemitism and Holocaust denial and distortion.

Furthermore, in October 2022, the Institute published a report on anti-Asian racism in Australian social media.[55] After collecting a sample of 182 instances of anti-Asian hate on Facebook and Instagram in Australia, the Institute consolidated the data to create a schema with four high-level categories of anti-Asian hate, each of which contains multiple subcategories. The following are the four high-level categories: (1) incitement to violence against Asians; (2) demonizing/dehumanizing Asians; (3) attacking Asians because of their culture; and (4) other xenophobia against Asians/Asian heritage. Each category points to the various ways anti-Asian hate may manifest itself on online platforms. According to Oboler, the categories are derived directly from the data that the Institute gathered, given that there is no widely accepted definition. Pulled from the report, Table 1 on the next page outlines the narrative schema for anti-Asian hate.

Table 1: Online Hate Prevention Institution Schema

1. Incitement to violence against Asians	
	1.1 Inciting violence against a specific Asian person / business / organisation
	1.2 Inciting violence against Asian businesses in general
	1.3 Inciting violence against Asian people generally
2. Demonising / dehumanising Asians	
	2.1 Yellow Peril / Yellow Terror / Yellow Specter
	2.2 Presenting Asians as spreaders of sickness
	2.3 Presenting Asians as animals or non-human
	2.4 Presenting Asians as violent / criminals
	2.5 Nazi analogies
3. Attacking Asians because of their culture (e.g. food)	
	3.1 Food related attack on culture
	3.2 Negative views of the value of Asian culture
	3.3 Negative views about Asian people blaming their culture
4. Other xenophobia against Asians / Asian heritage (i.e. separating them from mainstream society)	
	4.1 Statements of exclusion e.g. that they don't belong here
	4.2 Statements telling people to go back where they came from
	4.3 Statements telling people to give up their culture
5. Other forms of anti-Asian racism	

For both the antisemitism and anti-Asian hate schemas, the schematic categories are specific to the narrative, as opposed to being broad categories applicable across different narratives. These schemas highlight how the Institute uses context to shape the narrative. Indeed, for this organization, context is the key to understanding hate speech in its various forms.

The D-Lab's framework, which conceives of hate speech as a spectrum, provides another avenue for understanding hate speech. During the interview, the representative from the D-Lab expressed opposition to viewing hate speech as a binary—that is, where yes/no questions are asked as to whether a given piece of content constitutes hate speech. Believing a yes/no binary to be too simplistic, the representative argues that a spectrum is a more effective construct for understanding hate speech. To the representative, a spectrum allows for more flexibility and better captures the continuity and fluidity of hate speech.

The D-Lab has effectuated the spectrum by developing a methodology that uses ten questions to evaluate potential hate speech. The methodology decomposes the complex question—"Is this hate speech?"—to a simpler set of ten questions. The model identifies ten ordinal outcome variables: (1) sentiment; (2) (dis)respect; (3) insult; (4) humiliation; (5) inferior status; (6) violence; (7) dehumanization; (8) genocide; (9) attack/defense; and (10) hate speech benchmark.[56] For a given piece of content, an annotator would, for instance, provide information to determine if the given content meets the various elements of hate speech. The annotator would consider the ten variables, one at a time, and ask if the expressed sentiment of the content is positive or negative, if the content is dehumanizing, if the content is genocidal, etc. By decomposing hate speech into ten components, the model provides a means of unpacking the larger, abstract question—does this content constitute hate speech?—into tangible questions that point to the various elements of hate speech. Ultimately, the deep learning model can predict ratings as to each of the ten components of hate speech. The result is essentially a technological model that factors in all the responses to the ten questions, such that an answer is extrapolated as to a specific piece of content. The spectrum conceives of placing the most extreme, harmful version of hate speech on one end—for instance, genocidal hate speech—and counter speech of a positive nature on the other end. The low-to-medium-to-high variations of speech along the spectrum enable a more granular construction of hate speech.

The unique frameworks from the Institute and the D-Lab for understanding hate speech point to various considerations. The frameworks uniquely recognize the underlying complexities of hate speech. Their attention to the details that vary across contexts highlight how fixing a content-moderation policy around pre-established categories may be ill advised—at least, to the extent those categories leave little room for flexibility in understanding diverse contexts. The feedback from civil society introduces thought-provoking questions. How expansive should the scope of indirect hate speech be? What are the advantages and disadvantages to constructing the scope broadly? Would a broader conceptualization enable the recognition of overlapping themes? Or would it risk blurring boundaries between diverse contexts and overriding the interest in prioritizing contextualization?

Despite their diverse approaches, Oboler from the Institute and the representative from the D-Lab are similarly mindful of the need to avoid analyzing indirect hate speech in a simplistic or inflexible manner. Their narrative and spectrum-oriented frameworks, respectively, are cognizant of the complicated scenarios that different contexts may pose, suggesting that a one-size-fits-all approach is likely not feasible. Notably, during the interview, Oboler was skeptical of creating "proxy" categories for indirect hate speech—that is, proxy categories for political party/organization, proper name, ideology, and physical description. The narrative approach to which the Institute ascribes is catered to the specific details of a given factual scenario, so the pre-establishment of proxy categories would fail to capture the nuance central to the Institute's framework. The D-Lab's approach also suggests a need to adopt a more contextual or narrative-driven approach to hate speech. Though D-Lab's framework relies on a pre-established set of ten questions, it centers those questions on the specific facts or words of a case. The questions are oriented around the nature

of the harm that the content produces—for example, by inquiring whether the content is dehumanizing or genocidal.

Applying the two frameworks to the study of indirect hate speech reveals the complicated nature of online hate speech. Moving forward, if Meta adopts the term "indirect" hate speech to encompass everything that is not "direct" hate speech, doing so may come at the expense of contextualization. If everything is grouped together under an umbrella term, this may produce a situation where lines are blurred, such that content moderators ignore context-oriented factors and make sweeping generalizations. On the other hand, various forms of hate speech, including coded and ambiguous language, may be understood to fall along various parts of the same spectrum. And if these diverse forms of language manifest themselves along the same spectrum, there may be value to incorporating them in a collective understanding of indirect hate speech.

An interview with Professor Robert Post from Yale Law School was revealing.[57] When asked about additional categories to include in a potential list of proxy categories, Professor Post expressed that "millions exist" (*e.g.*, profession, food preferences, personality), far too many to reduce to a list, concluding that it would be counterproductive to center the analysis on a limited set of predefined categories. Rather than focus on the use of a word or on a preconceived identification category, Professor Post suggested that it would be more constructive to focus on the effect of the content. Observing that Facebook defines hate speech mechanically, Professor Post explained the need to develop a more dynamic system that accounts for the effects of the speech. Rather than being tied to a discrete set of categories, this system would incorporate new words and new meanings thereof as they arise in culture. This is a mammoth task, certainly, but it represents a more thorough approach than relying on categories that are incomplete and quickly obsolete.

As these interviews and the research above indicate, there is value to expanding beyond a category-based methodology to indirect hate speech. This white paper proceeds from that assumption and seeks to adopt a flexible, responsive, and creative approach to the way Meta may both conceptualize and address indirect hate speech. Interviews conducted thus far point to common themes and lessons: significant complexity underlies the issue of indirect hate speech, and civil society actors have already developed thoughtful frameworks for understanding and locating hate speech—frameworks that may prove instructive as the Board considers the most effective approach to addressing indirect hate speech.

1.3.2 Enforcement Issues

There is a need to situate this white paper within the broader conversation of addressing online hate speech effectively. Interviews with civil society have unveiled enforcement issues with respect to Meta's existing hate speech policies. Civil society actors have also cautioned against modifying Meta's policies in such a way that distracts from the overall purpose of countering hate speech at large. This paper is mindful of the sheer magnitude of Meta's reach, as well as the extent to which the focus on indirect hate speech may distract from the overall goal of removing harmful hate speech from Meta's platform. In other words, these entities have suggested that some of the perceived gaps in content moderation may be better addressed by urging Meta to devote more resources to current enforcement systems rather than finding policy solutions.

In the eyes of civil society, Meta has demonstrated a lack of willingness to enforce its hate speech policies. In an interview with representatives from the United Nations Office of the High Commissioner for Human Rights in Kenya (the "Office"),[58] a human rights analyst explained how the Office has brought cases to Meta that the Office perceived to clearly constitute hate speech and incite violence. However, the response that the Office received from Meta was that the cases did not violate Meta's policies. Likewise, in an interview with the National Cohesion and Integration Commission in Kenya (the "Commission"),[59] a representative explained how the Commission has attempted to speak with companies like Meta to take down hate speech on online platforms. However, according to the representative, companies like Meta have not been willing to

take down hateful speech. Similarly, in an interview with Amnesty International in Kenya,[60] a representative conveyed that social media companies have done very little to stop or even reduce hate speech in Kenya. The representative highlighted social media companies' lack of investment in content moderation in Kenya. A representative from Digital Threats at Global Witness echoed these sentiments.[61] Referencing the reports mentioned in the introduction of this paper, the representative asserted that there is a massive amount of content on social media that is clearly hate speech and that Meta has not generally succeeded at addressing even this low-hanging fruit.

Related to frustrations with Meta's enforcement of its current policies are concerns about Meta's future policies further impeding the mission of countering hate speech. In an interview with the founder and executive director of a prominent research and advocacy organization who asked to remain anonymous, the representative encouraged us to consider how focusing on the "subtleties" of online hate speech, including hate speech in its more indirect forms, may prove a distraction—namely, a distraction from taking down explicitly harmful hate speech that continues to permeate the Internet.[62] According to the representative, the basics have yet to be fully covered. Once content moderators attend to the "subtleties" of hate speech, they may overlook the reality that harmful hate speech in clear and direct violation of Meta's policies remains alarmingly frequent on the Internet.

Whether this is a matter of technical constraint or willingness remains unclear. Regardless, civil society actors have identified concerning issues with respect to Meta's enforcement. As the Board moves forward in conceptualizing indirect hate speech, it is imperative to recognize how indirect hate speech and direct hate speech interact and to not promote a system that addresses the former while neglecting the latter.

1.4 DEFINING HATE SPEECH

Given the relevant human rights framework and stakeholder insights above, Meta's definition of hate speech should be adjusted as follows to be properly inclusive of all hate speech, including indirect hate speech:

> Any harmful attack—direct or otherwise, intentional or otherwise—against people on the basis of protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.

Harmful should be defined as "carrying a substantial risk of violence, persecution, or discrimination." *Attack* should be defined, as in the current policy, as "violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing[,] and calls for exclusion or segregation."

By removing the requirement that attacks be "direct," this definition aligns with scholars', activists', and human rights authorities' understanding that hate speech often takes implicit forms. Hate speech by proxy, which is excluded under the "direct attack" definition, would be included here. In the same vein, removing the language that omits attacks against "concepts or institutions" reflects the fact that such language can—in practice—function as attacks against particular groups of people.

By adding the requirement that attacks be "harmful," the definition ensures that Meta only interferes with individuals' speech when doing so is necessary to protect the rights of others. The harmfulness criterion also brings Meta's notion of hate speech in line with human rights bodies that consider the effects of a statement when determining whether or not it constitutes hate speech. Finally, the expected harms of a statement depend largely on the context surrounding that statement (*e.g.*, the identity of the speaker, the political and social environment, and the size of the audience). Therefore, this definition coheres with the civil

society groups and the Rabat Plan of Action, which highlight the indispensability of context in assessing possible hate speech.

The proposed definition adds both nuance and precision to Meta's understanding of hate speech. Nevertheless, setting exact boundaries between what is and is not hate speech is flatly impossible. The harmfulness of a post and the extent to which it constitutes an attack are matters of degree. Just because hate speech exists on a spectrum, however, does not mean that it cannot be effectively moderated. Part II provides an enforcement framework that can help operationalize the definition above.

2 SIGNALS FOR DETERMINING HATE SPEECH

As discussed in Part I, hate speech is a wide-ranging concept. This white paper does not attempt to present an exhaustive typology of indirect hate speech terms or categories. Instead, it will outline a variety of signals that the Oversight Board and Meta should consider when deciding whether a particular post constitutes hate speech, including secondary hate speech, as well as which posts should be prioritized for human review. These signals are not part of the definition of hate speech proposed in Section I.D. Instead, they represent an enforcement framework for more effectively identifying posts that fall within that definition, taking into account the need for large-scale enforcement.

The analysis in this part will draw on examples of indirect hate speech from high-risk contexts, in which responding to such hate speech is particularly urgent. Specifically, it examines online hate speech in Brazil, Cameroon, Ethiopia, Kenya, Myanmar, and Ukraine.

There are two types of signals that can indicate that a post constitutes hate speech:

- **Online signals**, which relate to the content of the post itself and the ways in which users interact with it; and
- **Offline signals**, which relate to the real-world social and political context in which the post exists.

The signals within each type are listed in Table 2. However, this list is not necessarily exhaustive, and Meta or the Oversight Board may identify additional useful signals.

A. Online Signals	B. Offline Signals
a. Proxy Language b. Account History c. Reach and Engagement d. Explicit Disclaimers	a. Local Risk of Conflict b. Identity of the Target c. Identity of the Poster

Table 2: Hate Speech Signals

These signals can be used to triage potentially harmful speech to human moderators. Interpreters familiar with local language, politics, and customs are generally best placed to determine whether particular content amounts to indirect hate speech, though it is important to ensure that they are as unbiased as possible and not affiliated with individuals or groups engaged in the very hate speech that they are supposed to be moderating. Interviewees familiar with high-risk countries consistently reported that local readers are typically able to quickly identify the harmfulness of a post that an algorithm or foreign interpreter might consider benign.[63] Of course, manual review of every potential instance of hate speech is impossible. However, assessment of ambiguous cases as identified by the signals described here would help confirm whether a given post is likely to incite hostility, discrimination, or violence within the relevant context. Manual review may be especially valuable in contexts where Meta's hate speech detection algorithms are underdeveloped or in contexts in which hate speech language is evolving rapidly.

This paper does not assert whether any given signal should be dispositive in determining whether or not a post is more likely to constitute hate speech. Considering all of these factors holistically would allow for a nuanced understanding of hate speech. However, in practice Meta and/or the Board may determine that certain signals should automatically trigger human review or specific content moderation actions. Establishing such specific thresholds could make this signals framework more scalable and increase the predictability of Meta's content moderation decisions. Moreover, Meta may prioritize certain signals within this framework—*e.g.*, reach and engagement; local risk of conflict; and identity of the target—for guiding its content moderation processes. Meta would then choose how to moderate the post in question (*e.g.*, remove, reduce, or inform)[64] based on the likelihood that it amounts to indirect hate speech and the severity of its content. Ultimately, the details of Meta's enforcement process are best determined by the company's technical content moderation experts. This part simply provides a general enforcement framework that Meta and the Board would decide how to operationalize.

For each signal listed above, this part discusses:

- The relevance of the signal in determining whether or not a given post constitutes indirect hate speech; and
- Examples of hateful speech that illustrate how the signal operates in context

2.1 ONLINE FACTORS

2.1.1 Proxy Language

As discussed in Part I, hate speech often makes use of proxy language instead of directly referring to a protected characteristic.[65] Interviewees familiar with Ethiopia,[66] Cameroon,[67] and Myanmar[68] all described the presence of such proxy language online.

These are terms that at least some audiences will interpret as advocating hatred, discrimination, or violence based on a protected characteristic. Proxy language can vary in severity and explicitness. The following categorization captures this variation:

- **Slurs**, which are derogatory terms exclusively associated with a particular ethnic, religious, racial, or other group
- **Pejoratives**, which are terms associated with hatred toward vulnerable groups, but which also have other non-discriminatory and non-violent meanings
- **Code words**, which are words that do not ordinarily connote hatred, but that online communities co-opt to secretly refer to protected groups

Posts that contain slurs have a relatively high likelihood of constituting hate speech, given that those words are primarily used to denigrate or dehumanize vulnerable groups. By contrast, pejoratives and code words are more context-dependent. Determining whether uses of such language amount to hate speech may require a more detailed consideration of the other factors discussed in this section. As a representative from the Institute described, "You need local political knowledge and context to understand when something is indirect hate speech . . . [and] you have to take into consideration the entirety of the context."[69]

Measurement of the presence and frequency of coded language in individual posts requires maintaining an evolving and updated list of potentially hateful terms in each of the countries and languages in which Meta has users.

Some words are commonly deployed in hateful ways across many contexts. For example, instigators of ethnic hatred have referred to their targets as "insects" in Rwanda,[70] Ethiopia,[71] Bangladesh,[72] Korea,[73] and Canada.[74] That being said, even these words need to be translated into each language in which they are used to be identified on Facebook. Additionally, despite the existence of some common terms, coded language varies significantly between countries, languages, and cultures. Therefore, Meta should create and frequently update country-level lists of common proxy terms,[75] perhaps by leveraging its relationships with civil society organizations.[76] This research could begin with countries that are likely to experience violent conflict in the near future, where hate speech might be especially dangerous.

Examples

Below are examples of proxy language in various highly charged political contexts. Please note that not all of these examples necessarily amount to hate speech. We simply provide them to illustrate the use of these terms.

2.1.1.1 Terms or Phrases as Proxies

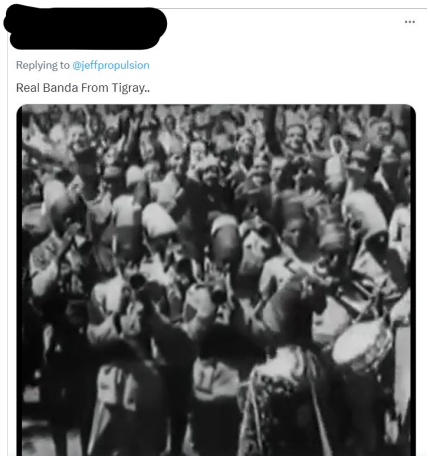
In Ethiopia, a common slur used to describe Tigrayan people is *tsila.*[77] The word imitates one of the main sounds in the Tigrinya language: the "tse."[78] It has come to be associated with stereotypes attached to Tigrayan people, such as "angering easily" or "becoming rich without hard work."[79] The word therefore carries a notion of threat in addition to being belittling.

Image 1: Use of Tsila *on Facebook*[80]

Translation: "You who says Tigray, Tigray, Tigray. Fuck you! Bunch of dirty people. All of you go to hell, you are all ***tsila*** *thieves."*

A common pejorative word is *banda*, which means traitor.[81] More specifically, it refers to people who betray their country out of self-interest. The term was used historically to refer to Ethiopians who fought alongside Italian occupiers as mercenaries between 1935 and 1941. Since the beginning of the conflict between the Ethiopian federal government and the TPLF, the term has often been used to denigrate and otherize Tigrayan people. However, it is worth noting that the word is also used to refer to members of other ethnic groups and is also often deployed between members of the same ethnic group, illustrating that proxy terms often have multiple meanings, not all of which are hateful.

Image 2: Use of Banda *on Twitter*[82]



The tweet contains an historical video of Tigrayans standing alongside Italians.

The Rohingya Muslim minority is often referred to as "illegal immigrants"[83] from Bangladesh by military and political authorities in Myanmar as a means of justifying policy changes to revoke their citizenship and strip them of their rights.[84] Such usage illustrates how immigration status often serves as a thinly-veiled stand-in for ethnicity, which is a protected characteristic.

In Cameroon, proxy language often refers to individuals' place of origin. The word *asuni*—meaning "someone who is not from the tribe"—is often used to denigrate people of other tribal affiliations.[85] The term *graffi* is used to refer negatively to people from grassland areas in the northwestern and western regions of the country.[86] Similarly, the phrase "Come No Go" is used to refer to people from the northwest who have settled in the southwest—originally to work on the plantations of the Cameroon Development Corporation—but have not returned.[87] *Anglofou* and *Francofou* refer to Anglophones and Francophones, between whom there are political tensions as a result of Cameroon's colonial past.[88]

2.1.1.2 Political Organizations as Proxies

In Ethiopia, the TPLF is often used as a stand-in for ethnically Tigrayan people. In the Oversight Board's *Alleged Crimes in Raya Kobo* case, a Facebook user made a post alleging that "the Tigray People's Liberation Front (TPLF) killed and raped women and children, as well as looted the properties of civilians in Raya Kobo and other towns in Ethiopia's Amhara region." Moreover, "[t]he user . . . claimed that ethnic Tigrayan civilians assisted the TPLF in these atrocities." Although the Oversight Board's analysis relied on the Violence and Incitement Community Standard, rather than the Hate Speech Community Standard, the case illustrates the potential for a political organization to represent a wider group of people.

The word *junta*—from the English word describing a military or political group that takes power over a country by force—is now frequently used to describe Tigrayan individuals.[89] This phrase further demonstrates that Tigrayan people not actually affiliated with the TPLF have become associated with the organization.

One case in which Tigrayan identity was conflated with TPLF membership was that of the killing of Meareg Amare Abreha, a professor at Bahir Dar University. Two posts on Facebook shortly before his death accused him of supporting the TPLF.[90] Abreha's son, Abrham Meareg, has filed suit against Meta, alleging human rights abuses and killings fueled by hate speech on its platforms.

Image 3: Use of Junta on Twitter[91]

2.1.1.3 Physical Descriptions as Proxies

In Yemen, the term "Dhabashi" is based on a character of the same name in a television show from the early 1990s who was from the north of Yemen and was portrayed as showing a disregard for the law and Yemeni traditions. It has since become an offensive and discriminatory term used to refer specifically to Yemenis from the northern area of the country.

2.1.1.4 Individuals as Proxies

Within Ethiopian social media networks, Dr. Tedros Adhanom Ghebreyesus—the Director-General of the World Health Organization ("WHO")—is often used as a symbol of the Tigrayan people. He is characterized as a traitor, human rights violator, and TPLF ally. The Ethiopian government has supported such depictions, for example by attempting to launch a WHO investigation of Tedros for allegedly supporting the TPLF.[92] Ethiopia's ambassador to the United Nations in Geneva, Zenebe Kebede Korcho, also attempted to deliver a virtual speech criticizing Tedros.

Image 4: Use of Tedros on Facebook[93]

Translation: "This donkey is the one who brought COVID upon us, he should go back to his relatives as shepherd of donkeys. He is an active member of homosexuals and brought homosexuality to us, he should be eradicated before adulterating others.

2.1.1.5 Illustrative Country Case Studies

There are two particularly useful case studies that are worth examining in this section to further explain various proxy categories: Ukraine and Kenya.

2.1.1.5.1 Ukraine

As the Russian war against Ukraine has continued, online hate speech against Ukrainians has spanned the Internet. According to an expert in the field, in the war against Ukraine, "hate speech exceeded the limits of verbal communication. Hate became visual: memes, cartoons, drawings. In Russia's war against Ukraine, visual hate speech became a tool of Russian hybrid warfare."[94] According to EUvDisinfo, a project developed to respond to Russia's ongoing disinformation campaigns affecting the European Union, "Russia's war against Ukraine demonstrates the deadly effect of hate speech, as it has served to dehumanise the opponent, in this case the legitimate, elected government in Kyiv and the wider Ukrainian population."[95]

Experts have distinguished between two interrelated prongs of anti-Ukrainian hate speech online: hate speech against the Ukrainians as a people and hate speech against Ukraine as a political community.[96] Although the former more directly relates to hate speech against the Ukrainian people as a vulnerable population, the latter still constitutes a form of indirect hate speech against the Ukrainian people that requires additional moderation.

By referring to the Ukrainian people as "Ukami," "Ukies," or "Ukes"—abbreviated and derogatory terms that are short for "Ukrainians"—pro-Russian users with anti-Ukrainian sentiment have expressed their "'non-recognition' of Ukrainian ethnicity."[97] Other users have combined the term "Ukies" with other offensive words, including "bastard," "scumbag," and "bonzo place." Furthermore, online users have often modified the famous Ukrainian slogan "Glory to Ukraine, Glory to Heroes!" into offensive phrases, such as "Glory to salo," where salo is cured pork fat that is common in Ukrainian cuisine.[98]

The term "Country-404" has been adopted as a derogatory term for Ukraine. A reference to the Internet's "404 Page Not Found" error, pro-war Russians have dubbed Ukraine "Country-404"—alluding to Ukraine as an "error," or illegitimate state.[99] The nature of this nickname reflects a "broader insistence among Russian politicians and propagandists that Ukraine has no right to exist."[100] Moreover, the pro-Russian, anti-Ukrainian narrative of Ukraine as a "failed state" has spread by drawing connections between Ukrainian and Nazi authorities—referring to Ukrainians as "Nazis," "fascists," "Nazi hunta," "punishers," "neonats," or "douchbags."[101] Pro-Russian, anti-Ukrainian online users may also refer to Ukrainians or people who support the Ukrainian language and tradition as "Banderites" or "Shukhevych's followers," in reference to Stepan Bandera and Roman Shukhevych, Ukrainian nationalist politicians and military leaders of the twentieth century.[102]

Although the online hate speech since the Russian war against Ukraine has primarily reflected the pro-Russian narrative, anti-Ukrainian rhetoric has also arisen beyond the geographical boundaries of the war in the countries to which Ukrainian refugees have fled. As millions of Ukrainians have crossed the Polish-Ukrainian border, "one seemingly benign, but actually dangerous word has proliferated in the Polish public discourse . . . . This word is '*ukrainizacja*' ('Ukrainization')."[103]

According to the Dangerous Speech Project, as far-right politicians in Poland have developed a narrative of "Ukrainians govern[ing] [in Poland], and the Polish people becom[ing] second-class citizens . . . , the word 'Ukrainization' [has] become[] an umbrella term for the changes that might be inspired by the presence of Ukrainian refugees." For example, during a meeting of the parliamentary group on internal relations in July 2022, far-right Polish politician Grzegorz Braun presented a "Stop the Ukrainization of Poland" pamphlet, "a sixty-four-page long document that describes the threats that Ukrainian immigrants pose to the "ethno-cultural structure" of the country and the actions the Polish government should undertake to "prevent the rapid depolonization of Poland."[104] Reportedly, the popularity of the term "Ukrainization" proliferated on social media after Braun shared the pamphlet: "One day after Braun's parliamentary team meeting, the Ukrainization topic reache[d] 5 million people on social media."[105] Likewise, the prolific hashtag to "Stop the Ukrainization of Poland"—#stopukrainizacjipolski—spread on social media.[106] Moreover, the phrase was displayed during the March of Independence, a march organized by far-right groups on Polish Independence Day, "televised nationwide and attended by tens of thousands."[107]

According to an expert on far-right extremism from Wroclaw, ultrareligious groups in Poland "look to Russia as a bulwark against secular Western values and denounce the 'Ukrainization' of Poland."[108] Relatedly, the *New York Times* reported that a group of fans at a soccer game in the Wroclaw stadium in October 2022 put up a large banner reading: "Stop the Ukrainization of Poland."[109] An image of the banner was circulated on social media, as reflected below.

Image 5: "Stop the Ukrainization of Poland" Banner[110]

Notably, the fear of "Ukrainization" is not limited to Poland. It has spread to other European countries, including Romania, Serbia, and Hungary.[111] According to a report by GlobalFocus, the Ukrainization narrative is present in these countries because of the "significant diaspora."[112] The report explains that the narrative "targets not only assimilationist policies but also post-2014 politics of limiting Russian influence and strengthening the Ukrainian language and culture."[113] The report also specifies that the narrative may refer to "the dissolution of the fundamentals of the state including its territory."[114] It additionally provides examples of how the Ukrainization vocabulary has permeated social media and nationalist and far-right online channels.

Ultimately, as the Russian war against Ukraine continues, it is crucial to recognize the rise of indirect hate speech against the Ukrainian community, not simply from Russian channels but also from areas to which Ukrainian refugees have fled.

2.1.1.5.2 Kenya

The Commission, first mentioned above in Section I.C.2, is a government agency in Kenya that was created to address and reduce ethnic conflict in the country. The Commission has conducted research on ethnically motivated hate speech in Kenya, and a representative from the Commission shared the following work product from the Commission: *Hatelex: A Lexicon of Hate Speech Terms in Kenya.*[115] Published by the Commission in April 2022, the document features results from a research study on hate speech and provides terms that people use to propagate hate speech in Kenya. In an interview, the representative explained that the document was controversial and received backlash from the public, noting that even people in the Commission could not agree upon certain terms.[116]

The main objective of the project was "to develop a lexicon of hate words that can be used as a resource to identify and analyze hate speech in physical spaces and social media texts in a multilingual

perspectives."[117] Specifically, the goal of the study was threefold, seeking to: (1) "[e]xplore the commonly used coded hate speech terms in Kenya"; (2) "[e]stablish how the coded terms are perceived by the users"; and (3) "[a]scertain how these coded terms are perceived by the target communities."[118] The study collected data from 523 respondents to online surveys; the study also involved four focus group discussions, featuring 187 participants. According to the study, "[t]he concern with ethnic stereotypes and coded language in Kenya generally and specifically stems from the fact that it is commonly used in the context of political campaigns to rally support of members of inner communities (and sometimes other communities) against target communities."[119]

The study identified terms and heavily coded messages "that can be used to incite hatred and deliberately exclude other communities" in various languages, including English, Swahili, Sheng, Kikuyu, Meru, and Kalenjin. For each commonly used hate term, the lexicon provides a translation and identifies the target community, the user community, and the meaning of the term. The chart below pulls examples from the Lexicon.

Table 3: Hate Terms in Kenya

Commonly Used Hate Term	**Translation**	**Target Community**	**User Community**	**Meaning**
Fumigation (English)	The action or process of disinfecting or purifying an area with the fumes of certain chemicals	Non-locals	Locals	Mete violence on non-locals so that they can vacate the area
Eliminate (English)	Completely remove or get rid of (something)	Communities perceived to support the dominant political party in the area in question	Communities seen to support opposing political parties	Kill members of the community which is perceived not to support the dominant political party
Kaffir (Swahili)	Derived from Arabic term *Kafir*, which means disbeliever or one who conceals the truth	Non-Muslims	Muslims	A negative referent majorly referring to the non-Muslim communities
Madoadoa (Swahili)	Dots	Non-locals	Political aspirants and supporters of the dominant political party in an area	Used to refer to people who are non-locals living in a place dominantly known to be inhabited by a certain community; considered as a way of inciting one tribe against the other

Wakuja (Swahili)	Those that come	Immigrant communities	Local communities that consider themselves native to an area	Used to refer to communities that migrated and settled into an area
Kama noma, noma. Kama mbaya, mbaya (*Sheng*)	If it is bad, then it is bad	Political aspirants and perceived supporters of the less dominant political party in an area	Political aspirants and supporters of the dominant political party in an area	This statement is perceived by communities that if an opposing political party plays foul, then party supporters should unleash the severest form of counterattack
Kihii (Kikuyu)	Uncircumcised man	Referred to communities which traditionally did not embrace circumcision as their rite of passage, *e.g.*, the Luo, Turkana, etc.	Kikuyu	Demeaning word used against political aspirants from some communities such as the Luo community with the aim to humiliate and terrorize not just the individual men, but their entire communities
Mwiji (Meru)	Uncircumcised man	See above	Meru	See above
Kimurkeldet (Kalenjin)	Brown teeth	Kikuyu	Kalenjin	This is a derogatory term referring to a person with brown teeth with the implication that they cannot even undertake personal hygiene effectively
Otutu labotonik (Kalenjin)	Uproot the weed	Non-Kalenjin communities	Kalenjin	May be used to mean that there are strangers who are a threat within the community and hence should be eliminated

2.1.2 Reach and Engagement

Posts that a large number of users view or with which many individuals interact are more likely to influence real-world behavior.[120] Therefore, if there is resource scarcity, Meta should prioritize such content for hate speech analysis by a human reviewer.

Additionally, the comments on a post can help elucidate whether that post is contributing to hatred or hostility on the basis of protected characteristics.[121] For example, if a post itself does not contain any overtly hateful language but many of the comments on the post do, the audience of the post may understand it as hate speech. In fact, the D-Lab uses the content of a post's comments to assess whether the original post constitutes hate speech.[122]

Of course, comments alone cannot reveal whether original post constituted hate speech. For example, hateful comments may actually target the poster of the content to which they are responding. A human reviewer could more readily determine whether or not the original post itself constitutes hate speech.

Nevertheless, as one interviewee said: "There are lots of posts that maybe technically do not meet [Meta's] definition of hate speech, but if . . . you can see the comments, you can understand that it is a scary situation."[123]

Examples

2.1.2.1 Ethiopia

In October 2021, Dejene Assefa, an Addis Ababa-based political activist and vocal supporter of Abiy Ahmed's government, posted a message to his more than 120,000 Facebook followers. The post implied that Ethiopians should engage in violence against their Tigrayan neighbors: "The war is with those you grew up with, your neighbors…If you can rid your forest of these thorns…victory will be yours."[124] The message was shared over 900 times and garnered more than 2,000 reactions before Facebook removed it.[125] Many of the comments on the original post called explicitly for violence against Tigrayans.[126]

The post illustrates the ways in which other users' engagement with a post can signal its potential for inciting harm as hate speech. Although Assefa's original post did not explicitly mention Tigrayans, many of the comments mentioned Tigrayans and the TPLF outright and thus made clear to whom Assefa was referring. Although the fact that many other accounts shared and reacted to the post does not itself imply that the post constituted hate speech, its wide circulation suggests that if it *were* hate speech, its potential to cause actual harm was significant.

Image 6: Dejene Assefa on Facebook[127]

Meanwhile, Getachew Reda, a senior advisor to the TPLF and the current president of the Tigray Interim Regional Administration, tweeted a statement referring to Abiy's government as "parasitic" and "predatory" and disclaiming liability for violence against "foreign citizens." To Reda's followers, this language might be interpreted as a call to harm non-Tigrayans within Tigray. Given his account's wide reach—Reda has more than 300,000 followers and this particular tweet was shared at least hundreds of times—such inflammatory language carries a high risk of violence.

Image 7: Getachew Reda on Twitter[128]

2.1.3 Account History

In situations where it is unclear whether a particular post amounts to indirect hate speech, the account's previous activity may provide a useful reference point. Prior posts that contain slurs, pejoratives, and/or code words would suggest that the post under examination may also be hate speech.

Similarly, if the account has commented on or shared hate speech posted by other users, its more recent activity may be more likely to constitute hate speech. Researchers have examined hate networks on social media platforms to predict which accounts were more likely to disseminate racist content during the COVID-19 pandemic[129] and Islamophobic content in the aftermath of terrorist attacks in Europe.[130] A representative from the Institute said: "If you find someone who comments on something that is clearly hate and then see what else they are engaging with, you are probably going to find more hate speech."[131]

Of course, considering account history might pose a risk of prejudicial enforcement of Meta's hate speech policy. One way to mitigate against this risk would be to not inform human reviewers that a particular post was flagged for review based on the account's previous activity. And as long as Meta maintains procedures for users to dispute its content moderation decisions, it can have remedies ex post for any incorrect content moderation decisions motivated by an account's prior offenses.

Examples

2.1.3.1 Ethiopia

Image 8 depicts two tweets from the same user referring to the TPLF and Tigrayan people, respectively. The first tweet alleges that all TPLF fighters bear responsibility for rape, murder, and other war crimes. In isolation, this tweet could be read as an attempt either to report on human rights abuses by TPLF combatants or to incite hatred against Tigrayan people. The second tweet, which claims that "Tigray traitors will die," provides evidence suggesting that the other tweet was indeed hate speech. While the content of the second tweet is not dispositive in assessing whether or not the first constitutes hate speech, it is a relevant data point.

Image 8: Two Tweets from the Same User[132]

2.1.4 Disclaimers

A post that features coded language or other indicators of indirect hate speech may not be harmful if it includes an official disclaimer. For example, a user might be sharing or quoting another user's hateful speech in order to raise awareness of or critique that speech. "Some technology is vulnerable to false positives . . . , [for example], counter-speech, which should be encouraged," a representative from the D-Lab said.[133]

Civil society representatives interviewed recommended removing posts containing hate speech unless they explicitly disavow or criticize that speech.[134] In other words, the default presumption should be that posts containing hate speech are themselves hate speech, unless there is clear indication to the contrary.

Examples

2.1.4.1 Ethiopia

Tigrayans, human rights activists, and other observers of the conflict between Ethiopia's federal government and the TPLF have often shared hate speech for the purposes of condemning it. Image 9 and Image 10 depict such instances. Notably, both tweets very explicitly disavow the content that they are sharing. The former asks Meta to remove the content; the latter labels the content "hate speech" and claims that such statements have led to genocide in Tigray. Had the disclaimers been less explicit, the users' sharing of the posts would have been more likely to harm Tigrayans, even if they did not intend to do so.

Image 9: Tweet Condemning Dejene Assefa's Call to Violence[135]

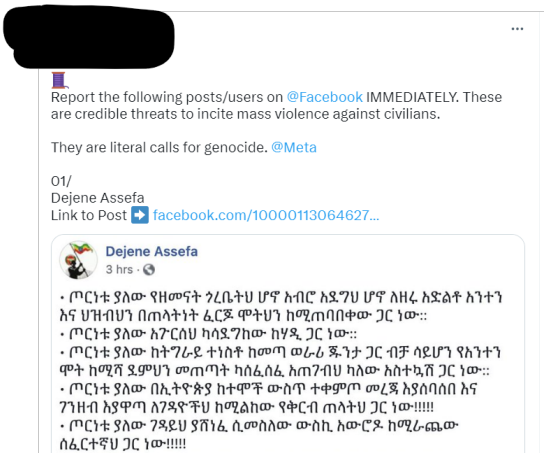


Image 10: Tweet Condemning Hate Speech Video[136]

2.2 OFFLINE FACTORS

2.2.1 Local Risk of Conflict

One factor that can distinguish hate speech from other types of speech is that it may contribute to real-world violence or persecution. The ICCPR prohibits speech "that constitutes incitement to discrimination, hostility, or violence" on the basis of nationality, race, or religion.[137] Meta's own policies prohibit similar content. Facebook's Community Standards prohibits targeting groups of people on the basis of "protected characteristics"[138] with violent speech, statements of inferiority, or calls for segregation.[139]

Of course, speech need not necessarily trigger violence or persecution to be considered hate speech. Many legal regimes[140] and civil society organizations[141] have adopted more expansive definitions of hate speech. Nevertheless, speech should be subject to additional scrutiny when it might contribute to real-world harms against particular groups.

One factor that affects the likelihood of online speech leading to real-world harms is the political situation in the area in which it takes place.[142] Specifically, inflammatory content posted in or around countries that are at a greater risk of violent conflict is more likely to contribute to violence or persecution. Therefore, Meta should prioritize content emanating from or discussing those countries for human review and potential moderation.

Measuring the risk of conflict in a particular country is a complicated exercise that depends on a variety of factors, including economic growth and the political participation of minority groups.[143] Fortunately, there are numerous indices that measure risk of conflict to which Meta could refer in determining whether certain content constitutes indirect hate speech. Some examples include:

- The European Commission's Global Conflict Risk Index, which expresses the statistical risk of violent conflict in a country within the next four years;[144]
- The Volatility and Risk Prediction Index, which assesses the immediate risk of an escalation of violence at the national level and within subnational administrative divisions;[145] and
- The Safety Perceptions Index, which measures citizens' worries and experiences of risk across 121 countries,[146] providing a subjective—rather than objective—measure of the risk of violence within a country.

This paper does not offer an opinion on which of these indices is most suitable to Meta's efforts to discern instances of indirect hate speech. These data sources are simply examples of existing efforts to quantify conflict risk.

Examples

2.1.1.1 Ethiopia

The fact that posts on Facebook and other social media have contributed to violence in politically sensitive areas is well documented.

In Ethiopia, political organizations, private citizens, and diaspora groups have engaged in social media campaigns in attempts to stoke violence against Tigrayan people.[147] Disaggregated dissemination of false information online has also bred conflict. For example, rumors on social media about "sleeper cells" leading to the federal government's defeat in Dessie and Kombolcha encouraged persecution of Tigrayans living outside of Tigray.[148] Civil plaintiffs have alleged that such activity on Facebook has led to concrete harms, including deaths.[149]

The example of Ethiopia illustrates the increased potential of social media content to harm vulnerable minorities during times of heightened political tensions and armed conflict.

2.1.1.2 Kenya

Kenya is an example of a country whose ethnic divisions produce local risks of conflict, especially during elections. Kenya has a history of ethnic violence and tribalism, which have produced violence and conflict around election cycles. The 2007–2008 Kenyan crisis emerged in the country when former President Mwai Kibaki was declared the winner of the 2007 presidential election. The disputed election "morphed into ethnic conflict"[150] and resulted in a violent political, economic, and humanitarian crisis. More than 1,200 Kenyans were reported killed with thousands more injured, over 300,000 people displaced, and approximately 42,000 houses and business looted or destroyed.[151] Five years after the 2007 post-election violence, Kenya had its first election under a new constitution.[152] Nevertheless, the shadow of the 2007 violence loomed over the 2013 elections.[153] Unfortunately, violence erupted again in the wake of the 2017 elections, leaving at least twenty-four people dead, "with some Kenyans fearing ethnic clashes similar to those triggered a decade" before.[154] As the election cycle was marked by murder and deadly protests, "much of the election chaos and violence stem[med] from tribal divisions."[155] In the aftermath of the 2017 elections, Kenya is a "bitterly divided country," as "[p]olitical and ethnic divisions exacerbated by the vote could lead to endemic violence and other destabilizing activity."[156]

In an interview with a representative from Amnesty International's branch in Kenya, the representative explained how the history in Kenya around ethnic polarization, especially around elections, is intense.[157] The representative described how Kenyans rarely think about tribes in the periods between elections, but when an election comes, they remember the strength of their tribal affiliations. Furthermore, the representative explained how Kenyans, especially in online circles, will use names and phrases to divide people. The representative pointed to examples of individuals using descriptive words for certain communities—words that are often degrading because of the polarized nature of the tribes and that are used to scare people from interacting with those communities. For example, the representative explained how agitated individuals of the Luo community uprooted a railway in 2007, and that action created a tagline to describe the Luo community: "*watu wa kung'oa reli*," meaning "people who destroy the railway."

Furthermore, the representative pointed to the shift to online hate during and after the 2017 elections. The representative detailed how there were online episodes on Facebook and Twitter exhibiting vitriol against certain communities but emphasized that no one was held accountable. Noting that 2017 was

an extended election, the representative described how the nation saw ethnic hatred all throughout that time, most of which occurred online. The representative explained the shift to online hate from 2013, when social media was not as embedded in the community, to 2017, when people understood the power of social media. Although Kenyans joined social media platforms between 2010 and 2013, the representative clarified that they had yet to develop literacy on the platforms or an understanding of the platforms' power. This eventual understanding marked the difference between 2013 and 2017. According to the representative, by 2017, people were pushing content online and falling into the trap of disinformation merchants, and social media's power to further divide communities became ever more apparent.

When asked whether ethnicity is only used as a political tool around elections, the representative explained how the online hate speech that Amnesty International has seen in Kenya is almost always on election-related issues. The representative discussed how, for example, ethnic hatred does not arise when dealing with budgets or raises for politicians; backlash in response to such issues are not perceived to be an ethnic issue, but rather are reflective of widespread anger against politicians. However, for issues directly involving citizens, politicians must go to their communities to build support, and in so doing, ethnic divides deepen between tribal communities.

Overall, the situation in Kenya is emblematic of the need for social media platforms to devote specific attention to countries with a history of ethnic violence. In the case of Kenya, this attention is particularly essential during election periods given that political divisions have the effect of deepening ethnic divisions. With a new election comes a new risk of local conflict, and individuals of the online community have become increasingly aware of the power for social media to solidify and exacerbate ethnic tensions.

2.1.1.3 Myanmar

In Myanmar, both the military and religious extremist groups used Facebook to disseminate hate speech during a period of escalating ethnic tension and the formalization of a Buddhist nationalist movement.[158] The armed forces established specialized social media units to spread propaganda and misinformation regarding Islam and Rohingya people.[159] These units included as many as seven hundred personnel.[160] Wirathu, a prominent monk whose anti-Muslim speech earned him the moniker "the Burmese Bin Laden," acknowledged the importance of Facebook in exposing people to his ideology. He told Buzzfeed News: "If the Internet had not come to [Myanmar], not many people would know my opinion and messages like now."[161]

In what was already a politically and ethnically sensitive context, content on Facebook contributed to widespread violence. According to an Amnesty International report, "the mass dissemination of messages that advocated hatred inciting violence and discrimination against the Rohingya . . . poured fuel on the fire of long-standing discrimination and substantially increased the risk of an outbreak of mass violence."[162]

2.2.2 Identity of the Speaker

The offline identity of the speaker is relevant both to understanding the meaning a given post and to assessing its potential to lead to harm. A person's offline statements or political affiliations can help shed light on ambiguous cases in which it is unclear whether the individual's post targets people on the basis of a protected characteristic. As one person who conducts social media research in Ethiopia plainly described: "If you know the person, you can [often] tell what they mean."[163]

Additionally, statements by prominent political figures, celebrities, or other individuals with large public followings may be more likely to influence their audiences' behavior. "Who says it matters a lot," said a Cameroon-based interviewee. "If some student goes to his Facebook page and calls [someone] names, it would not mean anything. But if a professor, a member of parliament, or a government minister made comments . . . , there will be a lot of people reacting."[164]

As such, the social media activity of these individuals should both be prioritized for review *and* potentially be subjected to a higher level of substantive scrutiny, unlike many of the other signals proposed here, given its especially potential for harm.

Of course, consideration of offline identities and statements is only possible for public figures.

Examples

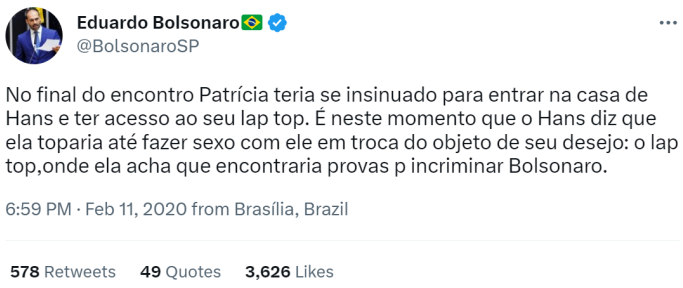
2.2.2.1 Brazil

Brazil offers a compelling example of how misinformation narratives by high-profile actors can function as hate speech—particularly when they target an individual based on their protected characteristic and exploit pre-existing prejudices. Specifically, Brazil provides an example of how online posts by high-profile politicians can produce a dangerous environment of misogyny against women, especially female journalists. Brazil has seen a rise in online misogyny against female journalists—namely, disinformation and online attacks[165]—a phenomenon that especially manifested itself during the 2018 election cycle. Specifically, the experience of award-winning journalist Patrícia Campos Mello reveals the danger of online hate speech against women, especially during election cycles. Relatedly, her experience points to the importance of recognizing how the behavior of high-profile actors online can affect vulnerable populations.

In 2018, Jair Bolsonaro sought the Brazilian presidency during the Brazilian general election. In October 2018, a week before Bolsonaro was elected, Campos Mello published an exposé[166] on his campaign's allegedly illegal use of WhatsApp to spread false new stories about his opponent in the election.[167] Following the release of the story, Campos Mello faced "a barrage of online threats against her and her family."[168] Her investigative reporting ultimately made her "the target of direct attacks from Bolsonaro, who fabricated sexist allegations that she 'tried to seduce' sources to aid her reporting, with misogyny being a key theme in the president's persistent attacks on journalists."[169] In 2020, during a congressional hearing on fake news in the nation's capital, Hans River Rio de Nascimento, a former employee of a digital marketing company, alleged in his testimony that Campos Mello lied in her investigative reporting and that she tried to extract information from him in exchange for sex.[170] According to Campos Mello, after Nascimento's statement went public, she "received hundreds of harassing messages on social media, and several politicians, including Eduardo Bolsonaro—a congressman and son of President Jair Bolsonaro—shared and repeated Nascimento's allegations on Twitter."[171]

The actions of Congressman Eduardo Bolsonaro have proven to be the most incriminating. After being criticized for stating in the chamber of deputies that he "[did] not doubt that Ms. Patrícia Campos Mello may have offered sexual favors, as Mr. Hans said, in exchange for information to try to harm President Jair Bolsonaro's campaign," he claimed that he had "done no more than repeat public testimony." However, he then proceeded to repeat the insinuations on his Twitter account, which has nearly two million followers.[172] The below reflects Eduardo Bolsonaro's Twitter activity against Campos Mello.

Image 11: Eduardo Bolsonaro's Twitter Activity Against Campos Mello[173]

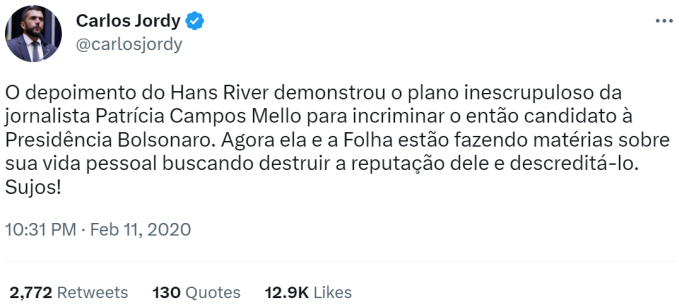


Translation from Google Translate: "At the end of the meeting, Patricia suggested that she enter Hans' house and gain access to his laptop. It is at this moment that Hans says that she would even agree to have sex with him in exchange for the object of her desire: the laptop, where she thinks she would find evidence to incriminate Bolsonaro."

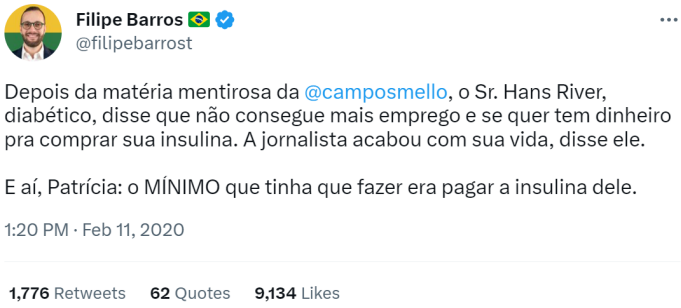
According to Reporters Without Borders ("RSF"): "The insinuations were then widely relayed on social media, triggering a new wave of sexist and misogynous insults and threats against her."[174] Emmanuel Colombié, the head of RSF's Latin America bureau, promulgated the following statement admonishing Bolsonaro's behavior: "The inflammatory behaviour of Eduardo Bolsonaro, whose statements deliberately sparked a new campaign of harassment against Patrícia Campos Mello, are completely unacceptable and unworthy of a parliamentarian."[175]

In addition to Eduardo Bolsonaro, two other congressmen, Carlos Jordy[176] and Filipe Barros,[177] repeated the accusations against Campos Mello on Twitter,[178] further fueling the online misogynist campaign against her. The below reflects their Twitter activity.

Image 12: Twitter Activity Against Campos Mello by Carlos Jordy and Filipe Barros[179]



Translation from Google Translate: "Hans River's testimony demonstrated the unscrupulous plan of journalist Patrícia Campos Mello to incriminate the then presidential candidate Bolsonaro. Now she and Folha are doing stories about his personal life, seeking to destroy his reputation and discredit him. Dirty!"



Translation from Google Translate: "After the lying article of @camposmello, Mr. Hans River, a diabetic, said he can no longer find a job and if he wants he has money to buy his insulin. The journalist ended her life, he said. So, Patrícia: the MINIMUM you had to do was pay for his insulin."

Notably, in 2020, Campos Mello sued Eduardo Bolsonaro in a civil court for moral damages and won.[180] Following the lawsuit, in an interview with Committee to Protect Journalists, Campos Mello

explained how "[m]embers of [the Brazilian] government have often used smear campaigns against journalists, including misogynistic [campaigns]."[181] When asked what journalists, especially women, can do to protect themselves from online harassment, Campos Mello responded:

> It is important to remember that online harassment does not just stay within the digital sphere—it spills over. You start receiving threats and being harassed on the street. Often, those who push online harassment say it's a joke, and that we have no sense of humor. But online harassment is used to intimidate journalists and has been used systematically. Social media platforms are making progress in regard to the spread of misinformation. But when it comes to online harassment, we have not seen significant advancement. They are not yet able to prevent a harassment campaign.[182]

In a separate interview, Campos Mello pointed to the dangers of being a woman and a journalist in Latin America, and to the potential for violence via online platforms:

> I realized that in Brazil—and I'm sure this is true for other parts of Latin America too—as a woman, people feel authorized to use all sorts of misogynistic and personal attacks against us. It's way more aggressive than anything that happens to male journalists. Professional media in general is under attack in Brazil, but for male journalists, it's always a different kind of attack: It's never that personal, and it's never against their family.[183]

After facing online harassment for years, Campos Mello has called for more action to address hate speech online. In an interview with *Foreign Policy*, Campos Mello shared:

> In 2018, when the attacks started, Facebook was not helpful at all—even though I explained that tons of fake content about me was being disseminated, they said they couldn't do anything. In 2020, Twitter was more proactive, whereas Facebook and other social media platforms were still very slow to react. I definitely think they need to do more regarding hate speech. I think they are not investing enough resources to counter hate speech, be it [artificial intelligence] or real people to moderate.[184]

Campos Mello's experiences being the target of online hate speech and harassment span the spectrum from direct to indirect attacks. Specifically, the attacks she faced from high-profile government officials in Brazil reveal the value of creating signals for the identity of the poster—and relatedly, here, for reach and engagement. On their face, the social media posts of the Brazilian congressmen would not necessarily be considered to constitute direct hate speech—arguably, as Congressman Eduardo Bolsonaro alleged, one could perceive such posts as merely repeating information from a testimony. However, greater evaluation reveals the inherent harm that exists in passing along such information. As shown by Campos Mello's experience, this is especially the case where the conveyor of information holds high-profile status in the political field and has followers who will respond to this information in a harmful manner, thus perpetuating the cycle of online hate speech.

By broadcasting and perpetuating harmful information on the Internet, the politicians' online activity had the effect of creating a harassment campaign against a female journalist and feeding into a dynamic of online misogyny, one already rampant in Latin America. However, the negative effects of such hate speech are not limited to female journalists. These effects likely will reach female politicians and women in other prominent positions. Notably, in Brazil, female politicians have been the subject of misogynist discourse and attacks, which "have rapidly increased since the election of Jair Bolsonaro. . . . The election of an openly misogynist president in 2018 opened the floodgates for violence against women in politics."[185] Because

elections function "like a trigger for the advancement of hate speech,"[186] female politicians are especially at risk of online hate speech attacks.

The content of the post alone may not always reveal the possibility of harmful effects on vulnerable groups, but the identity of the individual behind the post can be especially telling when that individual maintains high-profile status and has a significant following. Campos Mello's experience is merely one case study where the seemingly innocuous posts of high-profile politicians online can produce an environment of online and offline misogyny.

2.2.3 Identity of the Target

Just as not all speakers are equally likely to engage in hate speech, not all protected characteristics are equally likely to be the subject of hate speech. Especially in conflict settings, particular groups are likely to suffer violence. Insofar as the goal of identifying hate speech is to avoid concrete harms, Meta's content moderation efforts should be focused on the groups that are most likely to experience harm. For instance, both Uyghur and Han Chinese people are potential targets of hate speech on the basis of their ethnicities. However, in China in 2023, Uyghur people are more likely to suffer ethnic persecution. Online content moderation for a social media platform operating in China would therefore focus especially on speech targeting Uyghurs.

Considering the identity of the target is important both for prioritizing posts for human review and also for determining whether a particular post meets the proposed threshold of a "harmful attack." Violent language is more likely to cause offline harm to marginalized or vulnerable groups. Meta's enforcement of its hate speech policy should recognize this dynamic, rather than treating all protected characteristics as equally likely to be subjects of hate speech.

Of course, focusing on specific potential targets of hate speech does not mean that other groups should not receive any protection. Furthermore, the identity groups that are most vulnerable to hate speech shift over time, and keeping abreast of these changes is essential for leveraging this signal toward more effective content moderation.

Examples

In each of the contexts discussed in this section, there are particular groups that are most likely to suffer from violence or discrimination as a result of hate speech. These groups may be especially vulnerable because they are minorities within a country, politically or economically marginalized, or have suffered from discrimination and dehumanization in the past. In Cameroon, Anglophones are vulnerable in-part because of their minority status. In Ethiopia, Tigrayans are vulnerable, in part because of their loss of political power within the federal government in 2018. In Myanmar, the Rohingya people suffered decades of government discrimination and repression even before the ethnic cleansing campaign that began in 2017.

3 CONCLUDING REMARKS

Having laid out the signals framework, in this final part, this paper will provide an overview of this framework and discuss some recommendations that Meta should incorporate to effectively address hate speech in both its direct and indirect forms.

This paper makes two significant recommendations. First, it proposes a broader definition of hate speech in the Facebook Community Standards in light of relevant human rights law and expert insights. This definition should recognize that hate speech does not need to be either direct or intentional; rather, it can cause significant harm even if it is indirect or uses coded language. Notably, though, the definition would require that attacks be harmful in order to ensure that Meta preserves an environment of free expression and

only interferes with speech when doing so is necessary to protect the human rights of others. This new definition most clearly addresses the problem of ensuring that Meta does not, as a blanket policy matter, permit speech that is hateful under international legal standards and poses a significant risk of concrete harm. As to the second issue raised in this white paper—Meta's poor enforcement of hate speech clearly falling with its existing policies—the effects of this recommendation require a meaningful and substantial commitment on the part of Meta. Expanding the hate speech definition will provide greater clarity to its underlying principles and purpose, thereby assisting those who design the machine-learning content moderation tools as well as the human moderators. It cannot, however, detract from Meta's overall obligation to address the so-called low-hanging fruit of direct hate speech, which is clearly within the confines of the existing Meta definition and human rights norms and standards. A broader definition of hate speech is necessarily accompanied by tougher decisions on whether something constitutes hate speech: there are, of course, good-faith debates that can be had among people as to whether something is indirect hate speech or something more innocuous. That said, Meta must dedicate the necessary resources—whether intellectual or financial—to moderate both direct and indirect hate speech.

The second recommendation that this white paper makes is a more holistic approach to online hate speech that relies on a system of signals to help identify and prioritize content that should be flagged, removed, or otherwise sanctioned. This signals framework aims to capture both the complexity and real-world circumstances of hate speech as well as its variable impact on individuals, communities, and relationships. It also seeks to effectively address the hate speech hardest to identify because it is indirect or coded by using a holistic framework, comprised of various factors identified during desk research and the interviews with experts in this area. Using this framework is not intended to enable a clear determination of whether something is indirect hate speech in all instances; indeed, one of the main principles underlying the notion of indirect hate speech is that it often involves difficult cases, where reasonable people could disagree as to whether the post should be characterized as hate speech. However, this framework provides an actionable approach that would provide guidance as to which content to prioritize in hate speech enforcement. Beyond a determination by Meta's technical content moderation experts as to how to operationalize this framework, this signaling framework would likely require several additional policy steps and resource investiture, as identified below.

- **Invest in an expanded content moderation program that will produce a skilled, trained, unbiased, and stable cohort of moderators**: Even the best algorithms will prove unable to identify significant amounts of hate speech, particularly more indirect hate speech. And currently, Meta algorithms are letting through the vast majority of hate speech. A far higher number of human moderators are needed to review potential hate speech, particularly in high-risk contexts where the risk of violence or other severe consequences increases as long as the post remains viewable. These moderators must collectively speak and understand the local languages, not just the primary or official languages, but all languages spoken by a significant number of individuals who are on social media. Particularly in the context of emerging or active conflicts, the individuals involved with the content moderation process should be, as much as possible, unbiased and unaffiliated with actors that have demonstrated a record of significant human rights abuses, including governments.
- **Update algorithms routinely in light of new speech and uses for words**: Language is dynamic and unfortunately, that means that hate speech is as well. Algorithms must be constantly updated, not yearly but as often as technologically feasible, and certainly in response to new emerging or active conflicts. These changes should be publicized and promoted to civil society groups, particularly those working in the context of emerging or active conflicts to ensure that they are as accurate as possible.

- **Increase deferrals to humans in areas of emerging/active conflict or for high-profile political/other figures with a lot of engagement on Facebook**: Because the risk of harm of hate speech is higher in Ethiopia than the United States (and the importance of counter-speech is similarly heightened), Meta should filter more posts through their algorithms to humans in situations where there is emerging or active conflict. The same approach should be taken for high-profile political figures based on the same logic.
- **Examine what is already out there with respect to algorithms**: There are many organizations working on more robust algorithms. Even if it would require more resources and even if there would be some implementation curve, it is apparent that Meta may have technological options that may neatly integrate with the signals framework proposed herein.

Online hate speech is a problem of immense importance that has no easy solutions. In the midst of complex and intersecting questions of technology, law, and policy, this white paper proposes a multifaceted framework that aims to capture these complexities while still being workable in the context of the rapid, large-scale enforcement that is crucial in situations of emerging or present conflict. It also urges the Oversight Board not to focus on indirect hate speech at the expense of direct hate speech, which remains an immense concern for human rights activists and an area that still requires vast improvement from Meta.

[1] This memorandum was researched and written by Yale Law School students Justin Cole, Ali Hakim and Camilla Suarez under the supervision of Claudia Flores, clinical professor of law and director of the Yale Law School Lowenstein International Human Rights Clinic.

[2] *Alleged Crimes in Raya Kobo*, 2021-014-FB-UA, available at https://www.oversightboard.com/decision/FB-MP4ZC4CC.

[3] *See, e.g.*, Steve Stecklow, *Why Facebook Is Losing the War on Hate Speech in Myanmar*, REUTERS (Aug. 15, 2018), https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/; Shirin Ghaffary, *Civil Rights Leaders Are Still Fed Up with Facebook Over Hate Speech*, VOX (July 7, 2020, 6:54 PM), https://www.vox.com/recode/2020/7/7/21316681/facebook-mark-zuckerberg-civil-rights-hate-speech-stop-hate-for-profit; Barbara Ortutay, *Facebook's System Approved Dehumanizing Hate Speech*, PBS (June 9, 2022, 11:55 AM), https://www.pbs.org/newshour/world/facebooks-system-approved-dehumanizing-hate-speech; Noah Giansiracusa, *Facebook Uses Deceptive Math to Hide Its Hate Speech Problem*, WIRED (Oct. 15, 2021, 7:00 AM), https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/.; Global Witness, *New Investigation Shows Facebook Approves Ads Containing Hate Speech Inciting Genocide Against the Rohingya* (Mar. 21, 2022), https://www.globalwitness.org/en/press-releases/new-investigation-shows-facebook-approves-ads-containing-hate-speech-inciting-genocide-against-rohingya/ [hereinafter Global Witness, Myanmar Investigation]; Global Witness, *'Now Is the Time to Kill': Facebook Continues to Approve Hate Speech Inciting Violence and Genocide During Civil War in Ethiopia* (June 9, 2022), https://www.globalwitness.org/en/campaigns/digital-threats/ethiopia-hate-speech/ [hereinafter Global Witness, Ethiopia Investigation]; Global Witness, *Facebook Unable to Detect Hate Speech Weeks Away from Tight Kenyan Election* (July 8, 2022), https://www.globalwitness.org/en/campaigns/digital-threats/hate-speech-kenyan-election/ [hereinafter Global Witness, Kenya Investigation]; *see also* Elizabeth Dwoskin, Nitasha Tiku & Craig Timberg, *Facebook's Race-Blind Practices Around Hate Speech Came at the Expense of Black Users, New Documents Show*, WASH. POST (Nov. 21, 2021, 8:00 AM), https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race/ (commenting on how Meta executives opposed a plan that would target hate speech targeted against women of color in the interest of being "race-neutral"); Billy Perrigo, *Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch*, TIME MAG. (Nov. 26, 2019, 12:40 PM), https://time.com/5739688/facebook-hate-speech-languages/ (noting that Meta's hate speech algorithms only work in certain languages).

[4] Global Witness cooperated with Foxglove, a legal non-profit organization, on the Ethiopia and Kenya investigations.

[5] *See* Amnesty Int'l, *Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations* (Sept. 29, 2022), https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/; *see also* Alexandra Stevenson, *Facebook Admits It Was Used to Incite Violence in Myanmar*, N.Y. TIMES (Nov. 6, 2018), https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html (describing a Meta executive admitting that the company had failed to prevent its platform from being used to "foment division and incite online violence"); *cf.* Dan Milmo, *Rohingya Sue Facebook for £ 150bn over Myanmar Genocide*, GUARDIAN (Dec. 6, 2021), https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence (victims of the Rohingya genocide accusing Meta of being "willing to trade the lives of the Rohingya people for better market penetration in a small country in south-east Asia").

[6] Ctr. for Preventive Action, *War in Ethiopia*, COUNCIL ON FOREIGN RELS. (Feb. 7, 2023), https://www.cfr.org/global-conflict-tracker/conflict/conflict-ethiopia; *see also* Lauren Jackson, *Ending a Civil War*, N.Y. TIMES (Dec. 11, 2022), https://www.nytimes.com/2022/12/11/briefing/ethiopia-war-tigray.html.

[7] *See* Hum. Rts. Watch, *Ethiopia: Crimes Against Humanity in Western Tigray Zone* (Apr. 6, 2022, 1:00 AM), https://www.hrw.org/news/2022/04/06/ethiopia-crimes-against-humanity-western-tigray-zone; U.N. Off. of the High Comm'r on Hum. Rts., *UN Experts Warn of Potential for Further Atrocities Amid Resumption of Conflict in Ethiopia* (Sept. 19, 2022), https://www.ohchr.org/en/press-releases/2022/09/un-experts-warn-potential-further-atrocities-amid-resumption-conflict; Amnesty Int'l, *Ethiopia: Fears of Fresh Atrocities Loom in Tigray as Conflict Intensifies* (Oct. 24, 2022), https://www.amnesty.org/en/latest/news/2022/10/ethiopia-fears-of-fresh-atrocities-loom-in-tigray-as-conflict-intensifies/; Michael Crowley & Declan Walsh, *Blinken Calls for 'Accountability' on War Crimes in Ethiopia*, N.Y. TIMES (Mar. 15, 2023), https://www.nytimes.com/2023/03/15/world/africa/blinken-abiy-ethiopia.html; Lizzy Davies, *Ethiopia Accused of 'Serious' Human Rights Abuses in Tigray in Landmark Case*, GUARDIAN (Feb. 8, 2022, 8:10 AM), https://www.theguardian.com/global-development/2022/feb/08/ethiopia-human-rights-abuses-possible-war-crimes-

tigray; AMNESTY INT'L & HUM. RTS. WATCH, "WE WILL ERASE YOU FROM THIS LAND": CRIMES AGAINST HUMANITY AND ETHNIC CLEANSING IN ETHIOPIA'S WESTERN TIGRAY ZONE (2022), available at https://www.hrw.org/report/2022/04/06/we-will-erase-you-land/crimes-against-humanity-and-ethnic-cleansing-ethiopias. World Health Organization Director General Tedros Adhanom Ghebreyesus, who previously served as health minister and foreign affairs minister in Ethiopia, even expressed in October 2022 that there was a "very narrow window now to prevent genocide" in Tigray. Al Jazeera, *WHO Chief Warns Time Running Out to 'Prevent Genocide' in Tigray* (Oct. 19, 2022), https://www.aljazeera.com/news/2022/10/19/who-chief-tedros-preventing-genocide-tigray-ethiopia.

[8] *See* Int'l Crisis Grp., *Kenya's 2022 Election: High Stakes* (June 9, 2022), https://www.crisisgroup.org/africa/horn-africa/kenya/kenyas-2022-election-high-stakes; Clionadh Raleigh & Caleb Wafula, *Kenya's Political Violence Landscape in the Lead-Up to the 2022 Elections*, ARMED CONFLICT LOCATION & EVENT DATA PROJECT (Aug. 9, 2022), https://acleddata.com/2022/08/09/kenyas-political-violence-landscape-in-the-lead-up-to-the-2022-elections/; Saskia Brechenmacher & Nanjira Sambuli, *The Specter of Politics as Usual in Kenya's 2022 Election*, CARNEGIE ENDOWMENT FOR INT'L PEACE (July 27, 2022), https://carnegieendowment.org/2022/07/27/specter-of-politics-as-usual-in-kenya-s-2022-election-pub-87578.

[9] *See* Global Witness, Ethiopia Investigation, *supra* note 3; Global Witness, Kenya Investigation, *supra* note 3.

[10] *See supra* note 3.

[11] *See* Global Witness, Myanmar Investigation, *supra* note 3.

[12] *See* Global Witness, Ethiopia Investigation, *supra* note 3.

[13] *See* Global Witness, Kenya Investigation, *supra* note 3.

[14] Meta, https://www.facebook.com/business/about/ad-principles (last visited Apr. 24, 2023).

[15] *See* Global Witness, Ethiopia Investigation, *supra* note 3.

[16] *See supra* notes 11-13.

[17] *See infra* Section I.B.

[18] In response to an *Associated Press* article on the Global Witness Myanmar investigation, Raphael Frankel, director of public policy for emerging markets at Meta Asia Pacific, indicated: "We've built a dedicated team of Burmese speakers, banned the Tatmadaw, disrupted networks manipulating public debate[,] and taken action on harmful misinformation to help keep people safe. We've also invested in Burmese-language technology to reduce the prevalence of violating content." Victoria Milko & Barbara Ortutay, *'Kill More': Facebook Fails to Detect Hate Against Rohingya*, ASSOCIATED PRESS (Mar. 21, 2022), https://apnews.com/article/technology-business-bangladesh-myanmar-united-nations-f7d89e38c54f7bae464762fa23bd96b2. Frankel added: "This work is guided by feedback from experts, civil society organizations[,] and independent reports, including the [United Nations] Fact-Finding Mission on Myanmar's findings and the independent Human Rights Impact Assessment we commissioned and released in 2018." *Id.* In the context of Ethiopia, Mercy Ndegwa, Meta's public policy director for East & Horn of Africa said: "For more than two years, we've invested in safety and security measures in Ethiopia, adding more staff with local expertise and building our capacity to catch hateful and inflammatory content in the most widely-spoken languages, including Amharic, Oromo, Somali[,] and Tigrinya." Moreover, "[a]s the situation has escalated, we've put additional measures in place and are continuing to monitor activity on our platform, identify issues as they emerge, and quickly remove content that breaks our rules." Jasper Jackson, Lucy Kassa & Mark Townsend, *Facebook 'Lets Vigilantes in Ethiopia Incite Ethnic Killing'*, GUARDIAN (Feb. 20, 2022), https://www.theguardian.com/technology/2022/feb/20/facebook-lets-vigilantes-in-ethiopia-incite-ethnic-killing. Meta has described Ethiopia as "one of [its] highest priorities for country-specific interventions to keep people safe given the risk of conflict." *An Update on Our Longstanding Work to Protect People in Ethiopia*, META (Nov. 9, 2021), https://about.fb.com/news/2021/11/update-on-ethiopia/. In Kenya, the story is similar: Ndegwa has described "working closely with election authorities and trusted partners in the country" and "taking to the airwaves on local radio in Kenya[] to educate listeners on how to spot hate speech." Mercy Ndegwa, *How Meta Is Preparing for Kenya's 2022*

General Election, META (July 20, 2022), https://about.fb.com/news/2022/07/how-metas-preparing-for-kenyas-2022-general-election/.
[19] Global Witness, Ethiopia Investigation, *supra* note 3.
[20] Giansiracusa, *supra* note 3.
[21] Meta, *Facebook Community Standards*, https://transparency.fb.com/policies/community-standards/ (last visited Apr. 24, 2023).
[22] *Id.*
[23] Meta, *Hate Speech*, https://transparency.fb.com/policies/community-standards/hate-speech/ (last visited Apr. 24, 2023).
[24] *Id.*
[25] Meta, *How We Create and Use Market-Specific Slur Lists*, https://transparency.fb.com/enforcement/taking-action/how-we-create-and-use-market-slurs (last visited Apr. 24, 2023).
[26] *Id.*
[27] *South Africa Slurs*, 2021-011-FB-UA, available at https://www.oversightboard.com/decision/FB-TYE2766G/.
[28] *Russian Poem*, 2022-008-FB-UA, available at https://www.oversightboard.com/decision/FB-MBGOTVN8/.
[29] Interview with Mahlet Gebremedhin, Omna Tigray (Mar. 15, 2023) (interview recording on file, Lowenstein International Human Rights Clinic).
[30] International human rights conventions do not use the term "hate speech." This white paper uses the term here as a shorthand for the types of speech that international human rights law proscribes on the basis of its discriminatory content or effects.
[31] Several widely ratified conventions broadly protect free expression. The International Convention on Civil and Political Rights and the Universal Declaration of Human Rights enshrine the right to "impart information and ideas of all kinds." G.A. Res. 217 (III) A, Universal Declaration of Human Rights art. 19 (Dec. 10, 1948); International Covenant on Civil and Political Rights art. 19, Dec. 16, 1966, 999 U.N.T.S. 171 [hereinafter ICCPR]. Similar provisions appear in regional human rights treaties, including the European Convention on Human Rights and the American Convention on Human Rights. European Convention for the Protection of Human Rights and Fundamental Freedoms art. 10, Sept. 3, 1953, E.T.S. 5; American Convention on Human Rights "Pact of San Jose, Costa Rica" art. 13, Nov. 22, 1969, 1144 U.N.T.S. 123 [hereinafter ACHR]. The International Convention on the Elimination on All Forms of Racial Discrimination prohibits parties from denying these rights on the basis of race or ethnicity. International Convention on the Elimination of All Forms of Racial Discrimination art. 5, Dec. 21, 1965, 660 U.N.T.S. 195 [hereinafter ICERD]. For more on convergence and divergence on hate speech see Evelyn Aswad & David Kaye, *Convergence & Conflict: Reflections on Global and Regional Human Rights Standards on Hate Speech*, 22 NW. J. HUM. RTS. 165 (2022).
[32] ICCPR, *supra* note 31, art. 20(2).
[33] ACHR, *supra* note 31, art. 13(5).
[34] ICERD, *supra* note 31, art. 4(a).
[35] ICCPR, *supra* note 31, art. 19(3).
[36] Sanchez v. France, Eur. Ct. H.R. 724 (2021).
[37] *Id.*
[38] *Robert Faurisson v. France*, Communication No. 550/1993, U.N. Doc. CCPR/C/58/D/550/1993 (1996).
[39] Human Rights Council Res. 16/4, U.N. Doc. A/67/357, at ¶ 46 (Sept. 7, 2012).
[40] *Malcolm Ross v. Canada*, CCPR/C/70/D/736/1997, U.N. Hum. Rts. Comm., Oct. 26, 2000, available at https://www.refworld.org/cases,HRC,3f588efc0.html.
[41] Rabat Plan of Action, U.N. Doc. A/HRC/22/17/Add.4, at 11 (2013).
[42] *E.g.*, ICCPR, *supra* note 31, art. 20.
[43] According to the United Nations Special Rapporteur on Minority Issues, many governments do not maintain any data on instances of hate speech on social media. In many countries, there is no record of domestic legislation against hate speech being used in the context of social media, either because the legislative framework does not capture online hate speech or because the requirements for prosecution are too onerous. Furthermore, domestic law typically does not subject social media companies to fines or penalties for failing to moderate hateful content. On the flip side, states have

successfully "regulated" social media use in an attempt to suppress dissent as well. *See* Sanja Kelly et al., *Manipulating Social Media to Undermine Democracy*, FREEDOM HOUSE (2017), https://freedomhouse.org/report/freedom-net/2017/manipulating-social-media-undermine-democracy.

[44] U.N. Off. of the High Comm'r of Hum. Rts., *Guiding Principles on Business and Human Rights* (2011), https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf [hereinafter Guiding Principles].

[45] U.N. OFF. ON GENOCIDE PREVENTION AND THE RESPONSIBILITY TO PROTECT, UN STRATEGY AND PLAN OF ACTION ON HATE SPEECH 4 (2019), available at https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml.

[46] Guiding Principles, *supra* note 44.

[47] U.N. SPECIAL RAPPORTEUR ON MINORITY ISSUES, DRAFT EFFECTIVE GUIDELINES ON HATE SPEECH, SOCIAL MEDIA, AND MINORITIES 6 (2022), available at https://www.ohchr.org/sites/default/files/2022-06/Draft-Effective-Guidelines-Hate-Speech-SR-Minorities.pdf.

[48] *Id.*

[49] *Id.* at 7.

[50] *Id.* at 9.

[51] *Id.*

[52] *See, e.g.*, Special Rapporteur for Freedom of Expression, *Standards for a Free, Open, and Inclusive Internet*, INTER-AMERICAN COMM'N ON HUM. RTS. (Mar. 15, 2017), http://www.oas.org/en/iachr/expression/docs/publications/internet_2016_eng.pdf; Special Rapporteur on Freedom of Expression and Access to Info. in Afr., Press Release by the Special Rapporteur on Freedom of Expression and Access to Information in Africa on the Continuing Trend of Internet and Social Media Shutdowns in Africa, AFR. COMM'N ON HUM. AND PEOPLE'S RTS. (Jan. 29, 2019), https://www.achpr.org/pressrelease/detail?id=8.

[53] Interview with Andre Oboler, Chief Executive Officer, Online Hate Prevention Institute (Feb. 26, 2023) (interview recording on file, Lowenstein International Human Rights Clinic).

[54] Interview with Representative, University of California, Berkeley's D-Lab (Feb. 28, 2023) (interview recording on file, Lowenstein International Human Rights Clinic).

[55] *Anti-Asian Racism in Australian Social Media*, ONLINE HATE PREVENTION INST. (Oct. 2022), https://ohpi.org.au/anti-asian-racism-in-australian-social-media/.

[56] *Constructing Internal Variables via Faceted Rasch Measurement and Multitask Deep Learning: A Hate Speech Application*, CORNELL UNIV. (Sept. 22, 2020), https://arxiv.org/abs/2009.10277.

[57] Interview with Professor Robert Post, Yale L. Sch. (Feb. 27, 2023).

[58] Interview with Representatives, U.N. Off. of the High Comm'r for Hum. Rts. (Mar. 22, 2023).

[59] Interview with Representative, Nat'l Cohesion & Integration Comm'n Kenya (Mar. 23, 2023).

[60] Interview with Representative, Amnesty International Kenya (Mar. 24, 2023).

[61] Interview with Representative, Digital Threats at Global Witness (Mar. 24, 2023) (interview recording on file, Lowenstein International Human Rights Clinic).

[62] Interview with Founder & Exec. Dir., Rsch. & Advoc. Org. (Mar. 1, 2023) [hereinafter Interview with Anonymous].

[63] Interview with Anonymous, *supra* note 62; Interview with Mahlet Gebremedhin, *supra* note 29; Interview with Ngala Desmond, Country Project Manager, #DefyHateNow Cameroon (Mar. 28, 2023) (interview recording on file, Lowenstein International Human Rights Clinic).

[64] Meta, *Taking Action*, https://transparency.fb.com/enforcement/taking-action/.

[65] Mai ElSherief et al., *Latent Hatred: A Benchmark for Understanding Implicit Hate Speech*, (2021), http://arxiv.org/abs/2109.05322; Thomas Davidson et al., *Automated Hate Speech Detection and the Problem of Offensive Language*, 11 PROC. INT'L AAAI CONF. WEB SOC. MEDIA 512 (2017); Jherez Taylor, Melvyn Peignon & Yi-Shin Chen, *Surfacing Contextual Hate Speech Words within Social Media* (2017), http://arxiv.org/abs/1711.10093; Rijul Magu & Jiebo Luo, *Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks*, *in* PROCEEDINGS OF THE 2ND WORKSHOP ON ABUSIVE LANGUAGE ONLINE (ALW2) 93 (2018), https://aclanthology.org/W18-5112.

[66] Interview with Mahlet Gebremedhin, *supra* note 29.

[67] Interview with Ngala Desmond, *supra* note 63.
[68] Interview with Tom Andrews, Special Rapporteur on the Situation of Human Rights in Myanmar, United Nations (Mar. 24, 2023) (interview recording on file, Lowenstein International Human Rights Clinic).
[69] Interview with Andre Oboler, *supra* note 533.
[70] ROGER BROMLEY, BEAST, VERMIN, INSECT—HATE MEDIA AND THE CONSTRUCTION OF THE ENEMY: THE CASE OF RWANDA, 1990 – 1994 (2011).
[71] Interview with Anonymous, *supra* note 62.
[72] Davidson et al., *supra* note 65.
[73] Seung-Ho Kang & Sam-Kyu Kim, *Hate Speech and Usage of Japanese in Korean Insect Common Name*, 60 KOREAN J. APPL. ENTOMOL. 155 (2021).
[74] Erin Steuter & Deborah Wills, *Discourses of Dehumanization: Enemy Construction and Canadian Media Complicity in the Framing of the War on Terror*, 2 GLOB. MEDIA J. CAN. ED 7 (2009).
[75] Interview with Professor Post, *supra* note 57.
[76] BRINGING LOCAL CONTEXT INTO OUR GLOBAL STANDARDS, https://transparency.fb.com/policies/improving/bringing-local-context/ (last visited Apr. 24, 2023); Interview with Ngala Desmond, *supra* note 63.
[77] Nicola Barrach-Yousefi & Althea Middleton-Detzner, *Hate Speech and Conflict in the Federal Democratic Republic of Ethiopia*, PEACETECH LABS, at 19 (2021), https://static1.squarespace.com/static/54257189e4b0ac0d5fca1566/t/60bfaa27a19f0752ecd1426d/1623173770820/EthiopiaLexicon2021_web.pdf.
[78] *Id.* at 36.
[79] *Id.*
[80] *Id.*
[81] *Id.* at 10.
[82] Aman Jr. (@AmanJr42270485), TWITTER (Jan. 1, 2022, 5:56 AM), https://twitter.com/AmanJr42270485/status/1477232344935419906.
[83] Meta's current policy only protects immigrants from especially egregious forms of violent or dehumanizing speech, which it calls "Tier 1" attacks. *Hate Speech*, *supra* note 23.
[84] *Burma's Path to Genocide*, U.S. HOLOCAUST MEMORIAL MUSEUM (last visited Apr. 24, 2023), https://exhibitions.ushmm.org/burmas-path-to-genocide/chapter-3/hate-speech-that-claims-rohingya-do-not-belong-in-burma.
[85] Interview with Ngala Desmond, *supra* note 63.
[86] #defyhatenow, FIELD GUIDE CAMEROON 67 (2021), available at https://defyhatenow.org/wp-AAngcontent/uploads/2022/05/DHN_cameroon_field_guide_EN_2021_chapter6.pdf.
[87] *Id.*
[88] *Id.* at 66.
[89] Célian Macé, *"L'Ethiopie n'est plus un pays pour moi": La Confiance Brisée des Tigréens d'Addis-Abeba*, LIBÉRATION (Mar. 15, 2023).
[90] *Meta Sued for 2bn over Ethiopia Hate Speech Revealed by Bureau*, THE BUREAU OF INVESTIGATIVE JOURNALISM (Dec. 14, 2022), https://www.thebureauinvestigates.com/stories/2022-12-14/meta-sued-for-2bn-over-ethiopia-hate-speech-revealed-by-bureau.
[91] Ras Bin (@Bini45251286), TWITTER (Nov. 5, 2021, 4:31 AM), https://twitter.com/Bini45251286/status/1456539644947013672.
[92] *WHO Board Halts Ethiopia's Anti-Tedros Speech, Postpones Probe Decision*, REUTERS (Jan. 25, 2022), https://www.reuters.com/world/africa/who-sets-aside-ethiopias-request-probe-who-chiefs-links-rebellious-tigrayan-2022-01-24/.
[93] Barrach-Yousefi & Middleton-Detzner, *supra* note 62, at 10.
[94] *Fire and Fury: Anti-Ukrainian Hate Speech on Russian Social Networks*, UKRAINE WORLD (Sept. 13, 2019), https://ukraineworld.org/articles/infowatch/anti-ukrainian-hate-speech-VK.

[95] *When Words Kill—From Moscow to Mariupol*, EUVSDISINFO (June 17, 2022), https://euvsdisinfo.eu/when-words-kill-from-moscow-to-mariupol/.
[96] Mirce Adamcevski, *The War in Ukraine, Freedom of Expression and Hate Speech*, RESPUBLICA (Sept. 6, 2022), https://respublica.edu.mk/blog-en/mediums/the-war-in-ukraine-freedom-of-expression-and-hate-speech/?lang=en.
[97] *Fire and Fury*, *supra* note 94.
[98] *Id.*
[99] Jade McGlynn, *Russia's Imperial Arrogance Is Destroying Ukrainian Heritage*, FOREIGN POL'Y (May 30, 2022, 7:00 AM), https://foreignpolicy.com/2022/05/30/russias-imperial-arrogance-is-destroying-ukrainian-heritage/.
[100] *Id.*
[101] *Fire and Fury*, *supra* note 94.
[102] *Id.*
[103] Pawet Trzaskwoski, *'Ukrainization' Becomes Dangerous Word as Refugee Crisis Continues*, DANGEROUS SPEECH PROJECT (Feb. 16, 2023), https://dangerousspeech.org/ukrainization-becomes-a-dangerous-word-as-refugee-crisis-continues/.
[104] *Id.* The pamphlet is available at the following link: https://konfederacjakoronypolskiej.pl/wp-content/uploads/2022/07/SUP-luz.pdf.
[105] *Id.*
[106] #StopUkrainizacjiPolski, TWITTER, https://twitter.com/hashtag/stopukrainizacjipolski.
[107] *Independence March AD 2022*, KONFEDERACJA KORONY POLSKIEJ (Nov. 11, 2022), https://konfederacjakoronypolskiej.pl/bylismy-na-marszu/.
[108] Andrew Higgins, *How Poland, Long Leery of Foreigners, Opened Up to Ukrainians*, N.Y. TIMES (Feb. 22, 2023), https://www.nytimes.com/2023/02/22/world/europe/poland-ukraine-war-refugees.html.
[109] *Id.*
[110] @colonelhomsi, TWITTER (Feb. 28, 2023, 4:48 PM), https://twitter.com/colonelhomsi/status/1630686301543444485.
[111] *"Ukrainization" in Pro-Russian Propaganda in Romania, Poland, Serbia and Hungary*, GLOBALFOCUS (Aug. 8. 2022), https://www.global-focus.eu/2022/08/ukrainization-in-pro-russian-propaganda-in-romania-poland-serbia-and-hungary/.
[112] *Id.*
[113] *Id.*
[114] *Id.*
[115] *Hatelex: A Lexicon of Hate Speech Terms in Kenya*, NAT'L COHESION & INTEGRATION COMM'N KENYA (Apr. 2022) [hereinafter *Hatelex in Kenya*].
[116] Interview with Representative, Nat'l Cohesion & Integration Comm'n Kenya (Mar. 23, 2023).
[117] *Hatelex in Kenya*, *supra* note 115.
[118] *Id.*
[119] *Id.*
[120] Daniel Halpern, Sebastián Valenzuela & James E. Katz, *We Face, I Tweet: How Different Social Media Influence Political Participation through Collective and Internal Efficacy*, 22 J. COMPUT.-MEDIATED COMMC'N 320 (2017); Leticia Bode et al., *A New Space for Political Behavior: Political Social Networking and its Democratic Consequences**, 19 J. COMPUT.-MEDIATED COMMC'N 414 (2014).
[121] Interview with Andre Oboler, *supra* note 53.
[122] Interview with Representative, *supra* note 54.
[123] Interview with Anonymous, *supra* note 62.
[124] Zecharias Zelalem & Peter Guest, *Why Facebook Keeps Failing in Ethiopia*, REST OF WORLD (Nov. 13, 2021), https://restofworld.org/2021/why-facebook-keeps-failing-in-ethiopia/.
[125] *Id.*
[126] *Id.*
[127] sir arwa (@sirarwa), TWITTER (Oct. 30, 2021, 9:27 PM), https://twitter.com/sirarwa/status/1454620972330504197.
[128] Getachew Reda (@reda_getachew), TWITTER (Nov. 12, 2021, 8:00 AM),

http://web.archive.org/web/20211112161602/https:/twitter.com/reda_getachew/status/1459189337778708484.
[129] Joshua Uyheng & Kathleen M. Carley, *Characterizing Network Dynamics of Online Hate Communities Around the COVID-19 Pandemic*, 6 APPLIED NETWORK SCIS. 1 (2021).
[130] Naganna Chetty & Sreejith Alathur, *Hate Speech Review in the Context of Online Social Networks*, 40 AGGRESSIVE VIOLENT BEHAV. 108 (2018).
[131] Interview with Representative, *supra* note 53.
[132] Ashenafi Kassa (@Ashuyanis), TWITTER (Dec. 6, 2021, 3:45 AM), https://twitter.com/Ashuyanis/status/1467777165353705474.
[133] Interview with Representative, *supra* note 54.
[134] Interview with Andre Oboler, *supra* note 53.
[135] sir arwa (@sirarwa), TWITTER (Oct. 30, 2021, 9:27 PM), https://twitter.com/sirarwa/status/1454620972330504197.
[136] Axumawi Agametai (@Agameta11051624), TWITTER (Feb. 22, 2023, 8:08 AM), https://twitter.com/Agameta11051624/status/1628381275823251456.
[137] ICCPR, *supra* note 31, art. 20(2).
[138] The Community Standards list the following protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.
[139] *Hate Speech*, META, https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/ (last visited Apr. 24, 2023).
[140] ICERD, *supra* note 31.
[141] *See, e.g.*, Hate Speech and Hate Crime, Am. Library Ass'n, https://www.ala.org/advocacy/intfreedom/hate (last visited Apr. 24, 2023).
[142] *See* #DEFYHATENOW, SOCIAL MEDIA HATE SPEECH MITIGATION FIELD GUIDE 11 (2d ed. 2021); INT'L COMM. RED CROSS, HARMFUL INFORMATION: MISINFORMATION, DISINFORMATION AND HATE SPEECH IN ARMED CONFLICT AND OTHER SITUATIONS OF VIOLENCE (2021).
[143] Eric Min & Jacob Shapiro, *Determining Risk and Resilience to Violent Conflict*, WORLD BANK BLOGS (Apr. 12, 2018), available at https://blogs.worldbank.org/dev4peace/determining-risk-and-resilience-violent-conflict.
[144] *Global Conflict Risk Index*, DISASTER RISK MGMT. KNOWLEDGE CTR., https://drmkc.jrc.ec.europa.eu/initiatives-services/global-conflict-risk-index#documents/1059/list (last visited Apr. 24, 2023).
[145] *Volatility & Risk Predictability Index*, ARMED CONFLICT LOCATION & EVENT DATA PROJECT, https://acleddata.com/early-warning-research-hub/volatility-and-risk-predictability-index/ (last visited Apr. 24, 2023).
[146] *Safety Perceptions Index 2023: Understanding the Impact of Risk Around the World*, RELIEF WEB, https://reliefweb.int/report/world/safety-perceptions-index-2023-understanding-impact-risk-around-world (last visited Apr. 24, 2023).
[147] Elizabeth Mackintosh, *Facebook Knew it Was Being Used to Incite Violence in Ethiopia*, CNN BUS. (Oct. 25, 2021), https://www.cnn.com/2021/10/25/business/ethiopia-violence-facebook-papers-cmd-intl/index.html.
[148] Fabio Andres Diaz Pabon & Muna Shifa, *The Interaction of Mass Media and Social in Fuelling Ethnic Violence in Ethiopia*, 2021 CONFLICT AND RESILIENCE MON. 4 (2022).

[149] Alex Hern, *Meta Faces $1.6bn Lawsuit over Facebook Posts Inciting Violence in Tigray War*, GUARDIAN (Dec. 14, 2022), https://www.theguardian.com/technology/2022/dec/14/meta-faces-lawsuit-over-facebook-posts-inciting-violence-in-tigray-war.
[150] John Campbell, *What Went Wrong With Kenya's Elections?*, COUNCIL ON FOREIGN RELS. (Nov. 3, 2017, 1:58 PM), https://www.cfr.org/expert-brief/what-went-wrong-kenyas-elections.
[151] *UN Human Rights Team Issues Report on Post-Election Violence in Kenya*, OFF. OF THE HIGH COMM'R FOR HUM. RTS. (Mar. 18, 2008), https://www.ohchr.org/en/press-releases/2009/10/un-human-rights-team-issues-report-post-election-violence-kenya.
[152] Mwagi S. Kimenyi, *Kenya's Elections: Implications of Ethnic Rivalries and International Intervention*, BROOKINGS (Feb. 12, 2023), https://www.brookings.edu/opinions/kenyas-elections-implications-of-ethnic-rivalries-and-international-intervention/.
[153] David Smith, *Kenya Sees Huge Election Turnout But Violence Mostly Limited to Separatists*, GUARDIAN (Mar. 4, 2013), https://www.theguardian.com/world/2013/mar/04/kenya-vote-kenyatta-odinga-violence.
[154] Briana Duggan, Faith Karimi & Chandrika Narayan, *24 Killed in Post-Election Violence in Kenya, Rights Group Says*, CNN (Aug. 13, 2017, 3:57 AM), https://www.cnn.com/2017/08/12/africa/kenya-elections-protests/index.html.
[155] Eyder Peralta, *In Kenya, Much of the Election Chaos and Violence Stems from Tribal Divisions*, NPR (Oct. 24, 2017, 4:50 PM), https://www.npr.org/2017/10/24/559889613/in-kenya-much-of-the-election-chaos-and-violence-stems-from-tribal-divisions.
[156] John Campbell, *What Went Wrong With Kenya's Elections?*, COUNCIL ON FOREIGN RELS. (Nov. 3, 2017, 1:58 PM), https://www.cfr.org/expert-brief/what-went-wrong-kenyas-elections.
[157] Interview with Representative, Amnesty International Kenya (Mar. 24, 2023).
[158] Victoire Rio, *Myanmar: The Role of Social Media in Fomenting Violence*, *in* SOCIAL MEDIA IMPACTS ON CONFLICT AND DEMOCRACY (2021).
[159] Paul Mozur, *A Genocide Incited on Facebook, With Posts From Myanmar's Military*, N.Y. TIMES (Oct. 15, 2018), https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.
[160] *Id.*
[161] Sheera Frenkel, *This Is What Happens When Millions of People Suddenly Get the Internet*, BUZZFEED NEWS (Nov. 20, 2016), https://www.buzzfeednews.com/article/sheerafrenkel/fake-news-spreads-trump-around-the-world.
[162] AMNESTY INT'L, THE SOCIAL ATROCITY: META AND THE RIGHT TO REMEDY FOR THE ROHINGYA 6 (2022), available at https://www.amnesty.org/en/documents/ASA16/5933/2022/en/.
[163] Interview with Anonymous, *supra* note 62.
[164] Interview with Ngala Desmond, *supra* note 63.
[165] *Brazil: Disinformation and Online Attacks Against Women Journalists Pose Serious Challenges to the Exercise of Press Freedom in the Country*, REPS. WITHOUT BORDERS (Apr. 26, 2022), https://rsf.org/en/news/brazil-disinformation-and-online-attacks-against-women-journalists-pose-serious-challenges-exercise.
[166] Patrícia Campos Mello, *Businessmen Fund WhatsApp Campaign Against PT*, FOLHA DE S. PAULO (Oct. 18, 2018, 2:26 AM), https://www1.folha.uol.com.br/internacional/en/brazil/2018/10/businessmen-fund-whatsapp-campaign-against-pt.shtml.
[167] Augusta Saraiva, *Tackling Disinformation in Brazil*, FOREIGN POL'Y (Sept. 19, 2020, 8:00 AM), https://foreignpolicy.com/2020/09/19/tackling-disinformation-in-brazil-interview-patricia-campos-mello/.
[168] *Id.*
[169] Jem Bartholomew, *'Democracy on the Line': Brazil's Election and the Bolsonaro Disinformation Ecosystem*, COLUMBIA JOURNALISM REV. (Oct. 11, 2022), https://www.cjr.org/tow_center/democracy-on-the-line-brazils-election-and-the-bolsonaro-disinformation-ecosystem.php.
[170] *Brazilian Journalist Patrícia Campos Mello Faces Online Harassment Campaign*, COMM. TO PROTECT JOURNALISTS (Feb. 12, 2020, 5:41 PM), https://cpj.org/2020/02/brazilian-journalist-patricia-campos-mello-faces-o/.
[171] *Id.*
[172] *RSF Denounces Inflammatory Comments About Journalist by Brazilian Deputy*, REPS. WITHOUT BORDERS (Feb. 13, 2020), https://rsf.org/en/rsf-denounces-inflammatory-comments-about-journalist-brazilian-deputy.

[173] @BolsonaroSP, TWITTER (Feb. 11, 2020, 6:59 PM), https://twitter.com/BolsonaroSP/status/1227381635341070337.
[174] *RSF Denounces Inflammatory Comments About Journalist by Brazilian Deputy*, REPS. WITHOUT BORDERS (Feb. 13, 2020), https://rsf.org/en/rsf-denounces-inflammatory-comments-about-journalist-brazilian-deputy.
[175] *Id.*
[176] @carlosjordy, TWITTER (Feb. 11, 2020, 10:31 PM), https://twitter.com/carlosjordy/status/1227435051455000576
[177] @filipebarrost, TWITTER (Feb. 11, 2020, 1:20 PM), https://twitter.com/filipebarrost/status/1227296386460286978.
[178] *Brazilian Journalist Patrícia Campos Mello Faces Online Harassment Campaign*, COMM. TO PROTECT JOURNALISTS (Feb. 12, 2020, 5:41 PM), https://cpj.org/2020/02/brazilian-journalist-patricia-campos-mello-faces-o/.
[179] @BolsonaroSP, TWITTER (Feb. 11, 2020, 6:59 PM), https://twitter.com/BolsonaroSP/status/1227381635341070337.
[180] Renata Neder, *Brazilian Journalist Patrícia Campos Mello Sued President Bolsonaro's Son for Moral Damages—and Won*, COMM. TO PROTECT JOURNALISTS (Mar. 2, 2021, 11:33 AM), https://cpj.org/2021/03/brazilian-journalist-patricia-campos-mello-sued-president-bolsonaros-son-for-moral-damages-and-won/.
[181] *Id.*
[182] *Id.*
[183] Augusta Saraiva, *Tackling Disinformation in Brazil*, FOREIGN POL'Y (Sept. 19, 2020, 8:00 AM), https://foreignpolicy.com/2020/09/19/tackling-disinformation-in-brazil-interview-patricia-campos-mello/.
[184] Martina Bertam, *Brazil: Harassed Journalist Calls for More Action Against Hate Speech Online*, DW GLOB. MEDIA F. (July 29, 2021), https://corporate.dw.com/en/brazil-harrassed-journalist-calls-for-more-action-against-hate-speech-online/a-58660062.
[185] Manuela d'Ávila, *Violence in Social Media Threatens Women Active in Brazilian Politics*, HEINRICH BÖLL FOUND. (Sept. 27, 2022), https://us.boell.org/en/2022/09/27/violence-social-media-threatens-women-active-brazilian-politics.
[186] *Xenophobia, Religious Intolerance and Misogyny Were the Crimes Reported to Safernet that Grew the Most in the Elections*, SAFERNET, https://new.safernet.org.br/content/xenofobia-intolerancia-religiosa-e-misoginia-foram-os-crimes-denunciados-a-safernet-que-mais-cresceram-nas-eleicoes.