Ethical Scaling for Content Moderation:

Extreme Speech and the (In)Significance of Artificial Intelligence

JUNE 2022

Author

Sahana Udupa, Fall 2021 Joan Shorenstein Fellow

Antonis Maronikolakis

Hinrich Schütze

Axel Wisiorek

HARVARD Kennedy School
SHORENSTEIN CENTER
on Media, Politics and Public Policy

Abstract

In this paper, we present new empirical evidence to demonstrate the near impossibility for existing machine learning content moderation methods to keep pace with, let alone stay ahead of, hateful language online. We diagnose the technical shortcomings of the content moderation and natural language processing approach as emerging from a broader epistemological trapping wrapped in the liberal-modern idea of the 'human,' and provide the details of the ambiguities and complexities of annotating text as derogatory or dangerous, in a way to demonstrate the need for persistently involving communities in the process. This decolonial perspective of content moderation and the empirical details of the technical difficulties of annotating online hateful content emphasize the need for what we describe as "ethical scaling". We propose ethical scaling as a transparent, inclusive, reflexive and replicable process of iteration for content moderation that should evolve in conjunction with global parity in resource allocation for moderation and addressing structural issues of algorithmic amplification of divisive content. We highlight the gains and challenges of ethical scaling for AI-assisted content moderation by outlining distinct learnings from our ongoing collaborative project, AI4Dignity.

1 Corresponding author: sahana.udupa@lmu.de and sahanaudupa@hks.harvard.edu, https://orcid.org/0000-0003-3647-9570

Table of Contents

Keywords: *AI and content moderation, social media, extreme speech, ethical scaling, decoloniality*

"In the southern Indian state of Kerala, the right-wing group is a numerical minority. They are frequently attacked online by members of the communist political party. Should I then categorize this speech as exclusionary extreme speech since it is against a minority group?" asked a fact-checker from India, as we gathered at a virtual team meeting to discuss proper labels to categorize different forms of contentious speech that circulate online. For AI4Dignity, a social intervention project that blends machine learning and ethnography to articulate responsible processes for online content moderation, it was still an early stage of labeling. Fact-checkers from Brazil, Germany, India and Kenya, who participated as community intermediaries in the project, were at that time busy slotting problematic passages they had gathered from social media into three different categories of extreme speech for machine learning. We had identified these types as *derogatory extreme speech* (demeaning but does not warrant removal of content), *exclusionary extreme speech* (explicit and implicit exclusion of target groups that requires stricter moderation actions such as demoting) and *dangerous speech* (with imminent danger of physical violence warranting immediate removal of content). We had also drawn a list of target groups, which in its final version included ethnic minorities, immigrants, religious minorities, sexual minorities, women, racialized groups, historically oppressed caste groups, indigenous groups, large ethnic groups and any other. Under derogatory extreme speech, we also had groups beyond protected characteristics, such as politicians, legacy media, the state and civil society advocates for inclusive societies, as targets.

The Indian fact-checker's question about right-wingers as a numerical minority in an Indian state was quite easy to answer. "You don't seek to protect right-wing communities simply because they are a minority in a specific region. You need to be aware of the dehumanizing language they propagate, and realize that their speech deserves no protection," we suggested instantly. But questions from fact-checkers were flowing continuously, calling attention to diverse angles of the annotation problem.

"You have not listed politicians under protected groups [of target groups]," observed a fact-checker from Kenya. "Anything that targets a politician also targets their followers and the ethnic group they represent," he noted, drawing reference to a social media post with mixed registers of English and Swahili: "Sugoi thief will never be president. Sisi wakikuyu tumekataa kabisa, Hatuezii ongozwa na mwizi [We the Kikuyu have refused totally, we cannot be led by a thief]." In this passage, the politician did not just represent a constituency in the formal structures of electoral democracy but served as a synecdoche for an entire target community. Verbal attacks in this case, they argued, would go beyond targeting an individual politician.

In contrast, the German fact-checking team was more cautious about their perceptions of danger. "We were careful with the selection of dangerous speech," informed a fact-checker from Germany, "How can we designate something as dangerous speech when we are not too sure about the sender, let alone the influence they have over the audience?" In the context that we had not requested fact-checkers to gather information about the source of extreme speech instances, and for data protection reasons instructed them strictly to avoid adding any posters' per-

sonal identifiers, the problem of inadequate information in determining the danger levels of speech loomed over the annotation exercise.

The complex semantics of extreme speech added to the problem. "They don't ever use a sentence like, 'This kind of people should die.' Never." explained a fact-checker from Brazil, referring to a hoax social media post that claimed that United States' President Joe Biden had appointed an LGBTQI+ person to head the education department. Homophobic groups do not use direct insult, he explained:

> It's always something like, 'This is the kind of person who will take care of our children [as the education minister]'. Although it is in the written form, I can imagine the intonation of how they are saying this. But I cannot fact-check it, it's not fact-checkable. Because, you know, I don't have any database to compare this kind of sentence, it is just implicit and it's typical hate speech that we see in Brazil. Do you understand the difficulty?

In this paper, we present new empirical evidence to demonstrate the near impossibility for existing machine learning content moderation methods to keep pace with, let alone stay ahead of, hateful language online.

As questions poured in and extreme speech passages piled up during the course of the project, and as we listened to fact-checkers' difficult navigations around labeling problematic online content, we were struck by the complexity of the task that was staring at us. From missing parts of identity markers for online posters to the subtlety of language to the foundational premises for what constitutes the unit of analysis or the normative framework for extremeness in online speech, the challenge of labeling contentious content appeared insurmountable.

In this paper, we present new empirical evidence to demonstrate the near impossibility for existing machine learning content moderation methods to keep pace with, let alone stay ahead of, hateful language online. We focus on the severe limitations in the content moderation practices of global social media companies such as Facebook and Google as the context to emphasize the urgent need to involve community intermediaries with explicit social justice agendas for annotating extreme speech online and incorporating their participation in a fair manner in the lifecycle of artificial intelligence (AI) assisted model building. To advance this point, we present a set of findings from the AI4Dignity project that involved facilitated dialogue between independent fact-checkers, ethnographers and AI developers to gather and annotate extreme speech data.

We employ two methods to highlight the limitations of AI-assisted content moderation practices among commercial social media platforms. First, we compare the AI4Dignity extreme speech datasets with Perspective API's toxicity scores developed by Google. Second, using manual advanced search methods, we test a small sample of the annotated dataset to examine whether they continue to appear on Twitter. We layer these findings with the ethnographic observations of our interactions with fact-checkers during different stages of the project, to show how even facilitated exercises for data annotation with the close engagement of fact-checkers and ethnographers with regional expertise can become not only resource intensive and demanding but also uncertain in terms of capturing the granularity of extreme speech, although the binary classification between extreme and non-extreme as well as types of extreme speech that should be removed and those that warrant other kinds of moderation actions, such as downranking or counter speech, is agreed upon quite easily.

We argue that such interactions, however demanding, are the precise (and the only) means to develop an iterative process of data gathering, labeling and model

building that can stay sensitive to historically constituted and evolving axes of exclusion, and locate shifting, coded and indirect expressions of hate that ride on local cultural idioms and linguistic repertoire as much as global catchphrases in English. We highlight this exercise as a reflexive and ethical process through which communities with explicit social justice agendas and those most affected by hate expressions take a leading role in the process of annotation in ways that the gains of transparency and iteration in the 'ordering of data' and content moderation decisions are channeled back towards protecting communities. Such knowledge through iterative processes involves an appreciation not only for social media posts but also broader contextual factors including the vulnerability of target groups and the power differentials between the speaker and target.

Ethical scaling envisions a transparent, inclusive, reflexive and replicable process of iteration for content moderation that should evolve in conjunction with addressing structural issues of algorithmic amplification of divisive content.

This policy approach and the empirical evidence upon which it is built calls for some conceptual rethinking. The exercise of community intermediation in AI cultures highlights the importance of pushing back against the liberal framing of "the human versus the machine" conundrum. We therefore begin this essay with a critique of the liberal conception of the "human" by asking how the moral panics around human autonomy versus machine intelligence in AI-related discussions as well as its inverse—the ambitions to prepare machines as humans—hinge on the liberal-modern understanding of "rationality as the essence of personhood"[2] that obscures the troubled history of the human/subhuman/nonhuman distinction that colonial modernity instituted. We argue that the liberal-modern understanding of rationality that drives the ambitions to transfer rational personhood to the machine and the anxiety around such ambitions are conceptually unprepared to grasp the responsibility of community participation in the design and imagination of the machine. Such a view, for the problem of extreme speech discussed here, elides the responsibility of involving communities in content moderation. Critiquing the rationality-human-machine nexus and the colonial logics of the human/subhuman/nonhuman distinction that underwrite global disparities in content moderation as well as forms of extreme speech aimed at immigrants, minoritized people, religious and ethnic 'others,' people of color and women,[3] we propose the principle of "ethical scaling." Ethical scaling envisions a transparent, inclusive, reflexive and replicable process of iteration for content moderation that should evolve in conjunction with addressing structural issues of algorithmic amplification of divisive content. Ethical scaling builds on what studies have observed as "speech acts" that can have broad-ranging impacts not only in terms of their co-occurrence in escalations of physical violence (although causality is vastly disputed) but also in terms of preparing the discursive ground for exclusion, discrimination and hostility.[4] Far from an uncritical embrace of free speech, we therefore hold that responsible content moderation is an indispensable aspect of platform regulation. In the next sections of the essay, we substantiate the gains and challenges of "ethical scaling" with empirical findings.

2 Mhlambi, Sabelo. "From Rationality to Relationality." Carr Center for Human Rights Policy Harvard Kennedy School, Carr Center Discussion Paper, No. 009: 31. 2020, p 1.

3 These forms of extreme speech are analytically distinct but in reality come mixed with, amplify or differentially shape the outcomes of other kinds of extreme speech such as election lies and medical misinformation.

We conclude by arguing for a framework that treats the distribution and content sides of online speech holistically, highlighting how AI is insignificant in tackling the ecosystem of what is defined as "deep extreme speech."

AI in content moderation and the colonial bearings of human/machine

In countries where democratic safeguards are crumbling, the extractive attention economy of digital communication has accelerated a dangerous interweaving of corporate greed and state repression, while regulatory pressure has also been mounting globally to bring greater public accountability and transparency in tech operations.

As giant social media companies face the heat of the societal consequences of polarized content they facilitate on their platforms while also remaining relentless in their pursuit of monetizable data, the problem of moderating online content has reached monumental proportions. There is growing recognition that online content moderation is not merely a matter of technical capacity or corporate will but also a serious issue for governance, since regressive regimes around the world have sought to weaponize online discourse for partisan gains, to undercut domestic dissent or power up geopolitical contestations against "rival" nation states through targeted disinformation campaigns. In countries where democratic safeguards are crumbling, the extractive attention economy of digital communication has accelerated a dangerous interweaving of corporate greed and state repression, while regulatory pressure has also been mounting globally to bring greater public accountability and transparency in tech operations.

Partly to preempt regulatory action and partly in response to public criticism, social media companies are making greater pledges to contain harmful content on their platforms. In these efforts, AI has emerged as a shared imaginary of technological solutionism. In corporate content moderation, AI comes with the imagined capacity to address online hateful language across diverse contexts and political specificities. Imprecise in terms of the actual technologies it represents and opaque in terms of the technical steps that lead up to its constitution, AI has nonetheless gripped the imagination of corporate minds as a technological potentiality that can help them to confront a deluge of soul wrecking revelations of the harms their platforms have helped amplify.

AI figures in corporate practices with different degrees of emphasis across distinct content moderation systems that platform companies have raised, based on their technical architecture, business models and the size of operation. Robyn Caplan distinguishes them as the "artisanal" approach where "case-by-case governance is normally performed by between 5 and 200 workers" (platforms such as Vimeo, Medium and Discord); "community-reliant" approaches "which typically combine formal policy made at the company level with volunteer moderators" (platforms such as Wikipedia and Reddit); and "industrial-sized operations where tens of thousands of workers are employed to enforce rules made by a separate

4 See Butler, Judith. 1997. *Excitable Speech: A Politics of the Performative.* New York: Routledge. Dangers of regulatory overreach and clamping down freedom of expression require sound policies and procedural guidelines but the anxiety around overreach cannot become an excuse for unfettered defence of freedom of expression or to view content moderation as something that has to wait for imminent violence. For a review of this scholarship, see Udupa, Sahana, Iginio Gagliardone, Alexandra Deem and Laura Csuka. 2020. "Field of Disinformation, Democratic Processes and Conflict Prevention". Social Science Research Council, https://www.ssrc.org/publications/view/the-field-of-disinformation-democratic-processes-and-conflict-prevention-a-scan-of-the-literature/

policy team" (characterized by large platforms such as Google and Facebook).[5] Caplan observes that "industrial models prioritize consistency and artisanal models prioritize context."[6] Automated solutions are congruent with the objective of consistency in decisions and outcomes, although such consistency also depends on how quickly rules can be formalized.[7]

In "industrial-size" moderation activities, what is glossed as AI largely refers to a combination of a relatively simple method of scanning existing databases of labeled expressions against new instances of online expression to evaluate content and detect problems—a method commonly used by social media companies[8]—and a far more complex project of developing machine learning models with the 'intelligence' to label texts they are exposed to for the first time based on the steps they have accrued in picking up statistical signals from the training datasets. AI—in the two versions of relatively simple comparison and complex 'intelligence'—is routinely touted as a technology for the automated content moderation actions of social media companies, including flagging, reviewing, tagging (with warnings), removing, quarantining and curating (recommending and ranking) textual and multimedia content. AI deployment is expected to address the problem of volume, reduce costs for companies and decrease human discretion and emotional labor in the removal of objectionable content.

The need for cultural contextualization in detection systems is a widely acknowledged limitation since there is no catch-all algorithm that can work for different contexts.

However, as companies themselves admit, there are vast challenges in AI-assisted moderation of hateful content online. One of the key challenges is the quality, scope and inclusivity of training datasets. AI needs "millions of examples to learn from. These should include not only precise examples of what an algorithm should detect and 'hard negatives,' but also 'near positives'—something that is close but should not count."[9] The need for cultural contextualization in detection systems is a widely acknowledged limitation since there is no catch-all algorithm that can work for different contexts. Lack of cultural contextualization has resulted in false positives and over-application. Hate groups have managed to escape keyword-based machine detection through clever combinations of words, misspellings,[10] satire, changing syntax and coded language.[11] The dynamic nature of online hateful speech—where hateful expressions keep changing—adds to the complexity. As a fact-checker participating in the AI4Dignity project expressed, they are swimming against "clever ways [that abusers use] to circumvent the hate speech module."

A more foundational problem cuts through the above two challenges. This concerns the definitional problem of hate speech. There is no consensus both legally and culturally around what comprises hate speech, although the United Nations has set the normative parameters while acknowledging that "the characterization

5 Caplan, Robyn. "Content or Context Moderation?" *Data & Society.* Data & Society Research Institute. November 14, 2018, p 16. https://datasociety.net/library/content-or-context-moderation/.

6 Caplan, 2018, p 6.

7 Caplan, 2018.

8 Gillespie, Tarleton. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7 (2): 2053951720943234. 2020. https://doi.org/10.1177/2053951720943234.

9 Murphy, Hannah, and Madhumita Murgia. "Can Facebook Really Rely on Artificial Intelligence to Spot Abuse?" *FT.Com*, November, 2019. https://www.proquest.com/docview/2313105901/citation/D4DBCB03EAC348C7PQ/1.

for what is 'hateful' is controversial and disputed."[12] This increases the difficulties of deploying AI-assisted systems for content moderation in diverse national, linguistic and cultural contexts. A fact-checker from Kenya pointed out that even within a national context, there are not only regional and subregional distinctions about what is understood as hate speech but also an urban/rural divide. "In the urban centers, some types of information are seen as 'outlaw,' so it is not culturally accepted," he noted, "but if you go to other places, it's seen as something in the norm." As regulators debate actions against online extreme speech not only in North America, where big tech is headquartered, but also in different regions of the world where they operate, platform companies are reminded that their content moderation and AI use principles that are largely shaped by the "economic and normative...[motivations]...to reflect the democratic culture and free speech expectations of...[users]"[13] have to step beyond North American free speech values and negotiate the staggeringly diverse regulatory, cultural and political climates that surround online speech.[14]

There is no consensus both legally and culturally around what comprises hate speech, although the United Nations has set the normative parameters while acknowledging that "the characterization for what is 'hateful' is controversial and disputed."

Several initiatives have tried to address these limitations by incorporating users' experiences and opinions.[15] Google's Perspective API and Twitter's Birdwatch have experimented with crowdsourcing models to evaluate content. Launched in 2021 as a pilot, Birdwatch allows users to label information in tweets as misleading and provide additional context. Google's Perspective API offers "toxicity scores" to passages based on user inputs feeding the machine learning models. Such efforts have sought to leverage 'crowd intelligence' but the resulting machine learning models, while offering some promising results in terms of detecting evolving forms of extreme content, are prone to false positives as well as racial bias.[16] Studies have also found that crowdsourced models have the problem of differential emphasis. Whereas racist and homophobic tweets are more likely to be identified as hate speech in the North American context, gender-related comments are often brushed aside as merely offensive speech.[17] More critically, crowdsourced models have channelized corporate accountability and the onus of detection onto an undefined entity called 'crowd,' seeking to co-opt the Internet's promised openness to evade regulatory and social consequences of gross inadequacies in corporate efforts and investments in moderating problematic content.

Such challenges could be framed either as platform governance issues or the problem of technology struggling to catch up to the mutating worlds of words, thereby igniting the hope that they would be addressed as resources for content moderation

10 Gröndahl, Tommi, Luca Pajola, Mika Juuti, Mauro Contin, and N. Asokan. "All You Need Is 'Love': Evading Hate Speech Detection." ArXiv:1808.09115v3 [Cs.CL]. 2018.

11 See Burnap, Pete & Matthew L. Williams (2015). "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making". *Policy & internet,* 7(2), 223-242; Fortuna, Paula, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference,* pages 6786–6794, Marseille, France. European Language Resources Association; Ganesh, Bharath. "The Ungovernability of Digital Hate Culture." *Journal of International Affairs* 71 (2), 2018, pp 30–49; Warner, W., and J. Hirschberg. "Detecting Hate Speech on the World Wide Web." In P*roceedings of the Second Workshop on Language in Social Media,* 2012, pp 19–26. Association for Computational Linguistics. https://www.aclweb.org/anthology/W12-2103;

expand and political pressure increases. However, some fundamental ethical and political issues that undergird the problem prompt a more incisive critical insight. Across attempts to bring more "humans" for annotation, there is not only a tendency to frame the issue as a technical problem or platform (ir)responsibility but also a more taken-for-granted assumption that bringing "humans" into the annotation process will counterbalance the dangers and inadequacies of machine detection. This approach is embedded within a broader moral panic around automation and demands to assert and safeguard "human autonomy" against the onslaught of the digital capitalist data "machine." In such renderings, the concept of "the human" represents the locus of moral autonomy[18] that needs protection from the "machine."

More critically, crowdsourced models have channelized corporate accountability and the onus of detection onto an undefined entity called 'crowd,' seeking to co-opt the Internet's promised openness to evade regulatory and social consequences of gross inadequacies in corporate efforts and investments in moderating problematic content.

Conversely, the human-machine correspondence aspired to in the development of algorithmic machines takes, as Sabelo Mhlambi has explained, "the traditional view of rationality as the essence of personhood, designating how humans and now machines, should model and approach the world."[19] As he points out, this aspired correspondence obscures the historical fact that the traditional view of rationality as the essence of personhood "has always been marked by contradictions, exclusions and inequality."[20] In their decolonial reading, William Mpofu and Melissa Steyn further complicate "the human" as a category, highlighting the risks of its uncritical application:

> The principal trouble with the grand construction of the human of Euro-modernity…is that it was founded on unhappy circumstances and for tragic purposes. Man, as a performative idea, created inequalities and hierarchies usable for exclusion and oppression of the other…The attribute human…is not self-evident or assured. It can be wielded; given and taken away.[21]

"The human" as an attribute that is *wielded* rather than self-evident or assured brings to sharp relief the conceits and deceits of liberal-modern thought. The liberal weight behind the concept of the human elides its troubled lineage in European colonial modernity that racially classified human, subhuman and nonhuman,[22] institutionalizing this distinction within the structures of the modern nation-state

12 See https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

13 Klonick, Kate. "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review* 131, 2017, p 1603.

14 Sablosky, Jeffrey. "Dangerous Organizations: Facebook's Content Moderation Decisions and Ethnic Visibility in Myanmar." *Media, Culture & Society* 43 (6), 2021, pp 1017–42. https://doi.org/10.1177/0163443720987751.

15 See Online Hate Index developed by Berkeley Institute for Data Science https://www.adl.org

16 Sap et al., 2019.

17 Davidson et al., 2017.

18 Becker, Lawrence C., and Charlotte B. Becker. *A History of Western Ethics*. v. 1540. New York: Garland Publication. 1992.

19 Mhlambi, 2020.

20 Mhlambi, 2020, p 1.

21 Steyn, Melissa, and William Mpofu, eds. *Decolonising the Human: Reflections from Africa on Difference and Oppression.* Wits University Press. 2021, p 1. https://doi.org/10.18772/22021036512.

(that marked the boundaries of the inside/outside and minority/majority populations) and the market (that anchored the vast diversity of human activities to the logic of accumulation). As Sahana Udupa has argued, the nation-state, market and racial relations of colonial power constitute a composite structure of oppression, and the distinctive patterns of exclusion embedded in these relations have evolved and are reproduced in close conjunction.[23]

For online content moderation and AI, attention to colonial history raises four questions. A critical view of the category of the "human" is a reminder of the foundational premise of the human/subhuman/nonhuman distinction of coloniality that drives, validates and upholds a significant volume of hateful langauge online based on racialized and gendered categories and the logics of who is inside and who is outside of the nation-state and who is a minority and who is in the majority. Importantly, such oppressive structures operate not only on a global scale by defining the vast power differentials among national, ethnic or racialized groups but also within the nation-state structures where dominant groups reproduce coloniality through similar axes of difference as well as systems of hierarchy that "co-mingle with if not are invented" by the colonial encounter.[24] Importantly, extreme speech content is also driven by the market logics of coloniality, and as Jonathan Beller states, "Computational capital has not dismantled racial capitalism's vectors of oppression, operational along the exacerbated fracture lines of social difference that include race, gender, sexuality, religion, nation, and class; it has built itself and its machines out of those capitalized and technologized social differentials."[25] For instance, alongside active monetization of problematic content that deepens these divisions, biased training data in ML models has led to greater probability that African American English will be singled out as hateful, with "disproportionate negative impact on African-American social media users."[26] There is mount-

22 Wynter, Sylvia. "Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation—An Argument." CR: *The New Centennial Review* 3 (3), 2003, pp. 257–337.

23 Udupa, Sahana. "Decoloniality and Extreme Speech." In Media Anthropology Network E-Seminar. European Association of Social Anthropologists. 2020. https://www.easaonline.org/downloads/networks/media/65p.pdf.

24 Thirangama, Sharika, Tobias Kelly, and Carlos Forment. 2018. "Introduction: Whose Civility?" *Anthropological Theory* 18 (2–3), 2018, pp 153–74. https://doi.org/10.1177/1463499618780870.

25 Original emphasis. Beller, Jonathan. "The Fourth Determination". *e-flux,* 2017, Retrieved from https://www.e- flux.com/journal/85/156818/the-fourth-determination/

26 Davidson, Thomas, Debasmita Bhattacharya and Ingmar Weber. 2019. "Racial Bias in Hate Speech and Abusive Langauge Detection Datasets". *Proceedings of the Third Abusive Langauge Workshop,* pp. 25-35. Florence: Association for Computational Linguistics. See also Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. "The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* pp. 1668–1678, Florence: Association for Computational Linguistics. For problems in the category definitions, see also Fortuna, Paula, Juan Soler and Leo Wanner. 2020. "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? Empirical Analysis of Hate Speech Datasets". Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6786–6794. Marseille: European Language Resources Association.

ing evidence for how classification algorithms, training data, and the application of machine learning models are biased because of the limitations posed by the homogenous workforce of technology companies that employ disproportionately fewer women, minorities and people of color.[27] This is also reflected in the technical sciences. Natural Language Processing (NLP) and other computational methods have not only highlighted, but are also themselves weighed down by limited and biased data and labeling.

There is mounting evidence for how classification algorithms, training data, and the application of machine learning models are biased because of the limitations posed by the homogenous workforce of technology companies that employ disproportionately fewer women, minorities and people of color.

Epistemologies of coloniality also limits the imaginations of technological remedies against hateful language. Such thinking encourages imaginations of technology that spin within the frame of the "rational human"—the product of colonial modernity—as either the basis for the machine to model upon or the moral force to resist automation. Put differently, both the problem (extreme speech) and the proposed solution (automation) are linked to Euro-modern thinking.

At the same time, proposed AI-based solutions to hateful language that take the human as an uncontested category fail to account for how the dehumanizing distinction between the human/subhuman/nonhuman categories instituted by coloniality shape complex meanings, norms and affective efficacies around content that cannot be fully discerned by the machines. As Mhlambi sharply argues, "this is not a problem of not having enough data, it is simply that data does not interpret itself."[28] Computational processes will never be able to fully model meaning and meaning-making.

Even more, the dehumanizing distinction of coloniality also tacitly rationalizes the uneven allocation of corporate resources for content moderation across different geographies and language communities, and the elision of the responsibility of involving affected communities as an indelible principle of annotation and moderation. Based on the most recent whistleblower accounts that came to be described as the "Facebook Papers" in Western media, *The New York Times* reported that, "Eighty-seven percent of the company's global budget for time spent on classifying misinformation is earmarked for the United States, while only 13 percent is set aside for the rest of the world—even though North American users make up only 10 percent of the social network's daily active users."[29] In the news article, the company spokesperson was quoted claiming that the "figures were incomplete and don't include the company's third party fact-checking partners, most of whom are outside the United States," but the very lack of transparency around the allocation of resources and the outsourced arrangements around "third party partners" signal the severely skewed structures of content moderation that global social media corporations have instituted. Such disparities attest to what Denis Ferreira da Silva observes as the spatiality of racial formation characterized by a constitutive overlap between symbolic spatiality (racialized geographies of whiteness and privilege) and the material terrain of the world.[30]

28 Mhlambi 2020, p 5.

29 Frenkel, Sheera and Alba, Davey. "In India, Facebook Struggles to Combat Misinformation and Hate Speech." *The New York Times,* October 23, 2021. https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html.

30 Ferreira da Silva, Denis. *Toward a Global Idea of Race.* Minneapolis: University of Minnesota Press. 2007.

To summarize, the liberal-modern epistemology as well as racial, market and nation-state relations of coloniality significantly shape the 1) content of extreme speech 2) limitations in the imagination of technology 3) complexity of meaning of content and 4) disparities in content moderation. Both as a technical problem of contextualization and a political problem that conceals colonial classification and its structuring effects on content moderation, the dichotomous conception of "human vs machine" thus glosses over pertinent issues around who should be involved in the process of content moderation and how content moderation should be critically appraised in relation to the broader problem of extreme speech as a market-driven, technologically-shaped, historically inflected and politically instrumentalized phenomenon.

Ethical scaling

Far from recognizing the process of involving human annotators as a political issue rather than a mere technical one, the involvement of human annotators in corporate content moderation is framed in the language of efficiency and feasibility, and often positioned in opposition to the necessities of "scaling." While human annotators are recognized as necessary at least until the machines pick up enough data to develop capacities to judge content, their involvement is seen as fundamentally in tension with machine-enabled moderation decisions that can happen in leaps, matching, to some degree, the hectic pace of digital engagements and data creation. Reading against this line of thinking, Tarleton Gillespie offers some important clarifications around scale and size, and why they should not be collapsed to mean the same. Building on Jennifer Slack's[31] work, he suggests that scale is "a specific kind of articulation: …different components attached, so they are bound together but can operate as one—like two parts of the arm connected by an elbow that can now 'articulate' their motion together in powerful but specific ways."[32] Content moderation on social media platforms similarly involves the articulation of different teams, processes and protocols, in ways that "small" lists of guidelines are conjoined with larger explanations of mandates; AI's algorithms learnt on a sample of data are made to work on much larger datasets; and, if we may add, small public policy teams stationed inside the company premises in Western metropoles articulate the daily navigations of policy heads in countries far and wide, as governments put different kinds of pressure on social media companies to moderate the content that flow on their platforms. These articulations then are not only "sociotechnical scalemaking"[33] but also political maneuvering, adjustments and moving the 'parts' strategically and deliberately, so what is learnt in one context can be replicated elsewhere.

31 Slack, Jennifer. 2006. Communication as articulation. In: Shepherd G, St. John J and Striphas T (eds) Communication as . . . : Perspectives on Theory. Thousand Oaks: SAGE Publications, pp.223–231.

32 Gillespie, 2020, p 2 .

33 Seaver, Nick. "Care and Scale: Decorrelative Ethics in Algorithmic Recommendation." *Cultural Anthropology* 36 (3), 2021, pp. 509–37. https://doi.org/10.14506/ca36.3.11.

Gillespie's argument is insightful in pointing out the doublespeak of commercial social media companies. As he elaborates:

> The claim that moderation at scale requires AI is a discursive justification for putting certain specific articulations into place—like hiring more human moderators, so as to produce training data, so as to later replace those moderators with AI. In the same breath, other approaches are dispensed with, as are any deeper interrogations of the capitalist, 'growth at all costs' imperative that fuels these massive platforms in the first place.

The AI4Dignity project is built on the recognition that scaling as an effort to create replicable processes for content moderation is intrinsically a political practice and should be seen in conjunction with regulatory attention to what scholars like Joan Donovan and Evegny Morozov have powerfully critiqued as the algorithmic amplification and political manipulation of polarized content facilitated by extractive digital capitalism.

We take this critique of digital capitalism alongside the sociotechnical aspects of the annotation process, and argue for a framework that recognizes that scaling as a process that makes "the small…have large effects"[34] and proceduralizes this process for its replication in different contexts as also, and vitally, a political one. It is political precisely because of how and whom it involves as "human annotators," the extent of resources and imaginations of technology that guide this process, and the deeper colonial histories that frame the logics of market, race and rationality within which it is embedded (and therefore has to be disrupted).

The AI4Dignity project is built on the recognition that scaling as an effort to create replicable processes for content moderation is intrinsically a political practice and should be seen in conjunction with regulatory attention to what scholars like Joan Donovan[35] and Evegny Morozov[36] have powerfully critiqued as the algorithmic amplification and political manipulation of polarized content facilitated by extractive digital capitalism. We define this combined attention to replicable moderation process as political praxis and critique of capitalist data hunger as "ethical scaling." In ethical scaling, the replicability of processes is conceived as a means to modulate data hunger and channel back the benefits of scaling toward protecting marginalized, vulnerable and historically disadvantaged communities. In other words, ethical scaling imagines articulation among different parts and components as geared towards advancing social justice agendas with critical attention to colonial structures of subjugation and the limits of liberal thinking, and recognizing that such articulation would mean applying breaks to content flows, investing resources for moderation, and embracing an inevitably messy process of handling diverse and contradictory inputs during annotation and model building.

In the rest of the paper, based on the learnings gained from the AI4Dignity project, we will describe ethical scaling for extreme speech moderation by considering both the operational and political aspects of involving "human annotators" in the moderation process.

34 Gillespie, 2020, p 2.

35 Donovan, Joan. "Why Social Media Can't Keep Moderating Content in the Shadows." *MIT Technology Review.* 2020a. https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/; Donovan, Joan. "Social-Media Companies Must Flatten the Curve of Misinformation," April 14, 2020b. https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/d41586-020-01107-z.

36 Morozov, Evgeny. *The Net Delusion: The Dark Side of Internet Freedom.* New York: Public Affairs. 2011.

AI4Dignity

Building on the critical insights into liberal constructions of the "human" and corporate appeals to "crowds," the AI4Dignity project has actively incorporated the participation of community intermediaries in annotating online extreme speech. The project has partnered with independent fact-checkers as critical community interlocutors who can bring cultural contextualization to AI-assisted extreme speech moderation in a meaningful way. Facilitating spaces of direct dialogue between ethnographers, AI developers and (relatively) independent fact-checkers who are not employees of large media corporations, political parties or social media companies is a key component of AI4Dignity. Aware of the wildly heterogenous field of fact-checking that range from large commercial media houses to very small players with commercial interests as well as the political instrumentalization of the very term "fact-checks" for partisan gains,[37] the project has sought to develop relations with fact-checkers based on whether they are independent (enough) in their operations and with explicit agendas for social justice. The scaling premise here is to devise ways that can connect, support and mobilize *existing communities* who have gained reasonable access to meaning and context of speech because of their involvement in online speech moderation of some kind.

By involving fact-checkers, AI4Dignity has sought to draw upon the professional competence of a relatively independent group of experts who are confronted with extreme speech both as part of the data they sieve for disinformation and as targets of extreme speech.

Without doubt, fact-checkers are already overburdened with verification-related tasks, but there is tremendous social value in involving them to flag extreme speech as a critical subsidiary to their core activities. Moreover, for fact-checkers, this collaboration also offers the means to foreground their own grievances as a target community of extreme speech. By involving fact-checkers, AI4Dignity has sought to draw upon the professional competence of a relatively independent group of experts who are confronted with extreme speech both as part of the data they sieve for disinformation and as targets of extreme speech. This way, AI4Dignity has tried to establish a process in which the "close cousin" of disinformation, namely, extreme speech and dangerous speech, are spotted during the course of fact-checkers' daily routines, without significantly interrupting their everyday verification activities.

The first step in the implementation of AI4Dignity has involved discussions among ethnographers, NLP researchers and fact-checkers to identify different types of problematic content and finalize the definitions of labels for manually annotating social media content. After agreeing upon the definitions of the three types of problematic speech as derogatory extreme speech (forms that stretch the boundaries of civility but could be directed at anyone, including institutions of power and people in positions of power), exclusionary extreme speech (explicitly or implicitly excluding people because of their belonging to a certain identity/community), and dangerous speech (with imminent danger of physical violence),[38] fact-checkers were requested to label the passages under the three categories.

37 ` For instance, in the UK, media reports in 2019 highlighted the controversies surrounding the Conservative party renaming their Twitter account "factcheckUK" https://www.theguardian.com/politics/2019/nov/20/twitter-accuses-tories-of-misleading-public-in-factcheck-row. In Nigeria, online digital influencers working for political parties describe themselves as "fact-checking" opponents and not fake news peddlers. https://mg.co.za/article/2019-04-18-00-nigerias-propaganda-secretaries/

Each gathered passage ranged from a minimum sequence of words that comprises a meaningful unit in a particular language to about six to seven sentences. Fact-checkers from Brazil, Germany, India, and Kenya, who participated in the project, sourced the passages from different social media platforms they found relevant in their countries and those they were most familiar with. In Kenya, fact-checkers sourced the passages from WhatsApp, Twitter and Facebook; Indian fact-checkers gathered them from Twitter and Facebook; the Brazilian team from WhatsApp groups; and fact-checkers in Germany from Twitter, YouTube, Facebook, Instagram, Telegram and comments posted on the social media handles of news organizations and right-wing bloggers or politicians with large followings.

In the second step, fact-checkers uploaded the passages via a dedicated WordPress site on to a database connected in the backend to extract and format the data for NLP model building. They also marked the target groups for each instance of labeled speech. On the annotation form, they identified the target groups from a dropdown list that included "ethnic minorities, immigrants, religious minorities, sexual minorities, women, racialized groups, historically oppressed castes, indigenous groups and any other." Only under "derogatory extreme speech" were annotators also able to select "politicians, legacy media, the state and civil society advocates for inclusive societies" as target groups. Fifty percent of the annotated passages were later cross-annotated by another fact-checker from the same country to check the inter-annotator agreement score.

In the third step, we created a collaborative coding space called "Counterathon" (a marathon to counter hate) where AI developers and partnering fact-checkers entered into an assisted dialogue to assess classification algorithms and the training datasets involved in creating them. This dialogue was facilitated by academic researchers with regional expertise and a team of student researchers who took down notes, raised questions, displayed the datasets for discussion and transcribed the discussions. We also had a final phase of reannotation of over fifty percent of the passages from Kenya based on the feedback we received during the Counterathon about including a new category (large ethnic groups) in the target groups.

Through these steps, the project has aimed to stabilize a more encompassing collaborative structure for what might be called a "people-centric process model" in which "hybrid" models of human–machine filters are able to incorporate dynamic reciprocity between AI developers, academic researchers and community intermediaries such as independent fact-checkers on a regular basis, and the entire process is kept transparent with clear-enough guidelines for replication. Figure 1 illustrates the basic architecture and components of this people-centric content moderation process.

However, the exercise of involving communities in content moderation is time

38 Dangerous speech definition is borrowed from Susan Benesch's work (2012), and the distinction between derogatory extreme speech and exclusionary extreme speech draws from Udupa (2021). Full definitions of these terms are available at https://www.ai4dignity.gwi.uni-muenchen.de. See Benesch, Susan. "Dangerous Speech: A Proposal to Prevent Group Violence." New York: World Policy Institute. 2012; Udupa, Sahana. "Digital Technology and Extreme Speech: Approaches to Counter Online Hate." In *United Nations Digital Transformation Strategy.* Vol. April. New York: United Nations Department of Peace Operations. 2021a. https://doi.org/10.5282/ubm/epub.77473.

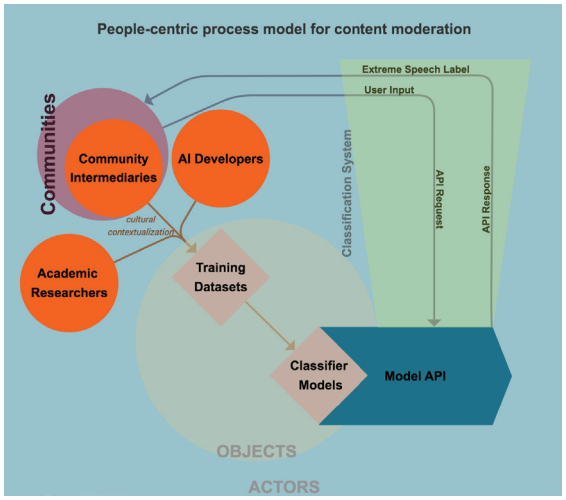
intensive and exhausting, and comes with the risks of handling contradictory inputs that require careful navigation and vetting. At the outset, context sensitivity is needed for label definitions. By defining derogatory extreme speech as distinct from exclusionary extreme speech and dangerous speech, the project has tried to locate uncivil language as possible efforts to speak against power in some instances, and in others, as early indications of exclusionary discourses that need closer inspection. Identification of target groups in each case provides a clue about the implications of online content, and whether the online post is merely derogatory or more serious. The three part typology has tried to bring more nuance to the label definitions instead of adopting an overarching term such as hate speech.

However, the exercise of involving communities in content moderation is time intensive and exhausting, and comes with the risks of handling contradictory inputs that require careful navigation and vetting.

However, even with a clear enough list of labels, selecting annotators is a daunting challenge. Basic principles of avoiding dehumanizing language, grounded understanding of vulnerable and historically disadvantaged communities, and knowledge around what kind of uncivil speech is aimed at challenging regressive power as opposed to legitimating harms within particular national or social contexts would serve as important guiding principles in selecting community annotators. AI4Dignity project has sought to meet the parameters by involving factcheckers with their close knowledge of extreme speech ecologies, professional training in factchecking, linguistic competence and a broad commitment to social justice (as indicated by their involvement in peace initiatives or a record of publishing factchecks to protect vulnerable populations).

By creating a dialogue between ethnographers, AI developers, and factcheckers, the project has tried to resolve different problems in appraising content as they emerged during the process of annotation and in delineating the target groups. However, this exercise is only a first step in developing a process of community intermediation in AI cultures, and it requires further development and fine tuning with future replications.

Figure 1



Architecture and components of the peope-centric process model for content moderation

Building on the learnings and findings from the project, we highlight below two distinct elements of the process model as critical aspects of ethical scaling in content moderation.

Iteration and experiential knowledge

As the opening vignettes indicate, the process of defining the labels and classification of gathered passages during the project was intensely laborious and dotted with uncertainty and contradiction. These confusions were partly a result of our effort to move beyond a binary classification of extreme and non-extreme and capture the granularity of extreme speech in terms of distinguishing derogatory extreme speech, exclusionary extreme speech and dangerous speech, and different target groups for these types. For instance, the rationale behind including politicians, media and civil society representatives who are closer to establishments of power (even if they hold opposing views) as target groups under "derogatory extreme speech" was to track expressions that stretch the boundaries of civility as also a subversive practice. For policy actions, derogatory extreme speech would require closer inspection, and possible downranking, counter speech, monitoring, redirection and awareness raising but not necessarily removal of content. However, the other two categories (exclusionary extreme speech and dangerous speech) require removal, with the latter (dangerous speech) warranting urgent action. Derogatory extreme speech also presented a highly interesting corpus of data for research purposes as it represented online discourses that challenged the protocols of polite language to speak back to power, but it also constituted a volatile slippery ground on which what is comedic and merely insulting could quickly slide down to downright abuse and threat.[39] For content moderation, such derogatory expressions can serve as the earliest cultural cues to brewing and more hardboiled antagonisms.

During the course of the project, instances of uncertainty about the distinction between the three categories were plentiful, and the Krippendorff (2003) intercoder agreement score (alpha) between two fact-checkers from the same country averaged 0.24.[40] However, two moments stand out as illustrative of the complexity.

39 Udupa, Sahana. "Gaali Cultures: The Politics of Abusive Exchange on Social Media." New Media and Society 20 (4), 2017, pp. 1506–22. https://doi.org/10.1177/1461444817698776.

40 Although, as mentioned, there was consensus that all selected passages were instances of extreme speech. The inter-coder agreement scores are also similar to other works in the field: In Ross, Björn, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, Michael Wojatzki. "Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis", ArXiv:1701.08118 [cs.CL]. 2017, a German dataset, α was between 0.18 and 0.29, in Sap, Maarten, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. "Social bias frames: Reasoning about social and power implications of language." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, the α score was 0.45, while in Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. "Multilingual and multi-aspect hate speech analysis", In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, a multilingual dataset, α was between 0.15 and 0.24. Also, a majority of these works include neutral examples as well.

During several rounds of discussion, it became clear that the list of target groups was itself an active political choice, and it had to reflect the regional and national specificities to the extent possible. In the beginning, we had proposed a list of target groups that included ethnic minorities, immigrants, religious minorities, sexual minorities, racialized groups, historically oppressed indigenous groups and any other. Fact-checkers from Brazil pointed out the severity of online misogyny and suggested adding "women" to the list. Fact-checkers from Kenya pointed out that "ethnic minorities" was not a relevant category since Kikuyu and Kalenjin ethnic groups around whom a large proportion of extreme speech circulated were actually large ethnic groups. Small ethnic groups, they noted, did not play a significant role in the country's political discourse. While this scenario itself revealed the position of minorities in the political landscape of the country, it was difficult to label extreme speech without giving the option of selecting "large ethnic groups" in the list of target groups. Fact-checkers from Germany pointed out that "refugees" were missing from the list, since immigrants—usually welcomed and desired at least for economic reasons—are different from refugees who are derided as unwanted. We were not able to implement this distinction during the course of the project, but we noted this as a significant point to incorporate in future iterations.

During the annotation process, fact-checkers brought up another knotty issue in relation to the list of target groups. Although politicians were listed only under the derogatory speech category, fact-checkers wondered what to make of politicians who are women or who have a migration background. The opening vignette from Kenya about the "Sugoi thief" signals a scenario, where politicians become a synecdoche for an entire targeted community. "Sawsan Chebli is a politician," pointed out a fact-checker from Germany, "but she also has migration background." Chebli, a German politician born to parents who migrated to Germany from Palestine, is a frequent target for right-wing groups. Fact-checkers from India highlighted the difficulty of placing Dalit politicians and Muslim politicians only under the category of "politicians" and therefore only under "derogatory speech" because targeting them could lead to exclusionary speech against the communities they represented. In such cases, we advised the fact-checkers to label this as exclusionary speech and identify the target groups of such passages as "ethnic minorities," "women," "historically disadvantaged caste groups," "immigrants," or other relevant labels.

Some fact-checkers and participating academic intermediaries also suggested that the three labels—derogatory, exclusionary, and dangerous—could be broken down further to capture the granularity. For instance, under derogatory speech, there could be "intolerance talk" (speech that is intolerant of opposition); "delegitimization of victimhood" (gaslighting and undermining people's experiences of threat and right to protection); and "celebratory exclusionary speech" (in which exclusionary discourse is ramped up not by using hurtful language but by celebrating the glory of the dominant group). Duncan Omanga, the academic expert on Kenya, objected to the last category and observed that "Mobilization of ethnic groups in Kenya by using glorifying discourses is frequent especially during the elections in the country. Labeling this as derogatory is complicated since it is internalized as the nature of politics and commonly legitimized." Although several issues could not be resolved partly because of the limitations of time and resources in the project, curating such observations has been helpful in highlighting the importance of

iteration in not only determining the labels but also linking the selection of labels with specific regulatory goals. In cases where removal of content versus retaining it is the primary regulatory objective, it is helpful to have a simpler classification, but breaking down the categories further would be important for research as well as for fine grained interventions involving counter speech and positive narratives targeting specific kinds of vitriolic exchange online.

Although several issues could not be resolved partly because of the limitations of time and resources in the project, curating such observations has been helpful in highlighting the importance of iteration in not only determining the labels but also linking the selection of labels with specific regulatory goals.

Moreover, the value of iteration is crucial for embedding embodied knowledge of communities most affected by extreme speech into the annotation process, and for ensuring that categories represent the lived experiences and accretions of power built up over time. Without doubt, stark and traumatizing images and messages can be (and should be) spotted by automation since it helps to avoid the emotional costs of exposure to such content in online content moderation. This does not discount responsible news coverage on violence that can sensitize people about the harms of extreme content, but in the day-to-day content moderation operations for online discourses, automation can provide some means for (precariously employed) content moderators to avoid exposure to violent content. Beyond such obvious instances of dehumanizing and violent content, subtle and indirect forms of extreme expression require the keen attention and experiential knowledge of communities who advocate for, or themselves represent, groups targeted by extreme speech.

Participating fact-checkers in the project—being immigrants, LGBTQI+ persons or members of the targeted ethnic or caste groups—weighed in with their own difficult experiences with extreme speech and how fragments of speech acts they picked up for labeling were not merely "data points" but an active, embodied engagement with what they saw as disturbing trends in their lived worlds. Indeed, ethical scaling as conceptualized in AI4Dignity's iterative exercise does not merely connect parts and components for actions that can magnify effects and enable efficiency, but grounds this entire process by connecting knowledges derived from the experiences of inhabiting and confronting the rough and coercive worlds of extreme speech. As the fact-checker from Brazil expressively shares their experience of spotting homophobic content in the opening vignettes of this essay, hatred that hides between the lines, conceals behind the metaphors, cloaks in 'humor' and mashups, or clothes itself in the repertoire of 'plain facts'—the subtleties of speech that deliver hate in diverse forms—cannot be fully captured by cold analytical distance, or worse still, with an approach that regards moderation as a devalued, cost-incurring activity in corporate systems. As the fact-checker in the opening vignette intoned by referencing the hoax message on Biden appointing an LGBTQI+ person to head the education department, it is the feel for the brewing trouble and insidious coding of hatred between the lines that helps him to flag the trouble as it emerges in different guises:

> As I told you, for example the transexual content was very typical hate speech included into a piece of misinformation, but not that explicit at all. So, you have to be in the position of someone who is being a target of hate speech/misinformation, to figure out that this piece is hate speech, not only misinformation. So that was making me kind of nervous, when I was reading newspapers every day and I was watching social media and I see that content spreading around, because this is my opinion on it and its much further, it's much more dangerous than this [a mere piece of misinformation]. You are…you are telling people that it's a problem that a transgender person, a transsexual is going to be in charge of education because somehow it's

a danger to our children. So, it makes me kind of uncomfortable and that's why we decided to join the project [AI4Dignity].

As we navigated extreme speech passages and the thick narratives around how factcheckers encountered and flagged them for the project, it became clear that iteration is an inevitably intricate and time intensive exercise. The AI4Dignity findings show that the performance of ML models (BERT) based on the datasets we gathered averaged performance metrics of other hate speech detection projects, but the model performance in detecting target groups was more than average.[41] These results underscore the point that ethical scaling is not merely about gauging the performance of the machine for its accuracy in the first instance but involves ethical means for scaling a complex process so that problems of cultural contextualization and bias are addressed through reflexive iterations in a systematic and transparent manner.

....hatred that hides between the lines, conceals behind the metaphors, cloaks in 'humor' and mashups, or clothes itself in the repertoire of 'plain facts'—the subtleties of speech that deliver hate in diverse forms—cannot be fully captured by cold analytical distance, or worse still, with an approach that regards moderation as a devalued, cost-incurring activity in corporate systems.

Name-calling as seed expressions

Such an iterative process, while grounding content moderation, also offers specific entry points to catch signals from types of problematic content that do not contain obvious watchwords, and instead employ complex cultural references, local idioms or multimedia forms. We present one such entry point as a potential scalable strategy that can be developed further in future projects.

Our experience of working with longer real world expressions gathered by fact-checkers rather than keywords selected by academic annotators[42] has shown the importance of name-calling as a useful shorthand to pick up relevant statistical signals for detecting extreme speech. This involves curating, with the help

41 The performance was also constrained by the fact that no comparison corpus for "neutral" passages was given, and instead only examples for the three labels of extreme speech were collected. However, our datasets are closer to real world instances of hateful language. Several hate speech detection projects have relied on querying of keywords, while AI4Dignity has sourced the passages from actual discussions online through community intermediaries. The performance of BERT on hate speech datasets is examined thoroughly in Swamy, Steve Durairaj, Anupam Jamatia, and Björn Gambäck. "Studying generalisability across abusive language detection datasets." In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), 2019. In Founta, Antigoni-Maria, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. "Large scale crowdsourcing and characterization of twitter abusive behavior", In 11th International Conference on Web and Social Media, ICWSM 2018. AAAI Press, 2018, the F1 score is 69.6. In Davidson, Thomas, Dana Warmsley, Michael Macy, Ingmar Weber. "Automated hate speech detection and the problem of offensive language", In International AAAI Conference on Web and Social Media, 2017, F1 is 77.3 while in Waseem, Zeerak, Dirk Hovy. "Hateful symbols or hateful people? predictive features for hatespeech detection on Twitter." In Proceedings of the NAACL Student Research Workshop, 2016, F1 score is at 58.4. In all those datasets, the majority of content is neutral, an intuitively easier task. In our work, multilingual BERT (mBERT) can predict the extreme speech label of text with an F1 score of 84.8 for Brazil, 64.5 for Germany, 66.2 for India and 72.8 for Kenya. When predicting the target of extreme speech, mBERT scored 94.1 (LRAP, label ranking average precision) for Brazil, 90.3 in Germany, 92.8 in India and 85.6 in Kenya.

of community intermediaries such as fact-checkers, an evolving list of putdowns and name-calling that oppressive groups use in their extreme speech attacks, and mapping them onto different target groups with a contextual understanding of groups that are historically disadvantaged (e.g., Dalits in India), groups targeted (again) in a shifting context (for instance, the distinction between 'refugees' and 'immigrants' in Europe), those instrumentalized for partisan political gains and ideological hegemony (e.g., different ethnic groups in Kenya or the religious majority/religious minority distinction in India) or groups that are excluded because of a combination of oppressive factors (e.g., Muslims in India or Europe). This scaling strategy clarifies that the mere identification of name-calling and invectives without knowledge of target communities can be misleading.

...interactions with fact-checkers helped us to sieve over twenty thousand extreme speech passages for specific expressions that can potentially lead to exclusion, threat and even physical danger.

For instance, interactions with fact-checkers helped us to sieve over twenty thousand extreme speech passages for specific expressions that can potentially lead to exclusion, threat and even physical danger. Most of these provocative and contentious expressions were still nested in the passages that fact-checkers labeled as "derogatory extreme speech", but, as mentioned earlier, derogatory expressions could be used to build a catalogue for early warning signals with the potential to normalize and banalize exclusion.

Interestingly, we found that such expressions are not always single keywords, although some unigrams are helpful in getting a sense of the discourse. They are trigrams or passages with a longer word count[43] often with no known trigger words but contain implicit meanings, indirect dog whistles and ingroup idioms. In Germany, exclusionary extreme speech passages that fact-checkers gathered had several instances of "gehört nicht zu" [does not belong] or "nicht mehr" [no more or a sentiment of having lost something], signaling a hostile opposition to refugees and immigrants. Some expressions had keywords that were popularized by right-wing politicians and other public figures, either by coining new compound words or injecting well-meaning descriptions with insidious sarcasm. For instance, in right-wing discourses, it was common to refer to refugees as "Goldstücke." A German politician from the center left SPD party, Martin Schulz, in a speech at Hochschule Heidelberg made the statement, "Was die Flüchtlinge zu uns bringen, ist wertvoller als Gold. Es ist der unbeirrbare Glaube an den Traum von Europa. Ein Traum, der uns irgendwann verloren gegangen ist [What refugees bring to us is

42 A large number of machine learning models rely on keyword-based approaches for training data collection, but there have been efforts lately to "leverage a community-based classification of hateful language" by gathering posts and extracting keywords used commonly by self-identified right-wing groups as training data. See Saleem, Haji Mohammed, Kelly P. Dillon, Susan Benesch, and Derek Ruths. "A Web of Hate: Tackling Hate Speech in Online Social Spaces." ArXiv Preprint ArXiv:1709.10159. 2017. The AI4Dignity project builds on the community-based classification approach instead of relying on keywords sourced by academic annotators. We have aimed to gather extreme speech data that is actively selected by community intermediaries, thereby uncovering characteristic complex expressions, including those containing more than a word.

43 The average word count for passages in German was 24.9; in Hindi was 28.9; in Portuguese was 16.2; in Swahili was 14.7; in English & German was 22.9; in English & Hindi was 33.0; in English & Swahili was 24.3; in English in Germany was 6.3, in English in India was 24.1; in English in Kenya was 28.0.

more valuable than gold. It is the unwavering belief in the dream of Europe. A dream that we lost at some point]."[44] In xenophobic circles, this expression was picked up and turned into the term "Goldstücke," which is sarcastically used to refer to immigrants/refugees. Similarly, academic intermediary Laura Csuka in the German team highlighted another interesting expression, "in der BRD [in the Federal Republic of Germany] as indicating an older age group whose nostalgia could give a clue about its possible mobilization for xenophobic ends.

In Kenya, the term "Jorabuon" used by Luos refers to Kikuyus, and hence, as one of the fact-checkers pointed out, "even if they are communicating the rest in English, this main term is in the mother tongue and could seed hostility." For communicative purposes, it also holds the value of in-group coding, since terms such as this one, at least for some time, would be intelligible to the community that coins it or appropriates it. In this case, Luos were sharing the term "Jorabuon" to refer to Kikuyus. The word "rabuon" refers to Irish potatoes. Kikuyu, in this coinage, are likened to Irish potatoes since their cuisine prominently features this root, and the mocking name marks them as a distinct group. "It is used by Luos when they don't want the Kikuyu to realize that they are talking about them," explained a fact-checker. Such acts of wordplay that test the limits of usage standards gain momentum especially during the elections when representatives of different ethnic groups contest key positions.

Hashtag #religionofpeace holds a similar performative power within religious majoritarian discourses in India. All the participating fact-checkers labeled passages containing this hashtag as derogatory. One of them explained, "#religionofpeace is a derogatory term [aimed at Muslims] because the irony is implied and clear for everybody." Certain keywords are especially caustic, they pointed out, since they cannot be used in any well-meaning context. One of them explained, "Take the case of 'Bhimte,' which is an extremely derogatory word used against the marginalized Dalit community in India. I don't think there is any way you can use it and say I did not mean that [as an insult]. This one word can convert any sentence into hate speech." Fact-checkers pointed to a panoply of racist expressions and coded allusions to deride Muslims and Dalits, including "Mulle," "Madrasa chaap Moulvi" [referring to Muslim religious education centers] and "hara virus" [green virus, the color green depicting Muslims], and the more insidious Potassium Oxide [K20 which phoenetically alludes to "Katuwon" and "Ola Uber" [two riding apps which together phonetically resemble Alla Ho Akbar].

Within online discourses, instrumental use of shifting expressions of name-calling, putdowns and invectives is structurally similar to what Yarimar Bonilla and Jonathan Rosa eloquently describe as the metadiscursive functions of hashtags in "forging a shared political temporality," which also "functions semiotically by marking the intended significance of an utterance."[45] Since name-calling in extreme speech contexts takes up the additional communicative function of coding the expressions for in-group sharing, some of them are so heavily coded that anyone outside the community would be confused or completely fail to grasp the

44 Stern.de. "Bremer Landgericht Gibt Facebook Recht: Begriff 'Goldstück' Kann Hetze Sein." June 21, 2019. https://www.stern.de/digital/bremer-landgericht-gibt-facebook-recht--begriff--goldstueck--kann-hetze-sein-8763618.html.

intended meaning. For instance, in the India list, fact-checkers highlighted an intriguing expression in Hindi, "ke naam par" [in the name of]. One of the participating fact-checkers understood this expression as something that could mean "in the name of the nation," signaling a hypernationalistic rhetoric, but thought it did not have any vitriolic edge. Another fact-checker soon interjected and explained: "'Ke naam par' is used for the scheduled caste community because they would say 'In the name of scheduled castes' when they are taking up the reservation in the education system and jobs.[46] This is a very common way to insult scheduled castes because they are called people who are always ready to take up everything that is coming free, mainly jobs or seats in medical and engineering institutes." Although "ke naam par" is invoked in a variety of instances including its use as a common connecting phrase in Hindi, its specific invocation in the right-wing discursive contexts revealed its function as a coded in-joke. During these instances of exchange between factcheckers, it became clear to us that iteration involved not only feeding the AI models with more data but also a meaningful dialogue between community intermediaries and academics so a fuller scope of the semiotic possibilities of coded expressions come into view.

During these instances of exchange between fact-checkers, it became clear to us that iteration involved not only feeding the AI models with more data but also a meaningful dialogue between community intermediaries and academics so a fuller scope of the semiotic possibilities of coded expressions come into view.

For sure, many of name-calling expressions and putdowns have an inevitable open-endedness and appear in diverse contexts, including well-meaning invocations for inclusive politics and news reportage, but they still serve as useful signaling devices for further examination. In most cases, participating fact-checkers brought their keen understanding of the extreme speech landscape, avowing that they have a "sense" for the proximate conversational time-space in which such expressions appeared online. As a fact-checker from India put it, they have "a grasp of the intentions" of users who posted them.

Are existing machine learning models and content moderation systems equipped to detect such expressions identified through collaborative dialogue and iteration? We carried out two tests and found several gaps and limitations in the extreme speech detection and content moderation practices of large social media companies such as Google and Twitter. Although Facebook and WhatsApp constituted prominent sources of extreme speech instances that fact-checkers gathered for the project, we were unable to include them in the tests due to severe restrictions on data access on these platforms and applications.

Perspective API test

For the first test, we ran relevant passages in the project database on Perspective API—a machine learning model developed by Google to assign a toxicity score (see Table 1).[47] We obtained an API key for Perspective[48] to run the test. Since Per-

45 Bonilla, Yarimar, and Jonathan Rosa. "#Ferguson: Digital Protest, Hashtag Ethnography and the Racial Politics of Social Media in the United States." *American Ethnologist* 42 (1), 2015, pp. 4–17. https://doi.org/10.1111/amet.12112, p 5.

46 Scheduled castes and scheduled tribes are bureaucratic terms to designate most oppressed caste groups for state affirmation policies in India.

47 https://www.perspectiveapi.com accessed 13 July 2021.

48 https://support.perspectiveapi.com/s/docs-get-started

spective API supports only English, French, German, Italian, Portuguese, Russian and Spanish for different attributes and Hindi only for the "toxicity" attribute, data for English (3,761 passages from all the countries), German (4,945 passages), Portuguese (5,245 passages), English/German (69 passages), Hindi (2,775 passages) and Hindi/English (1,162 passages) for a total of 17,957 passages were tested on available attributes. While accessing the API, the language of the input passages was not set, allowing Perspective to predict the language from the text. This is likely to be a more realistic scenario since content moderation tools often do not have the metadata on language. We computed six attributes that Perspective API identifies as toxicity, severe toxicity, identity attack, threat, profanity and insult for all the selected passages.[49]

We computed the averages for the three AI4Dignity labels (derogatory, exclusionary and dangerous speech) for the above languages. A major limitation is that mapping the three labels used in AI4Dignity to the Perspective API attributes is not straightforward. Perspective attributes are a percentage: the higher the percentage, the higher the chance a 'human annotator' would agree with the attribute. Based on the definitions of the attributes in both the projects, we interpreted correspondence between derogatory extreme speech in AI4Dignity and toxicity, profanity and insult in the Perspective model; between exclusionary extreme speech and severe toxicity and identity attack; and between dangerous speech and threat.[50]

Table 1 presents the breakdown of the score distribution for different attributes in AI4Dignity and Perspective. The derogatory passages in English across all the countries received a score of 43 (represented as 0.43 in the table) for toxicity and 41 for insult whereas exclusionary speech scored only 22 for severe toxicity and 32 for identity attack. Dangerous speech received a higher score of 50 for threat. A closer analysis also reveals that English language passages in Kenya received lower corresponding scores, especially for exclusionary speech. Exclusionary extreme speech in English from Kenya received a score of 14 for severe toxicity and 21 for identity attack; and dangerous speech in English received a score of 49. In other words, the threat level of dangerous speech passages in English language from Kenya was evaluated just at 49. English passages from India are assessed with 47/toxicity and 43/insult for derogatory speech; 36/severe toxicity and 51/identity attack for exclusionary speech; and 62/threat for dangerous speech. English passages from Germany also received lower scores for derogatory speech (28/toxicity and 24/insult) but scored higher for exclusionary speech (56/severe toxicity and 78/identity attack). There were no dangerous speech passages in English from Germany in the dataset. These results signal culturally specific uses of English, which the existing models find difficult to categorize. In comparison, the model performed better for

49 For descriptions of these categories, see https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages. Since the API restricts users to only one request per second, an artificial delay of 1.1 second was added between two requests so that all requests are processed. A 0.1 second buffer was added for any potential latency issues.

50 Derogatory, exclusionary and dangerous forms of extreme speech collected in our dataset do not correspond to mild forms of toxicity such as positive use of curse words as included in the "toxicity" class of Perspective API: "severe toxicity: This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words." https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages.

German-only and Portuguese-only passages for Germany and Brazil respectively across all the three categories. German derogatory passages received a score of 63 for toxicity and 69 for insult; exclusionary passages with 57 for severe toxicity and 80 for identity attack; and dangerous passages with 76 for threat. Brazilian Portuguese passages were correspondingly 85/toxicity and 86/insult for derogatory; 84/severe toxicity and 88/identity attack for exclusionary; and 74/threat for dangerous speech. However, Hindi passages in the derogatory extreme speech category received an average of just 53 for toxicity.

We compared the Perspective API scores with AI4Dignity categories using one more related metric. For this, we carried out a percentile test to assess different Perspective scores for each language. Table 2 shows the percentage of AI4Dignity passages that scored below 10 (denoted in the table as below_10), and similarly over_25, over_50, over_75 and over_90. This table shows that a number of examples were declared as "clean" by Perspective API (i.e., below_10) for the corresponding categories and also that a significant number of passages were just over_50. In

Table 1

	Toxicity	Severe Toxicity	Profanity	Identity	Insult	Threat	Total
eng_all	0,43	0,28	0,24	0,32	0,41	0,34	3490
eng_der	0,48	0,31	0,28	0,33	0,47	0,32	1972
eng_exc	0,35	0,22	0,19	0,32	0,33	0,28	916
eng_dan	0,39	0,28	0,19	0,30	0,33	0,50	602
eng_kenya_all	0,42	0,26	0,23	0,27	0,40	0,32	2680
eng_kenya_der	0,48	0,30	0,28	0,29	0,48	0,31	1560
eng_kenya_exc	0,27	0,14	0,13	0,21	0,28	0,19	585
eng_kenya_dan	0,38	0,25	0,17	0,28	0,32	0,49	535
eng_germany_all	0,60	0,50	0,45	0,71	0,53	0,34	6
eng_germany_der	0,28	0,18	0,16	0,39	0,24	0,15	1
eng_germany_exc	0,67	0,56	0,51	0,78	0,58	0,38	5
eng_germany_dan	0,00	0,00	0,00	0,00	0,00	0,00	0
eng_india_all	0,48	0,36	0,29	0,50	0,43	0,40	804
eng_india_der	0,47	0,34	0,29	0,50	0,43	0,34	411
eng_india_exc	0,48	0,36	0,29	0,51	0,43	0,44	326
eng_india_dan	0,53	0,44	0,34	0,49	0,42	0,62	67
deu_all	0,64	0,55	0,47	0,74	0,62	0,44	4903
deu_der	0,63	0,53	0,47	0,69	0,61	0,40	2602
deu_exc	0,66	0,57	0,48	0,80	0,63	0,48	2285
deu_dan	0,73	0,73	0,61	0,87	0,69	0,76	16
bra_all	0,84	0,80	0,79	0,81	0,85	0,61	5036
bra_der	0,85	0,81	0,80	0,81	0,86	0,60	4702
bra_exc	0,86	0,84	0,80	0,88	0,86	0,72	115
bra_dan	0,63	0,56	0,51	0,71	0,62	0,74	219
eng_deu_all	0,74	0,69	0,66	0,80	0,72	0,46	70
eng_deu_der	0,69	0,62	0,59	0,72	0,67	0,36	28
eng_deu_exc	0,77	0,74	0,70	0,86	0,74	0,52	42
eng_deu_dan	0,00	0,00	0,00	0,00	0,00	0,00	0
eng_hindi_all	0,51	-	-	-	-	-	1162
eng_hindi_der	0,45	-	-	-	-	-	207
eng_hindi_exc	0,44	-	-	-	-	-	132
eng_hindi_dan	0,54	-	-	-	-	-	823
hin_all	0,51	-	-	-	-	-	2755
hin_der	0,53	-	-	-	-	-	1532
hin_exc	0,46	-	-	-	-	-	860
hin_dan	0,57	-	-	-	-	-	363

Perspective scores for AIDignity passages across all types of extreme speech

other words, the problematic nature of a large majority of content was considered as mild. 73% of the derogatory passages in English received just over_25 score for toxicity, whereas 73% of dangerous passages in the same language received just over_25 score for threat. In contrast, 77% of exclusionary passages composed fully in Portuguese were scored as over_75 for severe toxicity and 90% of the exclusionary passages in the same language were rated as over_75 for identity attack. 74% of exclusionary extreme speech passages in German were similarly rated high (over_75) for identity attack but only 34% passages were rated with the same score for severe toxicity (and 60% of the passages scored over_50 for severe toxicity). For Hindi passages that can be assessed only for toxicity scores on Perspective API, the model's performance was weaker. A large majority of derogatory passages (93%) were rated as just over_25 for toxicity and 77% of English and Hindi mixed language passages were rated as over_25 for the same. Far fewer passages (13% in Hindi and 12% for English-Hindi) received a score of over_75. These results indicated that the model rated the instances as less than mild. Upon closer examantion of English language passages, we also found that English passages (from Kenya and India combined) received on average much lower values (45% toxicity) compared to the English passages from Germany (60%). Hence, this result also signaled the disparities in the model performance for English, especially in assessing culturally inflected features of English usage in countries in the global South (here India and Kenya) in the extreme speech context.

English trigger expressions in the German dataset

To examine one more aspect of Perspective API's model, we tested if this was more sensitive to common trigger words in English that have acquired some global momentum because of transnational social media and, by the same token, less equipped to detect problematic content that did not contain such words and phrases but were composed entirely in languages other than English. This qualitative analysis was prompted by our observation in the German language dataset that a higher proportion of passages in mixed language (German and English) were picked up by the model as severely toxic as opposed to fully German passages. If 59% of English-German mixed language passages received a score of more than 75 for severe toxicity, only 33% of German-only passages were above this score (See Table 3 and Figure 2).

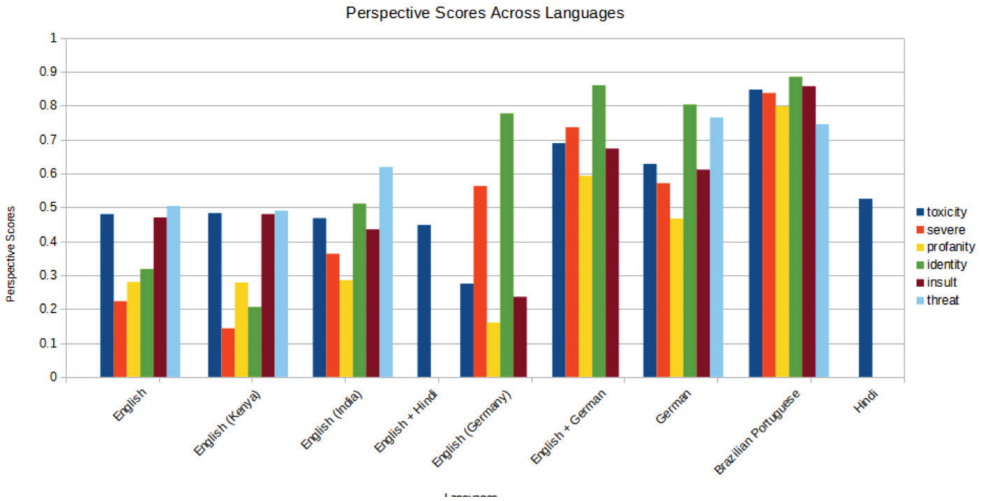
A subsequent qualitative analysis revealed that most of these mixed passages were German texts featuring one or more English trigger words or phrases that appear to have been picked by the Perspective model as cues for severe toxicity. In our dataset, we found them to be frequently-used hateful expressions in English also used in non-English extreme speech contexts, for example, "shithole countries," "black lies matter," "in cold blood," "new world order," and "wake up." On the one hand, the very salience of these English expressions in the German dataset revealed the global circulatory force of hateful catchphrases that now transcend national boundaries. On the other hand, with regard to content moderation, the existing models, as illustrated by Perspective API, tend to mark expressions with such catchphrases as hateful more extensively and clearly than those that contain more complex non-English expressions. For example, in terms of single words, most of the passages (92%) containing the most frequent English trigger word "shithole" (39 passages) have a high score for severe toxicity over_75 and

Table 2

INSULT	below_10	over_25	over_50	over_75	over_90	total
eng_der	0,07	0,71	0,44	0,21	0,08	1972
eng_kenya_der	0,07	0,71	0,45	0,23	0,09	1560
eng_germany_der	0	0	0	0	0	1
eng_india_der	0,08	0,69	0,41	0,14	0,03	411
deu_der	0,02	0,91	0,67	0,35	0,11	2602
bra_der	0,01	0,97	0,93	0,81	0,62	4702
eng_deu_der	0	0,93	0,75	0,54	0,21	28
TOXICITY	**below_10**	**over_25**	**over_50**	**over_75**	**over_90**	**total**
eng_der	0,05	0,73	0,47	0,2	0,07	1972
eng_kenya_der	0,05	0,73	0,47	0,22	0,08	1560
eng_germany_der	0	1	0	0	0	1
eng_india_der	0,07	0,75	0,46	0,15	0,03	411
deu_der	0,02	0,92	0,66	0,34	0,12	2602
bra_der	0,01	0,98	0,94	0,8	0,51	4702
eng_deu_der	0	0,96	0,75	0,39	0,14	28
hin_der	0,01	0,93	0,57	0,13	0,01	1532
eng_hindi_der	0,06	0,77	0,42	0,12	0,01	207
PROFANITY	**below_10**	**over_25**	**over_50**	**over_75**	**over_90**	**total**
eng_der	0.30	0.39	0.20	0.08	0.03	1972
eng_kenya_der	0.31	0.38	0.20	0.09	0.04	1560
eng_germany_der	0	0	0	0	0	1
eng_india_der	0.23	0.42	0.18	0.06	0.02	411
deu_der	0.13	0.66	0.49	0.26	0.08	2602
bra_der	0.01	0.94	0.88	0.70	0.50	4702
eng_deu_der	0.07	0.78	0.57	0.46	0.21	28
IDENTITY	**below_10**	**over_25**	**over_50**	**over_75**	**over_90**	**total**
eng_exc	0,22	0,45	0,24	0,1	0,02	916
eng_kenya_exc	0,3	0,28	0,07	0,02	0,01	585
eng_germany_exc	0	1	1	0,6	0,2	5
eng_india_exc	0,06	0,75	0,53	0,25	0,05	326
deu_exc	0,02	0,95	0,9	0,74	0,48	2285
bra_exc	0,01	0,98	0,97	0,9	0,71	115
eng_deu_exc	0	1	0,95	0,81	0,62	42
SEVERE	**below_10**	**over_25**	**over_50**	**over_75**	**over_90**	**total**
eng_exc	0,44	0,34	0,13	0,04	0	916
eng_kenya_exc	0,6	0,18	0,05	0,01	0	585
eng_germany_exc	0	1	0,4	0,4	0	5
eng_india_exc	0,18	0,63	0,27	0,09	0,01	326
deu_exc	0,07	0,85	0,6	0,34	0,15	2285
bra_exc	0,01	0,97	0,9	0,77	0,57	115
eng_deu_exc	0,02	0,9	0,83	0,67	0,24	42
THREAT	**below_10**	**over_25**	**over_50**	**over_75**	**over_90**	**total**
eng_dan	0,09	0,73	0,49	0,31	0,08	602
eng_kenya_dan	0,09	0,73	0,47	0,28	0,07	535
eng_germany_dan	0	0	0	0	0	0
eng_india_dan	0,09	0,78	0,63	0,51	0,21	67
deu_dan	0	0,88	0,75	0,69	0,69	16
bra_dan	0	0,96	0,85	0,59	0,24	219
eng_deu_dan	0	0	0	0	0	0

Percentile test for Perspective API scores for corresponding extreme speech types

Figure 2



Perspective scores for corresponding extreme speech types across languages

none of them are classified as clean (below_10; see Table 3).[51] This indicates that this expression was picked up by the German Perspective API model as a marker (trigger word) for severe toxic speech. Interestingly, similar results were obtained in an exemplary analysis for two German trigger words ("Homos" ['homosexuals'], "Scheiss" [shit]; see Table 3), which were selected from the most frequent words from the passages scoring over_90 toxicity (top words). These passages score high for severe toxicity (73% and 88% over_75) and do not have any false negatives (all are over_10 for severe toxicity as well as toxicity and almost all over_25 for severe toxicity).

These results are also corroborated by a manual test using the Perspective API Web Interface, where "shithole" as a single input scores high for toxicity, which indicates that this expression triggers the German Perspective model regardless of the context. Similar results were obtained for the German word, "Scheiss" [shit]. In contrast, exclusionary constructions that do not contain swear words like "X raus" [X out] receive varying toxicity scores depending on context. This shows that beyond the single words, more obvious exclusionary phrases like "Ausländer raus" [foreigners out] (which received a rather high score for severe toxicity as part of our data analysis, see Table 5) also act as trigger expressions.

In contrast, an analysis of passages containing other top words ("Deutschland," "Ausländer" [foreigner], "Moslems" [Muslims]) in the dataset shows that a number of these passages are classified as clean (below_10_severe = false negatives) by the Perspective API (see Table 4), which indicates that these words are not triggering the model on their own. Further qualitative analysis of these false negatives shows that these typically do not contain any trigger expressions but that they represent

51 Passages were obtained by applying simple search patterns covering variants and common misspellings of these trigger words.

more complex, concealed and argumentative forms of exclusionary discourse practices (see Table 4).

This analysis indicates that current models are not suitable for recognizing less obvious forms of exclusionary and dangerous extreme speech as documented in our AI4Dignity data sets. The results reveal gaps in machine detection models even within well-resourced language contexts such as Germany. Since advances in models gaining language understanding are promising but still much remains to be done[52] and extreme speech expressions keep shifting, it is important to make

This analysis indicates that current models are not suitable for recognizing less obvious forms of exclusionary and dangerous extreme speech as documented in our AI4Dignity data sets.

Table 3

	avg_severe_score:	over_75_severe:	below_10_severe:	below_10_toxic:	below_25_severe:	type:
deu_all	0,55	33%	9%	1%	18%	(German passages)
"Deutschland"	0,53	27%	5%	2%	14%	neutral word
"Ausländer"	0,57	32%	2%	1%	11%	neutral word
"Moslems"	0,68	46%	2%	0%	7%	neutral word
"Homos"	0,81	73%	0%	0%	0%	trigger word
"Scheiss"	0,87	88%	0%	0%	1%	trigger word
"Shithole"	0,85	92%	0%	0%	0%	trigger word
eng_deu_all	0,67	59%	3%	0%	13%	(mixed passages)

Trigger words in the extreme speech dataset in German language

Table 4

SEVERE TOXICITY	
0,09	"die deutschen sind in Deutschland in der unterzahl" ("germans are outnumbered in Germany")
0,09	"Was haben Ausländer in unserer Politik zu suchen !??" ("What are foreigners doing in our politics !??")
0,04	"Moslems sind anders sozialisiert." ("Muslims are socialized differently.")
0,09	"Scheinbare Fakten sind, dass die Moslems Schuld an dem Ausbreiten der Pandemie sind." ("Apparent facts are that muslims are to blame for the spread of the pandemic.")
0,09	"Ich will endlich lesen das Deutschland die Millionen "Ausländischen Hartz4 Empfänger" zurück schickt!" ("I finally want to read that Germany is sending back the millions of "foreign Hartz4 recipients"!"; Hartz4 = colloquial term for unemployment benefits in Germany)
0,09	"Alle Welt investiert in Fortschritt und Evolution, nur Deutschland in Flüchtlinge" ("The whole world is investing in progress and evolution, only Germany is investing in refugees")

Examples of German language passages with severe toxicity below_10.

52 Valentin Hofmann, Janet Pierrehumbert and Hinrich Schütze. "Dynamic Contextualized Word Embeddings." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL),* August 2021.

the process of identifying extreme speech with seemingly acceptable language usage (for example, as legitimate argument) or complex expression (for example, sarcasm) through greater human inspection and better training data. At the same time, monitoring trigger words and phrases could still be used as a possible scaling strategy (as we have discussed in the earlier section on "Name-calling as seed expressions"). This will enable machine learning models that learn from repeated speech patterns (in the German dataset, for example, "Ausländer raus," "Moslems hassen," "Deutschland zuerst," Table 5) to flag relevant and shifting catchphrases for human inspection. However, training data on such trigger phrases need to be constantly updated and reviewed across languages and beyond common English extreme speech expressions, based on the collaborative process we have been arguing for in this paper.

Table 5

SEVERE TOXICITY	
0,99	"Scheiß Ausländer 😡" ("Fucking foreigners 😡")
0,83	"Ausländer raus" ("Foreigners out")
0,94	"Moslems hassen (unter anderem) Hunde & Co." ("Muslims hate (among other things) dogs & Co.")
0,92	"[...] Deutschland zuerst [...] wir brauchen euch Ausländer nicht # geht dorthin wo ihr hergekommen seid" ("[...] Germany first [...] we don't need you foreigners # go where you came from")

Examples of German language passages with severe toxicity over_75.

Twitter persistence test

In the second test, we randomly sampled 70 passages from each country (a total of 280 passages for four countries) from the whole dataset to check if these passages continued to appear on Twitter. We used the advanced search function[53] and selected either the first five words of the passage or the entire passage if it did not contain five words and entered them in the "this exact phrase" field. From the returned tweets, we manually scrolled the list to find the sampled passages. If we found the original tweet, we took a screenshot and stored the web address of the tweet. The screenshot was then cropped to remove the username of the poster, retaining only the date and the full tweet. If the tweet was not found, we moved to the next one.

This search came with several limitations, most significantly, lack of access to how Twitter filters and organizes its search results. In some instances, advance search queries did not give the result when we pasted the full passages although we had found the same passages in earlier searches. Another limitation is the lack of knowledge on online platforms and messenger services as sources for extreme

53 https://twitter.com/search-advanced

speech passages that fact-checkers gathered for the AI4Dignity project. Not to make the task too cumbersome for fact-checkers, as well as recognizing the fact that different social media platforms are relevant to a different extent across countries, we requested the fact-checkers to locate extreme speech expressions on any social media platform they found relevant in their specific national and linguistic context. The "persistence test" we carried out is therefore not definitive of whether the AI4Dignity's curated extreme speech passages continued to appear on Twitter and, moreover, on other social media platforms. However, we understand the results as a good indication of the limitations of corporate content moderation practices beyond platform specificities, especially since similar extreme speech expressions tend to travel between social media platforms and what persists on a prominent platform such as Twitter is a good signal for its presence and resonance in online discourses more broadly.

Interestingly, only 13 out of 70 passages (18%) in the sampled data from Germany were found on Twitter, and all of them were in the German language. This indicated stronger corporate content moderation systems at work, in the context of far tighter regulatory controls over online speech in the country and greater resources allocated for content moderation.[54] 24 out of 70 sampled passages (34%) from Kenya were spotted on Twitter, but the language breakup of those that continued to appear on the platform revealed that English passages were picked up for moderation far more frequently than those composed in Swahili. Out of 31 tweets written entirely or partly in Swahili, 15 passages continued to appear on Twitter, while only 9 out of 39 English passages could be traced on Twitter. From the Indian dataset, 36 out of 70 sampled passages (51%) continued to appear on Twitter. Half of them were written in Hindi and the other half in English. For Brazil, 63% of the sampled passages (44 out of 70)—all written in Portuguese—continued to be found with advanced search on Twitter.

These findings on regional and language are corroborated by news reports and studies on content moderation on other platforms, especially Facebook.[55] In the case of India, for instance, as The New York Times reported based on the "Facebook Papers": "Of India's 22 officially recognized languages, Facebook said it has trained its A.I. systems on five. (It said it had human reviewers for some others). But in Hindi and Bengali, it still did not have enough data to adequately police the content, and much of the content targeting Muslims 'is never flagged or actioned,' the Facebook report said."[56] Such vast disparities between countries and across the languages bear evidence of unequal and inadequate allocation of resources and lack of processual depth in corporate content moderation and, especially, how hateful expressions in non-Western languages are more likely to escape content filters and other moderation actions.

54 https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html

55 Perrigo, Billy. "Facebook Says Its Removing More Hate than Ever before: But There's a Catch." *Time.* 2019. https://time.com/5739688/facebook-hate-speech-languages; Sablosky, 2021; Murphy and Murgia, 2019; Barrett, Paul M. "Tech - Content Moderation June 2020." NYU Stern Center for Business and Human Rights. 2020. https://bhr.stern.nyu.edu/tech-content-moderation-june-2020.

56 Frenkel and Davey, 2021.

Conclusions: Deep extreme speech and the insignificance of AI

In this paper, we have built on the findings of AI4Dignity, an interdisciplinary and collaborative social action project, to emphasize the need for establishing procedural benchmarks for a people-centric process model for AI-assisted content moderation. An analysis of the curated datasets of the project and two tests around Perspective API scores and the persistence of extreme speech expressions on Twitter has also revealed the limitations in the content moderation practices of big tech, especially the massive gaps in detecting problematic content in peripheralized languages as well as culturally specific use of English in countries beyond the West. It has shown gaps in machine detection even within well-resourced languages such as German, as extreme speech actors find complex and coded expressions to engage in exclusionary discourses against marginalized people. We have also highlighted the challenges of involving community intermediaries in annotation, including the very selection of community members. We have proposed some basic principles for selection—public record of social justice advocacy, linguistic competence, knowledge of vulnerable groups in a national/social context, and experiential knowledge to distinguish between exclusionary extreme content and forms of incivility that seek to challenge repressive power. The project has also sought to build this context sensitivity into label definitions (derogatory, exclusionary and dangerous) and to create a vetting process by involving academic intermediaries with regional expertise and normative commitment to protecting vulnerable and historically disadvantaged communities. However, despite this, the process of annotation comes with the challenges of disagreements over labels and target groups. This difficulty underscores the need for making academic intermediation and professional training more robust by developing clarity about the social consequences of online speech in ways to avoid what could easily slip into false negatives (when hate annotators are not a potential target) and false positives (when annotators are a target).

This analysis returns to our argument for ethical scaling—expensive and exhausting as it may be—to stress the importance of inclusive datasets and a reflexive and iterative process of involving communities in content labeling as critical steps towards modulating and challenging corporate hunger for data.

This analysis returns to our argument for ethical scaling—expensive and exhausting as it may be—to stress the importance of inclusive datasets and a reflexive and iterative process of involving communities in content labeling as critical steps towards modulating and challenging corporate hunger for data. The foregoing analysis of language variation and vast gaps in AI-assisted detection of problematic content in peripheralized languages also highlight the importance of parity in the resources allocated for content moderation.

Ethical scaling, as implemented in AI4Dignity, develops from a conception of AI that does not mirror the inhuman, logical reduction of personhood and the denial of personhood to the marginalized that comprise the ideological edifice of colonial modernity. Instead, through its collaborative process model, it foregrounds what Mhlambi eloquently elaborates as the ethic of "interconnectedness," inspired by the Sub-Saharan African philosophy of ubuntu, in which "Personhood…[is]… extended to all human beings, informed by the awareness that one's personhood is directly connected to the personhood of others."[57] Ethical scaling challenges "AI's quest for a mechanical personhood"[58] and its mooring in the Enlightenment idea of liberty (and the attendant market logic of accumulation) that relies on structures of inequality and dispossession not only to sustain itself but also in its very conception.

Tied to the market logics of data commodification that amplify polarizing content and to a philosophy of logical personhood that denies the principle of cooperation built into "relational personhood,"[59] AI models, when employed to contain harm, also suffer from systemic bias in training data and lack of transparency in AI-assisted decision making. Equally, in terms of practical implementation, AI-based content moderation struggles to keep pace with online speech as the ever-ready means for expressive, suggestive and concealed forms of hate and exclusion that keep evolving. Ethical scaling stresses the need for an experience-near approach to annotation by involving community intermediaries who have a keen understanding of the historical forces of exclusion and the current conjuncture of extreme speech.[60]

Ethical scaling stresses the need for an experience-near approach to annotation by involving community intermediaries who have a keen understanding of the historical forces of exclusion and the current conjuncture of extreme speech.

While highlighting the limitations of AI-based systems on the content side of extreme speech and its embeddedness in the oppressive structures of coloniality and the need for collaborative AI, we conclude this paper by briefly outlining the challenges posed by the distribution side. We suggest that AI is insignificant in addressing intricate networks of distribution that make inroads into the everyday worlds of online users by centering community allegiances in the logics of sharing. This form of distribution, described by Udupa as "deep extreme speech," is built upon tapping community-based trust in ways that content is felt, evaluated and shared not only because of the meanings it might hold but also, more importantly, because it flows through social and community ties that shape the experience of communication as natural, obligatory or simply fun.[61]

Politically partisan content on WhatsApp groups in India provides an illustrative example. Across urban and rural India, WhatsApp is hewn and hammered to create intrusive channels for inflamed rhetoric of different kinds. Political parties have remodeled WhatsApp to serve a heady concoction of top-down "broadcasts" and "organic bottom-up messaging" by installing "party men" within WhatsApp groups of family members, friends, colleagues, neighbors and other trusted communities. "WhatsApp penetration"—defined as the extent to which party people "organically" embed themselves within trusted WhatsApp groups—is seen as a benchmark for a political party's community reach. Local musicians, poets, cinema stars and other "community influencers" have been recruited to develop and expand such "organic" social media networks for party propaganda. Similar trends are observed in Brazil where local influencers, whom one of our participating fact-checkers described as "the guy who is taking a look at the community and telling people what's going on, alerting the community on where the police operation is taking place in the neighborhood, which streets to avoid because of bang, bang [fights between organized crime gangs]," have been drawn into WhatsApp groups and other social media to spread divisive content. According to Brazilian fact-checker Gilberto Scofield, who collaborated on the AI4Dignity project, such "hyperlocal influencers" as human conduits for extreme speech also include pop-

57 Mhlambi, 2020, p 7.
58 Mhlambi, 2020, p 12.
59 Mhlambi, 2020, p 18.
60 Udupa, Sahana, Iginio Gagliardone and Peter Hervik. Digital Hate: *The Global Conjuncture of Extreme Speech.* Bloomington: Indiana University Press, 2021.
61 Udupa, 2021

ular hairdressers who are trusted and admired in the locality. In such circulatory milieus, content develops trustworthiness or at least the efficacy of attention precisely because it attaches to social trust embedded within kin or kin-like networks.

Although automation solutions might help to address the distribution and amplification aspects of extreme speech by tracking influential human "super spreaders," bot activities and trending devices such as hashtags that whip up and organize divisive discussions, AI-based systems are simply incapable of addressing networks of deep extreme speech that lie at the interstices of offline and online, meaning and affect, and technology and the thick contexts of social distribution.

Equally gravely, manipulation of online discourses by repressive and populist regimes around the world have raised the risk of dual use of advanced technologies around AI and their direct instrumentalization for state surveillance. Repressive regimes in the global South, for instance, have begun to copycat strict regulatory mechanisms for social media that developed economies with stable democratic systems have begun to adopt, for authoritarian controls over speech in their own countries.[62] Such risks not only underscore the importance of strict protocols for data protection but also global efforts to monitor AI deployments for targeted surveillance—concerns that have emerged as key topics for the expanding policy and regulatory discussions around AI.[63]

It is critical that AI's promise is tempered with grounded attention to the cultural and social realities of extreme speech distribution and the political dangers of surveillance and manipulation, while also harnessing the potentiality of automation for moderating content through a people-centric process that is transparent, inclusive and responsible, and one that stays close to those that are least protected. ✖

62 Ong, Jonathan Corpus. "Southeast Asia's Disinformation Crisis: Where the State is the Biggest Bad Actor and Regulation is a Bad Word", Items, Social Science Research Council, 2021. https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/southeast-asias-disinformation-crisis-where-the-state-is-the-biggest-bad-actor-and-regulation-is-a-bad-word/

63 Almeida, Patricia Gomes Rêgo de, Carlos Denner dos Santos, and Josivania Silva Farias. "Artificial Intelligence Regulation: A Framework for Governance." *Ethics and Information Technology* 23 (3), 2021, pp. 505–25. https://doi.org/10.1007/s10676-021-09593-z; Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. "What's Next for AI Ethics, Policy, and Governance? A Global Overview." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 153–58. AIES '20. New York, NY, USA: Association for Computing Machinery. 2020. https://doi.org/10.1145/3375627.3375804; High-Level Expert Group on Artificial Intelligence. 2019. "Ethics Guidelines for Trustworthy AI." European Commission.

Acknowledgments

This paper was written during the Joan Shorenstein Fellowship (Fall 2021) that Sahana Udupa received at the Shorenstein Center on Media, Politics and Public Policy, Harvard Kennedy School, and draws on research funded by the European Research Council proof of concept grant (2021-22) under the Horizon 2020 program, grant agreement number 957442. We thank Leah Nann, Miriam Homer, Aleksander Szymanski, Swantje Kastrum and Marc-Anthony Bauer for their excellent research assistance, and all the fact-checkers and partner organizations who have generously given their time and expertise for the project. We thank Anmol Alphonso, Anna-Sophie Barbutev, Anshita Batt, Clara Becker, Boom (Fact-checker), Eva Casper, Fact Crescendo, Mayur Deokar, Aylin Dogan, Govindraj Ethiraj, Fact Crescendo, Nidhi Jacob, Erick Kashara, Thays Lavor, Julia Ley, Chico Marés, Rahul Namboori, Lupa News, Geofrey Omondi, Vinod Rathi, Gilberto Scofield, Cristina Tardáguila and Marita Wehlus for collaborating with us on the project. We are grateful to Joan Donovan for her insightful review of the paper, which helped to clarify several points throughout the essay. Sahana also thanks Joan for her warmth and intellectual support during the fellowship and colleagues at the Center for engaged discussions.

About the Author

Sahana Udupa is professor of media anthropology at LMU Munich and the principal investigator of the AI4Dignity project. Her latest publications include the UN research paper, "Digital Technology and Extreme Speech: Approaches to Counter Online Hate".

Antonis Maronikolakis, Hinrich Schütze and Axel Wisiorek are with the Center for Information and Language Processing, LMU Munich, and research NLP models for different languages.

References

Ali, Syed Mustafa. 2016. "A Brief Introduction to Decolonial Computing." XRDS: *Crossroads, The ACM Magazine for Students* 22 (4): 16–21. https://doi.org/10.1145/2930886.

Almeida, Patricia Gomes Rêgo de, Carlos Denner dos Santos, and Josivania Silva Farias. 2021. "Artificial Intelligence Regulation: A Framework for Governance." *Ethics and Information Technology* 23 (3): 505–25. https://doi.org/10.1007/s10676-021-09593-z.

Barrett, Paul M. 2020. "Tech - Content Moderation June 2020." NYU Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/tech-content-moderation-june-2020.

Becker, Lawrence C., and Charlotte B. Becker. 1992. *A History of Western Ethics.* v. 1540. New York: Garland Publication.

Beller, Jonathan. (2017, October). "The Fourth Determination". *e-flux.* Retrieved from https://www.e- flux.com/journal/85/156818/the-fourth-determination/

Benesch, Susan. 2012. "Dangerous Speech: A Proposal to Prevent Group Violence." New York: World Policy Institute.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code.* Cambridge, UK: Polity Press.

Bonilla, Yarimar, and Jonathan Rosa. 2015. "#Ferguson: Digital Protest, Hashtag Ethnography and the Racial Politics of Social Media in the United States." *American Ethnologist* 42 (1): 4–17. https://doi.org/10.1111/amet.12112.

Burnap, Pete and Matthew L. Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & internet, 7(2), 223-242.

Caplan, Robyn. 2018. "Content or Context Moderation?" *Data & Society.* Data & Society Research Institute. November 14, 2018. https://datasociety.net/library/content-or-context-moderation/.

Davidson, Thomas, Debasmita Bhattacharya and Ingmar Weber. 2019. "Racial Bias in Hate Speech and Abusive Langauge Detection Datasets". *Proceedings of the Third Abusive Langauge Workshop*, pp. 25-35. Florence: Association for Computational Linguistics.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Langauge." ", In International AAAI Conference on Web and Social Media. ArXiv:1703.04009v1 [Cs.CL].

Donovan, Joan. 2020a. "Why Social Media Can't Keep Moderating Content in the Shadows." *MIT Technology Review.* 2020. https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/.

———. 2020b. "Social-Media Companies Must Flatten the Curve of Misinformation," April 14, 2020. https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/d41586-020-01107-z.

Ferreira da Silva, Denis. 2007. *Toward a Global Idea of Race.* Minneapolis: University of Minnesota Press.

Fortuna, Paula, Juan Soler, and Leo Wanner. 2020. "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets". In *Proceedings of the 12th Language Resources and Evaluation Conference,* pp. 6786–6794, Marseille, France. European Language Resources Association.

Founta, Antigoni-Maria, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2-18. "Large scale crowdsourcing and characterization of twitter abusive behavior". In 11th International Conference on Web and Social Media, ICWSM 2018, AAAI Press.

Frenkel, Sheera and Alba, Davey. 2021. "In India, Facebook Struggles to Combat Misinformation and Hate Speech." *The New York Times,* October 23, 2021. https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html.

Ganesh, Bharath. 2018. "The Ungovernability of Digital Hate Culture." *Journal of International Affairs* 71 (2): 30–49.

Gillespie, Tarleton. 2020. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7 (2): 2053951720943234. https://doi.org/10.1177/2053951720943234.

Gröndahl, Tommi, Luca Pajola, Mika Juuti, Mauro Contin, and N. Asokan. 2018. "All You Need Is 'Love': Evading Hate Speech Detection." ArXiv:1808.09115v3 [Cs.CL].

High-Level Expert Group on Artificial Intelligence. 2019. "Ethics Guidelines for Trustworthy AI." European Commission.

Hofmann, Valentin, Janet B. Pierrehumbert, and Hinrich Schütze. 2021. "Dynamic Contextualized Word Embeddings." ArXiv:2010.12684 [Cs], June. http://arxiv.org/abs/2010.12684.

Klonick, Kate. 2017. "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review* 131: 73.

Lee, Ronan. 2019. "Extreme Speech| Extreme Speech in Myanmar: The Role of State Media in the Rohingya Forced Migration Crisis." *International Journal of Communication* 13 (0): 22.

Mhlambi, Sabelo. 2020. "From Rationality to Relationality." *Carr Center for Human Rights Policy Harvard Kennedy School,* Carr Center Discussion Paper, No. 009: 31.

Mignolo, Walter D. 2007. "Introdution: Coloniality of Power and Decolonial Thinking." *Cultural Studies* 21 (2–3): 155–67.

Morozov, Evgeny. 2011. *The Net Delusion: The Dark Side of Internet Freedom.* New York: Public Affairs.

Murphy, Hannah, and Madhumita Murgia. 2019. "Can Facebook Really Rely on Artificial Intelligence to Spot Abuse?" *FT.Com,* November. https://www.proquest.com/docview/2313105901/citation/D4DBCB03EAC348C7PQ/1.

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

Ong, Jonathan Corpus. 2021. "Southeast Asia's Disinformation Crisis: Where the State is the Biggest Bad Actor and Regulation is a Bad Word." *Items,* Social Science Research Council. https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/southeast-asias-disinformation-crisis-where-the-state-is-the-biggest-bad-actor-and-regulation-is-a-bad-word/

Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. "Multilingual and multi-aspect hate speech analysis". 2019. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Perrigo, Billy. 2019. "Facebook Says Its Removing More Hate than Ever before: But There's a Catch." *Time.* https://time.com/5739688/facebook-hate-speech-languages/.

Quijano, Anibal. 2007. "Coloniality and Modernity/Rationality." *Cultural Studies* 21 (2): 168–78.

Ross, Björn, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, Michael Wojatzki. 2017. "Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis". ArXiv:1701.08118 [cs.CL].

Sablosky, Jeffrey. 2021. "Dangerous Organizations: Facebook's Content Moderation Decisions and Ethnic Visibility in Myanmar." *Media, Culture & Society* 43 (6): 1017–42. https://doi.org/10.1177/0163443720987751.

Saleem, Haji Mohammed, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2017. "A Web of Hate: Tackling Hate Speech in Online Social Spaces." ArXiv Preprint ArXiv:1709.10159.

Sap, Maarten, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. "Social bias frames: Reasoning about social and power implications of language." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. "The Risk of Racial Bias in Hate Speech Detection." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 1668–78. Florence, Italy.

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. "What's Next for AI Ethics, Policy, and Governance? A Global Overview." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 153–58. AIES '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375804.

Seaver, Nick. 2021. "Care and Scale: Decorrelative Ethics in Algorithmic Recommendation." *Cultural Anthropology* 36 (3): 509–37. https://doi.org/10.14506/ca36.3.11.

Slack, Jennifer. 2006. Communication as articulation. In: Shepherd G, St. John J and Striphas T (eds) Communication as . . . : Perspectives on Theory. Thousand Oaks: SAGE Publications, pp.223–231.

Swamy, Steve Durairaj, Anupam Jamatia, and Björn Gambäck. 2019. "Studying generalisability across abusive language detection datasets." In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL).

Stern.de. 2019. "Bremer Landgericht Gibt Facebook Recht: Begriff 'Goldstück' Kann Hetze Sein." June 21, 2019. https://www.stern.de/digital/bremer-landgericht-gibt-facebook-recht--begriff--goldstueck--kann-hetze-sein-8763618.html.

Steyn, Melissa, and William Mpofu, eds. 2021. *Decolonising the Human: Reflections from Africa on Difference and Oppression.* Wits University Press. https://doi.org/10.18772/22021036512.

Thirangama, Sharika, Tobias Kelly, and Carlos Forment. 2018. "Introduction: Whose Civility?" *Anthropological Theory* 18 (2–3): 153–74. https://doi.org/10.1177/1463499618780870.

Udupa, Sahana. 2017. "Gaali Cultures: The Politics of Abusive Exchange on Social Media." *New Media and Society* 20 (4): 1506–22. https://doi.org/10.1177/1461444817698776.

———. 2020. "Decoloniality and Extreme Speech." In Media Anthropology Network E-Seminar. European Association of Social Anthropologists. https://www.easaonline.org/downloads/networks/media/65p.pdf.

———. 2021. "Digital Technology and Extreme Speech: Approaches to Counter Online Hate." In *United Nations Digital Transformation Strategy.* Vol. April. New York: United Nations Department of Peace Operations. https://doi.org/10.5282/ubm/epub.77473.

Udupa, Sahana, Iginio Gagliardone and Peter Hervik. 2021. *Digital Hate: The Global Conjuncture of Extreme Speech.* Bloomington: Indiana University Press.

Warner, W., and J. Hirschberg. 2012. "Detecting Hate Speech on the World Wide Web." In *Proceedings of the Second Workshop on Language in Social Media,* 19–26. Association for Computational Linguistics. https://www.aclweb.org/anthology/W12-2103.

Waseem, Zeerak, Dirk Hovy. 2016. "Hateful symbols or hateful people? predictive features for hatespeech detection on Twitter". In Proceedings of the NAACL Student Research Workshop.

Wynter, Sylvia. 2003. "Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation—An Argument." *CR: The New Centennial Review* 3 (3): 257–337.