

# Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence

Big Data & Society  
January–June: 1–15  
© The Author(s) 2023  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20539517231172424  
journals.sagepub.com/home/bds



Sahana Udupa<sup>1,2</sup> , Antonis Maronikolakis<sup>2</sup>   
and Axel Wisiosek<sup>2</sup>

## Abstract

In this article, we present new empirical evidence to demonstrate the severe limitations of existing machine learning content moderation methods to keep pace with, let alone stay ahead of, hateful language online. Building on the collaborative coding project “AI4Dignity” we outline the ambiguities and complexities of annotating problematic text in AI-assisted moderation systems. We diagnose the 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”. Presenting a decolonial critique of the “human vs machine” conundrum and drawing attention to the structuring effects of coloniality on extreme speech, we propose “ethical scaling” to highlight moderation process as political praxis. As a normative framework for platform governance, ethical scaling calls for a transparent, reflexive, and replicable process of iteration for content moderation with community participation and global parity, which should evolve in conjunction with addressing algorithmic amplification of divisive content and resource allocation for content moderation.

## Keywords

AI, extreme speech, ethical scaling, decoloniality, social media content moderation, ethnography and algorithm auditing

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. In countries where democratic safeguards are crumbling, the extractive attention economy of digital communication has accelerated a dangerous interweaving of corporate profit and state repression, while regulatory pressure has also been mounting globally to bring greater public accountability and transparency in tech operations (George, 2016; Hervik 2019; Lee, 2020; Ong, 2021; Sablosky, 2021; Wasserman and Madrid-Morales, 2022).

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, artificial intelligence (AI) has emerged as a shared imaginary of technological solutionism. Ambitious in its scope and opaque in terms of the technical steps that lead up to its constitution, AI has gripped the imagination of corporate minds as a technological potentiality that can help them to confront a deluge of revelations of the harms their platforms have

helped amplify (Gorwa et al., 2020; Matsakis, 2018; Murphy and Murgia, 2019).

In this article, we present new empirical evidence to demonstrate the severe difficulties for existing machine learning content moderation methods to keep pace with, let alone stay ahead of, hateful language online. We present a set of findings from the project, “AI4Dignity” (2021–22), which involved facilitated dialogue between independent fact-checkers, ethnographers and AI developers to gather and annotate online vitriolic expressions in Brazil, Germany, India, and Kenya. We examine the limitations in the content moderation practices of global social media companies in terms of linguistic and contextual knowledge by comparing datasets from AI4Dignity with

<sup>1</sup>University of Munich, Munchen, Germany

<sup>2</sup>Ludwig-Maximilians-Universitaet Muenchen (LMU Munich), Munchen, Germany

## Corresponding author:

Sahana Udupa, Professor of Media Anthropology, LMU Munich, Oettingenstrasse 67, Munich 80538, Germany.  
Email: Sahana.udupa@lmu.de



Perspective API's toxicity scores developed by Google as an illustrative case. We layer these findings with ethnographic observations of our interactions with fact-checkers during different stages of the project, to develop a methodology of combining ethnography with "algorithm auditing" (Sandvig et al., 2014) that can articulate a critique and contribute to policy in an interrelated way. The combined evidence of ethnographic and algorithm auditing methods shows the importance of community involvement, and furthermore, 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.

The empirical evidence about the difficulties of annotation underscores the need for conceptual rethinking. We discuss the Perspective API test results and our experiences in the project to demonstrate that community involvement in annotation is not merely a technical necessity for developing effective training datasets, but it raises questions around some of the foundational assumptions and aspirations around AI. Presenting a decolonial critique of AI-assisted moderation and the liberal framing of the "human versus machine" conundrum that animates AI imaginaries, we propose the framework of "ethical scaling" to highlight moderation process as political praxis. We conclude by discussing the limitations and risks of automation, especially how a focus on message level intervention constitutes just a small part of the larger problem concerning the social distribution, multimedia amplification and broad-ranging impacts of extreme speech (Udupa et al. 2021). Our critique of AI-assisted content moderation and ethical scaling as a policy intervention are therefore an attempt at bringing accountability and ethnographic nuance to an important, if not a decisive, element in extreme speech ecologies, and to temper the claims around corporate AI's capacities in content moderation and technological imaginations that underlie them.

## AI in content moderation

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 policy team" (characterized by large platforms such as Google and Facebook)

(2018: 16). Caplan observes that "industrial models prioritize consistency and artisanal models prioritize context" (16). 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.

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 (Gillespie, 2020)—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.

While the capacities of AI language models to perform different communicative tasks have been expanding, there are vast challenges in AI-assisted moderation of hateful content online, as companies and natural language processing (NLP) researchers also admit. 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" (Murphy and Murgia, 2019). 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. Hate groups have managed to escape keyword-based machine detection through clever combinations of words, misspellings, satire, changing syntax and coded language (Burnap and Williams, 2015; Gröndahl et al., 2018; Warner and Hirschberg, 2012). The dynamic nature of online hateful speech—where hateful expressions keep changing—adds to the complexity. As a fact-checker participating in AI4Dignity commented, they are swimming against "clever ways [that abusers use] to circumvent the hate speech module".

Several initiatives have tried to address these limitations by incorporating users' experiences and opinion, such as crowdsourcing models experimented by Google's Perspective API and Twitter's Birdwatch. Such efforts have sought to leverage 'crowd intelligence' but the resulting machine learning models, while offering some

promising results, are prone to false positives as well as racial bias (Sap et al., 2019). More critically, crowdsourced models have channelized the onus of detection onto an undefined entity called ‘crowd’, seeking to coopt the Internet’s promised openness to evade regulatory and social consequences of gross inadequacies in corporate efforts and investments in moderating problematic content. Third-party moderation reveals a similar structure of devalued work and undercompensated labor (Roberts 2019).

## AI4Dignity and the problem of annotation

To test the efficacy of existing corporate content moderation practices as well as to develop inclusive training datasets and foreground the interpretative and emotional labor of annotation work that is needed in creating them, we designed the project AI4Dignity, building on multiyear ethnographic research (2013-ongoing) on online vitriolic cultures and their on-ground manifestations. In this project, collaborating factcheckers dialogued with ethnographers and NLP researchers to identify problematic content on social media and finalize the definitions of labels for annotating types of problematic content and target groups. These labeled passages were used for establishing baselines and novel tasks for large pretrained models and traditional machine learning models (multilingual mBERT, XLM-R and monolingual langBERT models for Hindi, Swahili and Portuguese). Subsequently, a user interface was created to input the text and receive outputs by the model for the extreme speech type and target group for each country as a proof-of-concept.

The main point of departure for the project has been the framework of “extreme speech” and its interventions in highlighting the limits of “hate speech” as a discourse (Udupa 2018; Udupa and Pohjonen 2019). While recognizing the significance of “hate speech” as a regulatory concept built on longer legal debates over speech restrictions (Nockleby, 2000; Warner & Hirschberg, 2012), the extreme speech framework has simultaneously pointed out that the discourse of hate speech predefines the effects of hate speech as negative and damaging, and its regulatory rationale is, thus, of control and containment. The state is the largest actor in this effort, but internet intermediaries also increasingly monitor and restrict speech on their platforms based on different articulations of harm and international conventions on hate speech.

As it jostles between state regulation, the capitalist market and political fields, hate speech has become what Brubaker and Cooper would describe as a thick concept with a “tangle of meanings” and an evaluative load (Brubaker and Cooper, 2000, p. 14). In everyday conversational contexts, “hate speech” is often used as a charge or an accusation that closes off, rather than opens up, avenues for change and dialogue (Boromisza-Habashi, 2013). Highlighting the risks of contextual flattening and rhetorical

impacts of hate speech to obstruct the very goals of mitigation that it sets out to accomplish, the framework of extreme speech has advanced ethnographic attention to cultural variation in speech and evolving practices in online communities, and historical awareness and emic perspectives to assess their valence, implications and possible countermeasures. Importantly, it argues that the current conjuncture of vitriol is not a sudden crisis triggered by social media expansion but is underwritten by longer processes of racialization and coloniality (more about this point in the next sections). Ethnographic sensibility to historical and local inflections emphasized in the extreme speech framework sets the benchmark for developing procedures in which community inputs become the cornerstone for operationalizing the conception of online harms and organizing the practicalities of moderation.

In turn, the very first step has been to identify community intermediaries, and the project partnered with independent factcheckers as a key stakeholder community because of their professional training in handling contentious content. For fact-checkers, the collaboration also offered ways to foreground their own grievances as a target community of extreme speech. 13 factcheckers (8 female and 5 male) from four countries—Brazil, Germany, India, and Kenya—with fluency in English and a major local language participated in the first phase of the project backed by their senior colleagues who participated in project discussions at various times (2020–2021).

The next step was to develop the labels for annotation. The ethnographers on the team discussed the definitional scope of different categories with the factcheckers and NLP researchers before finalizing the three labels (Table 1). In these discussions, factcheckers’ requests to include more fine-grained labels had to be balanced against NLP researchers’ preference for a smaller list for better model performance, given the limitations of the size of data gathered. The final list of three labels represented a gradation of severity—from derogatory to exclusionary to dangerous forms of speech—with corresponding recommendations for regulatory actions.

After agreeing upon the definitions of the three types of problematic speech, fact-checkers were requested to gather and label online passages. 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. The project adopted a platform agonistic approach, allowing platform selection to evolve through on-ground expertise of factcheckers. In Kenya, fact-checkers gathered extreme passages largely from WhatsApp, Twitter, and Facebook; Indian fact-checkers 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.

**Table 1.** Definitions of types of extreme speech and recommended moderation actions.

Type of extreme speech	Definition	Recommended actions
Derogatory extreme speech	Expressions that do not conform to accepted norms of civility within specific regional/local/national contexts and target persons/groups based on racialized categories or protected characteristics (caste, ethnicity, gender, language group, national origin, religious affiliation, sexual orientation) as well as other groups holding power (state, media, politicians) (Uduba, 2018). It includes derogatory expressions not only about people but also about abstract categories or institutions that they identify targeted groups with. It includes varieties of expressions that are considered within specific social-cultural-political contexts as “the irritating, the contentious, the eccentric, the heretical, the unwelcome, and the provocative, as long as such speech...[does]...not tend to provoke violence”.	Closer inspection, and downranking, counter speech, monitoring, redirection, and awareness raising but not necessarily removal of content.
Exclusionary extreme speech	Expressions that call for or imply exclusion of historically disadvantaged and vulnerable people/groups from the “in-group” based on caste, ethnicity, gender, language group, national origin, racialized categories, religious affiliation, and/or sexual orientation. These expressions incite discrimination, abhorrence and delegitimization of targeted groups. The label does not apply to abstract ideas, ideologies, or institutions, except when there are reasonable grounds to believe that attacks against ideas/ideologies/institutions amount to a call for or imply exclusion of vulnerable groups associated with these categories. For example, if attacking a particular religion in a specific context has a reasonable chance to incite hatred and exclusion of people who practice this religion, such expressions would fall under ‘exclusionary extreme speech’. In terms of exclusionary extreme speech, the analysis builds on existing definitional standards around hate speech set up by the United Nations (2020). <sup>8</sup>	Closer inspection and possible removal.
Dangerous speech	Dangerous speech refers to expressions that have a reasonable chance to trigger/catalyze harm and violence against target groups (including ostracism, segregation, deportation, and genocide) (Benesch, 2012).	Immediate removal

**Table 2.** Language distribution across extreme speech types and targets.

Language	Brazil	Germany	India	Kenya
English	0	6	1056	2695
Local Language	5109	4922	2778	404
English and local language	0	71	1174	2081

In all the cases, factcheckers sourced the passages from popular “surface” platforms, thereby bringing widely available online content for training and testing the models. A total of 20,297 passages were obtained from this collaboration covering English, Swahili, Brazilian Portuguese, German, and Hindi (see Table 2).

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. 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. Following this, and at each step of the annotation process, we clarified the categories with new rounds of discussions with factcheckers, calibrated the target group list, and took forward major disagreements for further discussion.

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 in four country teams 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. Counterathon allowed ethnographers, NLP researchers and factcheckers to discuss the reliability of labels in each country case, explore different strategies to gather better statistical signals from curated expressions, and foreground the difficulties that fact-checkers faced both in terms of analytical clarity around the labels and emotional labor involved in gathering and sifting problematic content.

The project revealed distinct challenges facing AI-assisted moderation and the process of establishing human supervision. At the outset, selecting annotators was a daunting challenge. Although factcheckers bring contextual knowledge, their autonomy cannot be taken for granted. We built on international standards for fact-checking by collaborating with participants affiliated to the International Fact-Checkers' Network or involved in civil society peace activism and ensured that they were not employed directly by commercial social media companies or political parties.

The process of defining the labels and classification of gathered passages during the project proved to be intensely laborious and dotted with uncertainty and contradiction. Such 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. Although there was consensus that all selected passages were extreme speech cases, instances of uncertainty about the distinction between the three categories and target groups were plentiful, and the Krippendorff (2003) intercoder agreement score (alpha) between two fact-checkers from the same country averaged 0.24 (similar to other hate speech detection projects).<sup>1</sup>

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" is not a relevant category since Kikuyu and Kalenjin ethnic groups around whom a large proportion of extreme speech circulates are actually large ethnic groups. Based on this discussion, we included "large ethnic groups" as a target community to refer to extreme speech instances between the Kikuyu and Kalenjin groups and capture the specific demographic and power constellation defining Kenyan politics. Fact-checkers from Germany pointed out that "refugees" were missing from the list, since the general term "immigrants"—some of whom are

welcomed and desired at least for economic reasons—are different from refugees and asylum seekers who are derided as unwanted.

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 in Germany wondered what to make of politicians who are women or who have a migration background. "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." Citing the expression "Sugoi thief," he pointed out that in such expressions, politicians become a synecdoche for an entire target community. Fact-checkers from India highlighted the difficulty of placing Dalit politicians and Muslim politicians 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.

As these exchanges bear out, delineating target groups is a crucial step in the annotation process, requiring community calibration at regular intervals—the specific rhythms of which should evolve based on ground realities, especially critical transitions such as regime change. Grounded focus on target groups as a core principle for collaborative annotation builds on a particular point of emphasis in the "extreme speech" framework's critique of the hate speech discourse. As a form of power, the discourse of hate speech is inextricably tied to the state and its political economies of violence. Historically, it emerged from projects of civility that coincided with (and partly constituted) the state's monopolization of violence (Thiranagama et al., 2018). Under these conditions, the pressure to speak polite language has been an act of domination—moral injunctions linked to assertions of privilege. Civility, thus, is an "effect of political recognition and of a responsive structure of authority" (Mitchell, 2018: 217). In other words, the implications of incivility—or the extremeness of speech more broadly—cannot be comprehended without analyzing forms of recognition and responsiveness to demands that are available to diverse groups (Udupa et al., 2021).

Importantly, alongside drawing correspondence between context sensitive lists of target groups and extreme expressions in ways to account for the ambiguities and historical lineages of "extremeness," 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. During the project, a leading factchecker

and LGBTQI + activist in Brazil told us how hate speakers “don’t ever use a sentence like, ‘This kind of people should die.’ Never”. 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, he went on to elaborate, “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.” Discussions such as this prompted us to pry open the meanings of some complex expressions, as factcheckers 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 (see Tables 7–10).

Are existing machine learning models and content moderation systems equipped to detect such subtleties identified through collaborative dialogue, iteration and historically sensitive analytical frameworks? Using the project dataset (Tables 2–4), we carried out qualitative tests of an initiative by Google. 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.

### Perspective API test

For the first test, we ran relevant passages in the project database on Perspective API<sup>2</sup>—a machine learning model developed by Jigsaw to assign toxicity scores to texts (Table 5

and Figure 1), as part of an “attempt to package the identification of ‘toxic’ speech into a service that can be used by websites to help moderate their forums and comment sections” (Rieder and Skop, 2021, 2). Methodologically, the platform agnostic multilingual dataset of AI4Dignity is congruent with Perspective API’s approach as a tool focused on applications in different languages across platforms and media. We obtained an API key for Perspective<sup>3</sup> to run the test. Based on existing language support of the Perspective system, data for English (3761 passages from all the countries), German (4945 passages), Portuguese (5245), English/German (69), Hindi (2775), and Hindi/English (1162) 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 the model to predict the language from the text. We computed six attributes that Perspective identifies as toxicity, severe toxicity, identity attack, threat, profanity, and insult.<sup>4</sup> 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 attributes of Perspective API 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.

Table 5 presents the breakdown of the score distribution for different attributes in AI4Dignity and Perspective. Derogatory passages in English across all the countries received a score of 48 (represented as 0.48 in the table) for toxicity and 47 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

**Table 3.** General distribution of labeled passages.

Extreme speech	Brazil	Germany	India	Kenya
Derogatory extreme speech	4774	2643	2226	3386
Exclusionary extreme speech	115	2340	1421	966
Dangerous speech	220	16	1361	828

**Table 4.** Total number (n) and percentage (%) of messages directed at target groups.

Target groups	Brazil		Germany		India		Kenya		Total	
	n	%	n	%	n	%	n	%	n	%
Religious minorities	16	0.5	1269	23.8	3522	64.7	111	2.2	4918	25.4
Any other	1066	30.5	34	0.6	356	6.5	1534	30.3	2990	15.5
Immigrants	28	0.8	2355	44.1	109	2	292	5.8	2784	14.3
Women	1479	42.3	367	6.9	418	7.7	396	7.8	2660	13.8
Large ethnic groups	0	0	0	0	0	0	2273	44.8	2273	11.8
Sexual minorities	674	19.3	347	6.5	89	1.6	80	1.6	1190	6.2
Historically oppressed caste groups	45	1.3	1	0	853	15.7	33	0.7	932	4.8
Racialized groups	78	2.2	527	9.8	3	0.1	80	1.6	688	3.6
Ethnic minorities	58	1.7	430	8.1	89	1.6	77	1.5	654	3.4
Indigenous Groups	50	1.4	6	0.1	5	0.1	195	3.8	256	1.3

**Table 5.** Perspective scores for AIDignity passages across all types of extreme speech.

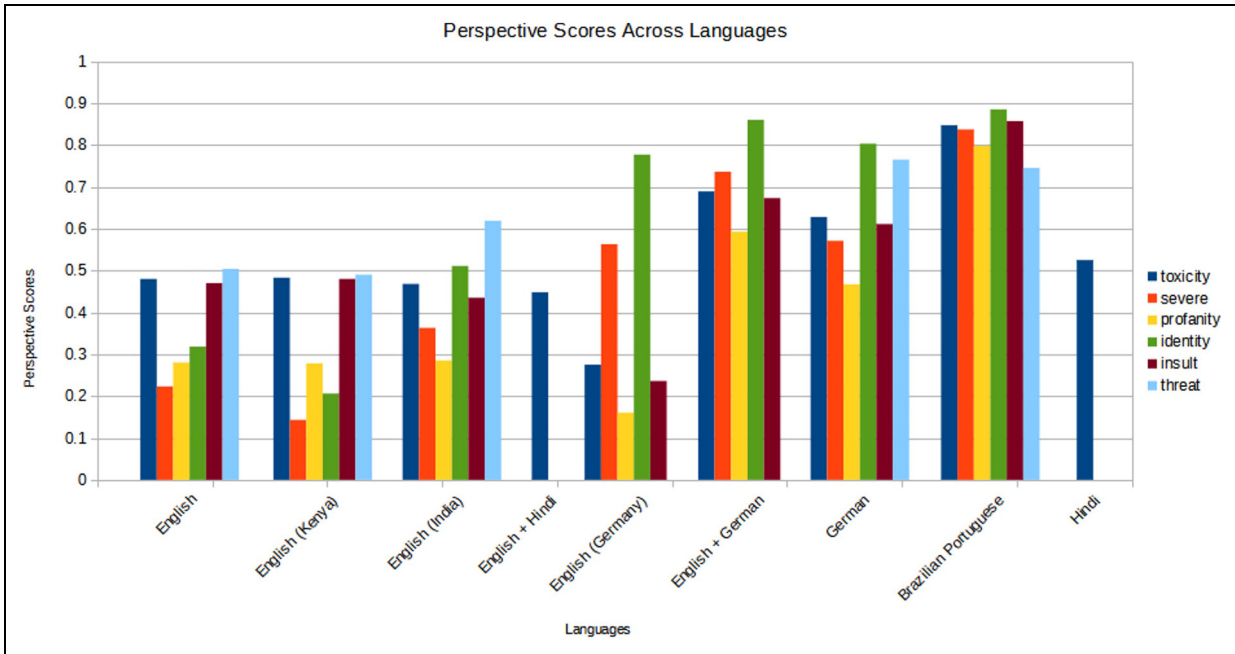
	Toxicity	Severe Toxicity	Profanity	Identity	Insult	Threat	Total
eng_all	0.43	0.28	0.24	0.32	0.41	0.34	3490
eng_der	0.48	0.31	0.28	0.33	0.47	0.32	1972
eng_exc	0.35	0.22	0.19	0.32	0.33	0.28	916
eng_dan	0.39	0.28	0.19	0.30	0.33	0.50	602
eng_kenya_all	0.42	0.26	0.23	0.27	0.40	0.32	2680
eng_kenya_der	0.48	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

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). These results signal disparities in the model performance for English, especially in assessing culturally shaped English usage in countries in the global South (here, India and Kenya) in extreme speech contexts.

In comparison, the model performed better for German-only and Portuguese-only across all the three categories. German derogatory passages received a score of 63 for toxicity and 61 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.

### *Culturally coded expressions and complex statements*

To examine one more aspect of the Perspective 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



**Figure 1.** Perspective scores for corresponding extreme speech types across languages.

**Table 6.** Trigger words in the German dataset.

	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)

equipped to detect problematic content that do not contain such expressions but composed entirely in languages other than English and with local cultural references. We selected German and Indian datasets for closer inspection. In the German dataset, mixed language passages that received a high score on Perspective had English expressions such as “shithole countries,” “black lies matter,” “in cold blood,” “new world order” and “wake up.” In terms of single words, a majority of passages (92%) containing the most frequent English trigger word “shithole” (39 passages) have a high score for severe toxicity (over\_75) and none of them are classified as clean (below\_10; Table 3).<sup>5</sup> Interestingly, similar results were obtained in an exemplary analysis for two German trigger words (“Homos” [‘homosexuals’], “Scheiss” [shit]; see Table 6). 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). Beyond the single words, more obvious exclusionary phrases like “Ausländer raus” [foreigners out] (which received a high score 0.83 for severe toxicity) 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), which indicates that these words are not triggering the model on their own. Further qualitative analysis of these false negatives shows that they represent more complex, concealed, and argumentative forms of exclusionary discourse practices (Table 7). The results reveal gaps in machine detection models even within well-resourced language contexts such as Germany.



**Table 7.** Examples of German language passages with severe toxicity below .10.

SEVERE TOXICITY	Examples
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”)

**Table 8.** Examples of passages in English that scored below 0.1 on Perspective API.

TOXICITY	SEVERE TOXICITY	
0.06	0.03	“It Was Nehru Who I st Found #GreenFungus!” [accompanying image shows Pakistan’s flag]
0.07	0.03	Coronavirus in India: 600-case spike takes tally over 3000, govt says 30% of cases linked to Tablighi Jamaat. #BioWeapon #BioJihad #Tableeghijamaat
0.08	0.04	“Remember Tienanmen Square, Beijing in 1988? And how Deng Xiaoping handled it? Perhaps there is a lesson there on how to handle the engineered disturbances of NE Delhi! I’m sure all comrades will agree!”
0.09	0.06	“We are 80%, you are 18%. If we hit back, imagine what will happen to you”
0.05	0.02	LGBT; It wears down and corrupts the family institution, which is the basic building block of society, disrupts the structure of the society, and prevents the growth of physically and mentally healthy individuals and generations.

We carried out further qualitative analysis with the India dataset by examining the list of extreme speech passages from India that scored below 0.1 for toxicity and severe toxicity on Perspective. Tables 8, 9 and 10 show expressions in English, Hindi/English, and Hindi with no common hateful words but contain implicit meanings, culturally coded expressions including invocations of mythological figures and narratives, indirect dog whistles, and ingroup idioms to deride Muslim minorities and Dalits. A unique feature is the range of hashtags and namecalling with satire and wordplay that have escaped the model in the cited instances.

## A decolonial critique

The above test shows limitations in detecting problematic content even when it circulates on widely used platforms, but in relation to a particular corporate initiative. The mandate of this initiative is situated within the political economy of open technical standards that can safeguard the “web-focused business model” of the parent company and negotiations with news providers to organize a complex system of dependencies on its services (Rieder and Skop 2021, 5). As such, it does not represent the entire spectrum of AI-assisted moderation systems and language models more broadly, since the goals,

design, use cases and articulations of problematic content vary considerably across companies, and between state and corporate actors. However, the unique architecture of Perspective—“a cooperative, multi-polar model...[involving] ... a degree of openness, transparency and adaptability” (2)—invites attention to gaps in machine learning even when well-funded systems curated by corporate interests are kept “open” and “cooperative” at least in the limited sense of open source code, technical interface possibilities, publication of label definitions and public availability of model application with a process to link user feedback with model adaptation and optimization. Such limitations, which are compounded by the fact that Perspective has kept the core training dataset as “proprietary,” 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 political pressure increases and resources for content moderation, including human supervision, expand. However, the vast complications of annotation that AI4Dignity has highlighted—at the level of classification (label definitions and identification of target groups), content (interpretation of meanings) as well as process (methods and frequency of community involvement)—stress the point that some fundamental ethical and political

**Table 9.** Examples of passages in mixed registers of English and Hindi that scored below 0.1 on Perspective API.

TOXICITY	Examples
0.05	"The problem is with our constitution, in 1976 by 42nd amendment act word "Secular" was added fortunately it was "not the part of the constitution" in 1947. when we can remove art 370, then why not this word." #मुस्लिमो_का_संपूर्ण_बहिष्कार [#complete_ban_on_Muslims]
0.02	"New Examination Pattern in India 1. General students : Ans ALL questions 2. OBC : Write ANY question. 3. SC : Read ONLY questions. 4. ST : Thanks for coming to exam. #SCSTAct_गलत_है [is wrong]"
0.07	"Arjuna : Who are these people asking Azadi? What is Azadi? Krishna : Arjuna, Ask not what they are asking for and why they are asking. Just Pick up your bow and arrows and give them what they want. #देशद्रोहियों_शर्म_करो [Shame on traitors]"
0.07	"#IPL2021 Lot of #Peacefuls in IPL teams. Very dangerous trend. Most of the teams have 3 to 4 Pislamists players. #Cricketjihad?"
0.06	"I support #GoliMaaro! You? [Shoot them – refers to a slogan targeting Muslims]"

**Table 10.** Examples of passages in Hindi that scored below 0.1 on Perspective API.

TOXICITY	Passage
0.03	"(लड़का + लड़की) x ववाह = हम दो हमारे दो। {बकरा + (बुरखाx4) + बकरी} x नकिह = हम पाँच हमारे पैतीस शुअर!" ("Boy + Girl x Marriage = We two, ours two {Male goat + (Burqa x 4) + Female goat} x Nikah = We five, ours thirty-five pigs!")
0.04	"#भारतवर्ष को #यवनो और #म्लेच्छों से मुक्त करो!! ("Free #Bharatvarsh [India] from #Yavanon [derogatory term for Muslims] and #Mlechhas [derogatory term for the Dalits])"
0.00	"Bharat mein Bharat Mata ki Jai karne wala rahega... aur jo nahi bolega, jo Bharat ka virodh karega, sanvidhanik vyavasthaon ka sammaan nahi karega, niraadar karega, baar baar karega, unke baare mei nischit roop se vichar karne ki aavyashakta hai" ("In India those who raise salutary slogans to Mother Bharat will remain. Those who do not raise the slogan, those who oppose Bharat, disrespect constitutional systems, and do it again again—it is important to decisively think about them")
0.05	"Are bhimte kbhi to ghr se bhr nikl kr chittorgarh aakr dikha neele se pila ho jayega" ("Arey Bhimte [derogatory term for the Dalits] come out of the house, appear in Chittorgarh, from blue you will turn yellow")
0.10	"मुल्ते से फल सब्जी मत खरीदो।" ("Don't buy fruits and vegetables from Mulley [derogatory term for Muslims]")

issues undergird the problem of content moderation and AI, which require critical insight beyond individual empirical cases. These complications prompt an inquiry into abstract normative constructions such as "the human" that guide and temper the development of AI systems, and how such abstractions point to a larger structural problem linked to Euro-modern thinking and systems of oppression it helped create.

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 the 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 (Becker and Becker, 1992) that needs protection from the "machine" (Zuboff 2019).

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" (2020, 1). 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" (1).

The liberal weight behind the concept of the human elides its troubled lineage in European colonial modernity that racially classified human, subhuman and nonhuman (Wynter, 2003), institutionalizing this distinction within the structures of the modern nation-state (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). 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.

While a decolonial critique of AI encompasses broad ranging criticisms to ask “...why AI as a field *depends on* and was made possible by the logics of race and coloniality” (Adams 2021, 179), for online content moderation and AI—the focus of this article—attention to colonial history raises three 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 language 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. 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 and antagonism that “mingle” with if not are “invented” by the colonial encounter (Thiranagama et al., 2018: 165). In Germany, anti-immigrant and Islamophobic messages constitute the largest proportion of extreme speech in our dataset (Table 4, see also Tables 6–7 for examples). In India, religious minorities are the primary targets; in Brazil, women and sexual minorities are the most frequently targeted groups, and in Kenya, a major part of extreme speech surrounds exchanges between large ethnic groups (Table 4, see also Tables 8–10). Importantly, extreme speech content is also driven by the capitalist logics of coloniality now manifest, among other things, as data monetization, albeit in different degrees—for instance between stricter platform regulations in Europe as opposed to the First Amendment protections and laissez faire approach in the US. Therefore, while each country case is distinct in its history, media systems (Hallin and Mancini 2012) and current constellation of power, an overarching pattern is how coloniality’s nation-state structure, racialization and market rationality have prepared the ground for online extreme speech and its harms to proliferate.

At the same time, epistemologies of coloniality limit 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. This thinking is conceptually unprepared to grasp the responsibility of community participation in the design and imagination of the machine. Put succinctly, both the problem (extreme speech) and the proposed solution (automation) are intrinsically linked to Euro-modern thinking.

Even more, the dehumanizing distinction of coloniality tacitly rationalizes the uneven allocation of corporate resources for content moderation across different geographies and language communities. 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” (Frenkel and Alba, 2021). 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 skewed structures of content moderation that global social media corporations have instituted. Perspective API—the corporate initiative examined in this study—is currently limited to evaluating passages in English, French, German, Italian, Portuguese, Russian and Spanish for different attributes and Hindi only for the “toxicity” attribute. Studies have revealed similar disparities across countries, platforms and languages in terms of machine learning capabilities (Barrett, 2020; Murphy and Murgia, 2019; Perrigo, 2019; Sablosky, 2021). For instance, citing the hate campaign by the Assamese-speaking Hindu majority against the Bengali-speaking Muslim minority in Assam in eastern India, Perrigo (2019) shows how messages that described Bengali Muslims as “parasites,” “rats,” and “rapists” and viewed at least 5.4 million times were not picked up by Facebook, because the company did not have an algorithm to detect hate speech in Assamese. These findings bear evidence of unequal and inadequate allocation of resources and how hateful expressions in non-Western languages as well as specific ways of using English in non-Western contexts (see Table 5, Figure 1) are more likely to escape content filters and other moderation actions. Such disparities attest to what Denis Ferreira da Silva (2007) 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.

To summarize, the liberal-modern epistemology as well as racial, market and nation-state relations of coloniality significantly shape the 1) content and targets of extreme speech 2) limitations in the imagination of technology and 3) disparities in content moderation. Both as a technical problem of contextualization and a political problem that obscures colonial classification and its structuring effects on content moderation, the dichotomous conception of “human vs machine” glosses over pertinent issues around who should be involved in the process of annotation and moderation beyond the reified category of the “human”, 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”. Whereas companies acknowledge that human annotators are necessary (Murphy and Murgia, 2019), 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 (2020) offers some important clarifications around scale and size, and why they should not be collapsed to mean the same. Building on Jennifer Slack’s (2006) 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” (Gillespie, 2020: 2). 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 metropolises 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. Gillespie’s argument points out the doublespeak of commercial social media companies. Content moderation efforts, including hiring human moderators and the use of AI, are discursive means of circumventing the “growth at all costs imperative that fuels these massive platforms in the first place” (2). More gravely, algorithmic amplification and political manipulation of polarized content are inextricably entwined with the logics of extractive digital capitalism (Donovan 2020; Morozov 2011).

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” (Gillespie, 2020: 2) 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).

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. It 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 “interconnect-edness”, 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” (7).

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.

The project 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.<sup>6</sup> The results of the ML models and the collaborative design underscore the point that ethical scaling is not merely about gauging the performance of the model for its accuracy in the first instance but involves ethical means for scaling a complex process through reflexive iterations.

When implemented with more resources for different languages and communities, systematic collaborative process—as envisaged in the initial effort of AI4Dignity—can provide unique entry points for technical scaling. For instance, the prevalence of culturally coded namecalling and hashtags reveals how they transform otherwise innocent expressions into exclusionary extreme speech. Datasets from India, for instance, have a panoply of racist expressions and coded allusions to deride Dalits (e.g., “Bhimte”) and Muslims, including “Mulle”, “Madrassa chaap Moulyvi” [referring to Muslim religious education centers], “hara virus” [green virus, the color green depicting Muslims], “Green Fungus” and the more insidious Potassium Oxide [K2O which phonetically alludes to “Katuwon” [derogatory term for Muslims] and “Ola Uber” [two riding apps which together phonetically resemble Alla Ho Akbar]. With community participation, such invectives coined against vulnerable communities can be catalogued as statistical cues for further human inspection, since automated models struggle to catch them.

Future development of ethical scaling should involve developing guidelines for selecting community intermediaries, involvement of communities in the process of training AI systems *and* defining the platform rules that those AI systems must enforce to reflect local and cultural realities and context, and developing standards for a review process

that considers the impact on vulnerable communities, benefit sharing (Birhane et al., 2022) and potential discrimination and bias. Such efforts can emerge only with a radical rethinking around AI by de-centering Euromodern rationality, deconstructing the paradigm of “the human” and associated moral panics, and embracing a collaborative, people-centric ethos that inevitably brings with it a confusing terrain of conflicting positions on speech and power but ultimately grounds annotation in experience-near, embodied knowledges and historically contextualized realities.

## Conclusions: risks and limitations of AI-assisted content moderation

Highlighting the limitations of AI-based systems on the content side of extreme speech in the preceding sections, 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. 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 extreme speech that are distributed via channels such as WhatsApp that tap community trust and penetrate via existing social ties such as neighborhood communities and kin groups. Such scenarios of distribution are common in countries like India, South Africa and Brazil (Udupa et al. 2021; Wasserman and Madrid-Morales 2022). The problem of content moderation addressed in this article through ethical scaling represents a specific, and admittedly, a small part in the broader set of issues concerning the dissemination, impacts and regulation of extreme speech.

With regard to the use of AI, studies have also raised concerns that the 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. 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 (de Almeida et al., 2021; High-Level Expert Group on Artificial Intelligence, 2019; Schiff et al., 2020).

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.

## Acknowledgments

This article draws on a discussion paper on ethical scaling which 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. We thank Leah Nann, Miriam Homer, Aleksander Szymanski, Swantje Kastrup 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 FactCheck, 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, Geoffrey Omondi, Vinod Rathi, Gilberto Scofield, Cristina Tardáguila, and Marita Wehlus for collaborating with us on the project.


## Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement number 957442).

## ORCID iDs

Sahana Udupa  <https://orcid.org/0000-0003-3647-9570>

Antonios Maronikolakis  <https://orcid.org/0009-0000-2463-2588>

Axel Wisioerek  <https://orcid.org/0000-0002-6058-0460>

## Notes

1. In Ross et al 2017, a German dataset,  $\alpha$  was between 0.18 and 0.29; in Sap et al. 2019, the  $\alpha$  score was 0.45; in Ousidhoum et al. (2019), a multilingual dataset,  $\alpha$  was between 0.15 and 0.24. Also, a majority of these works include neutral examples as well.
2. <https://www.perspectiveapi.com> accessed 13 July 2021.
3. <https://support.perspectiveapi.com/s/docs-get-started>
4. For descriptions of these categories, see <https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages>.
5. Passages were obtained by applying simple search patterns covering variants and common misspellings of these trigger words.
6. 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. The performance of BERT on hate speech datasets is examined thoroughly in Swamy et al. 2019. In Founta et al. 2018, the F1 score is 69.6. In Davidson et al. 2017, F1 is 77.3; in Waseem

- and Hovy 2016, F1 score is 58.4. In all these datasets, a major proportion of the content is neutral.
7. Redmond Bate vs Director of Public Prosecutions before the Lord Justice Sedley and Justice Collins on July 23, 1999; *The Times*, July 28, 1999.
  8. “Any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender, or other identity factor. This is often rooted in, and generates, intolerance and hatred, and in certain contexts can be demeaning and divisive” (United Nations, 2020). However, community involvement and contextual knowledge would be necessary to track the misuse of “hate speech” discourses.

## References

- Adams R (2021) Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews* 46(1/2): 176–197.
- Barrett PM (2020) *Who Moderates the Social Media Giants? A Call to End Outsourcing. Report, NYU Stern Center for Business and Human Rights, June*. NYU Stern Center for Business and Human Rights. Available at: <https://bhr.stern.nyu.edu/tech-content-moderation-june-2020> (accessed 16 December 2021).
- Becker LC and Becker CB (1992) *A History of Western Ethics*. New York: Garland Publication.
- Benesch S (2012) *Dangerous speech: A proposal to prevent group violence*. New York: World Policy Institute.
- Birhane A, Isaac W, Prabhakaran V, et al. (2022) Power to the People? Opportunities and Challenges for Participatory AI. In: *Equity and Access in Algorithms, Mechanisms, and Optimization*, 6 October 2022, pp. 1–8. DOI: 10.1145/3551624.3555290.
- Brubaker R and Cooper F (2000) Beyond ‘identity’. *Theory and Society* 29(1): 1–47.
- Boromisza-Habashi D (2013) *Speaking Hatefully: Culture, Communication, and Political Action in Hungary*. University Park, PA: Pennsylvania State University Press.
- Burnap P and Williams ML (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 R (2018) Content or Context Moderation? Data & Society Research Institute. Available at: <https://datasociety.net/library/content-or-context-moderation/> (accessed 21 November 2021).
- Davidson T, Warmesley D, Macy M, et al. (2017) Automated hate speech detection and the problem of offensive language. In: *Proceedings of the International AAAI Conference on Web and Social Media* 11(1), pp. 512–515.
- de Almeida PGR, dos Santos CD and Farias JS (2021) Artificial intelligence regulation: A framework for governance. *Ethics and Information Technology* 23(3): 505–525.
- Donovan J (2020) Social-media companies must flatten the curve of misinformation. DOI: 10.1038/d41586-020-01107-z
- Ferreira da Silva D (2007) *Toward a Global Idea of Race*. Minneapolis: University of Minnesota Press.
- Founta A, Djouvas C, Chatzakou D et al. (2018) Large scale crowdsourcing and characterization of twitter abusive behavior. In: *11th International Conference on Web and Social Media, ICWSM 2018*, AAAI Press, pp 2–18.
- Frenkel S and Alba D (2021) In India, Facebook Grapples With an Amplified Version of Its Problems. *The New York Times*, 23 October. Available at: <https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html> (accessed 11 December 2021).
- George C (2016) Managing the dangers of online hate speech in South Asia. *Media Asia* 42(3–4): 144–156.
- Gillespie T (2020) Content moderation, AI, and the question of scale. *Big Data & Society* 7(2): 1–5.
- Gorwa R, Binns R and Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7(1): 1–15.
- Gröndahl T, Pajola L, Juuti M, et al. (2018) All you need is ‘love’: Evading hate speech detection. In: *AISeC ’18: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, October 2018, pp. 2–12. New York: Association for Computing Machinery.
- Hallin D and Mancini P (2012) *Comparing Media Systems Beyond the Western World*. New York: Cambridge University Press.
- Hervik P (2019) Ritualized opposition in Danish online practices of extremist language and thought. *International Journal of Communication* 13: 3104–3121.
- High-Level Expert Group on Artificial Intelligence (2019) *Ethics guidelines for trustworthy AI*. European Commission.
- Krippendorff K (2003) *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications.
- Lee E (2020) Moderating content moderation: A framework for nonpartisanship in online governance. *American University Law Review* 70(3): 913–1059. Washington, United States: American University Law Review.
- Matsakis L (2018) Facebook’s AI Can Analyze Memes, but Can It Understand Them? *Wired*. Available at: <https://www.wired.com/story/facebook-rosetta-ai-memes/> (accessed 20 June 2022).
- Mhlambi S (2020) *From Rationality to Relationality: Carr Center for Human Rights Policy Harvard Kennedy School* (009). Carr Center Discussion Paper: 31.
- Mitchell L (2018) Civility and collective action: Soft speech, loud roars, and the politics of recognition. *Anthropological Theory* 18(2–3): 217–247.
- Morozov E (2011) *The Net Delusion: The Dark Side of Internet Freedom*. New York: Public Affairs.
- Murphy H and Murgia M (2019) Can Facebook really rely on artificial intelligence to spot abuse? *Financial Times*, London, November 8, 2019.
- Nockleby JT (2000) Hate speech. In: Levy WL, Karst KL and Winkler A (eds) *Encyclopedia of the American Constitution*. New York: Macmillan, 1277–1279.
- Ong JC (2021) *Southeast Asia’s disinformation crisis: Where the State is the biggest bad actor and regulation is a bad word*. Available at: <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/> (accessed 15 January 2021).
- Ousidhoum N, Lin Z, Zhang H, et al. (2019) 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), November 2019, pp. 4675–4684. Hong Kong, China: Association for Computational Linguistics.
- Perrigo B (2019) Facebook says its removing more hate than ever before: But there's a catch. *Time*. Available at: <https://time.com/5739688/facebook-hate-speech-languages/> (accessed 29 April 2020).
- Rieder B and Skop Y (2021) The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society* 8(2): 1–16.
- Roberts S (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.
- Ross B, Rist M, Guillermo CG, et al. (2017) Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In: *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pp. 6–9.
- Sablosky J (2021) Dangerous organizations: Facebook's content moderation decisions and ethnic visibility in Myanmar. *Media, Culture & Society* 43(6): 1017–1042.
- Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In: *64th Annual Meeting of the International Communication Association*, Seattle.
- Sap M, Card D, Gabriel S, et al. (2019) The risk of racial bias in hate speech detection. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 1668–1678. Association for Computational Linguistics.
- Schiff D, Biddle J, Borenstein J, et al. (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*, pp. 153–158. New York: ACM.
- Slack J (2006) Communication as articulation. In: *Communication as...: Perspectives on Theory*. Thousand Oaks: Sage Publications Ltd, 223–231.
- Swamy SD, Jamatia A and Björn Gambäck B (2019) Studying generalisability across abusive language detection datasets. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 940–950. Association for Computational Linguistics.
- Thiranagama S, Kelly T and Forment C (2018) Introduction: Whose civility? *Anthropological Theory* 18(2–3): 153–174.
- Udupa S (2018) Gaali cultures: The politics of abusive exchange on social media. *New Media & Society* 20(4): 1506–1522. March 2017.
- Udupa S, Gagliardone I and Hervik P (2021) *Digital Hate: The Global Conjunction of Extreme Speech*. Bloomington: Indiana University Press.
- Udupa S and Pohjonen M (2019) Extreme speech | extreme speech and global digital cultures. *International Journal of Communication* 13: 3049–3067.
- United Nations (2020) United Nations strategy and plan of action on hate speech: A detailed guidance on implementation for United Nations field presences. Available at: [www.digitallibrary.un.org](http://www.digitallibrary.un.org) (accessed 10 August 2021).
- Warner W and Hirschberg J (2012) Detecting hate speech on the world wide web. In: *Proceedings of the Second Workshop on Language in Social Media*, June 2012, pp. 19–26. Association for Computational Linguistics.
- Waseem Z and Hovy D (2016) Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: *Proceedings of the NAACL-HTL*, pp. 88–93. Association for Computational Linguistics.
- Wasserman H and Madrid-Morales D (2022) *Disinformation in the Global South*. Newark: Wiley.
- Wynter S (2003) Unsettling the coloniality of being/power/truth/freedom: Towards the human, after Man, its overrepresentation—an argument. *The New Centennial Review* 3(3): 257–337.
- Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.