DIGITAL TECHNOLOGY AND EXTREME SPEECH: APPROACHES TO COUNTER ONLINE HATE

Sahana Udupa

April 2021

Commissioned Research Paper for the
United Nations Peacekeeping Technology Strategy

United Nations Peacekeeping

CONTENT

EXECUTIVE SUMMARY

To address the spread and severity of online hate and disinformation, and the mammoth challenge they have posed to human rights protection, this strategy paper calls for keen attention toward dynamic scenarios where online vitriolic expressions, actors, practices, networks, and technologies are in a state of constant flux and evolution, and therefore, evasive and slippery for regulatory action and policy making. Keeping this in view, the paper proposes four priority areas for UN entities:

- tackling *global unevenness* in platform governance
- *connecting* critical communities
- monitoring *'gray' zones, fringe actors*, and *smaller/domestic platforms*
- engaging *repressive states* to tackle coordinated disinformation and hate campaigns.

To explore the dynamism of online vitriol and policy measures in the priority areas, the paper builds on the framework of 'extreme speech' rather than the more commonly invoked term, 'hate speech'. 'Extreme speech' emphasizes the importance of longer histories of exclusion, racialization and dispossession that underpin contemporary digital manifestations of hate. At the same time, it draws attention to rapidly mutating online user practices including recent trends of hateful language that comes cloaked in 'funny' memes and wordplays, and intricate networks of political manipulation that draw not only on technology but also social trust.

The normative emphasis of the 'hate speech' discourse hinges on the imperative for immediate action, and hence raises the risk of glossing over historical trajectories, cultural subtleties and evolving ground realities.

Moving beyond technological solutionism, crisis driven actions and moral panics about digital communication, the framework of extreme speech offers a way to develop culturally appropriate and holistic interventions. Such interventions can be grouped under four interconnected levels (global, national, bilateral and local/community) and a mix of five high-level action frames (intermediation, policy pressure, connection, monitoring, and training/awareness) relevant for each level.

Global Interventions

Global unevenness in platform governance

Existing social media platform governance practices around online extreme speech range from very strict regulations to very lax scenarios globally. Taking note of this vast unevenness, UN entities should mobilize political and diplomatic tools to facilitate the implementation of best practices across Member States, drawing lessons from some of the latest regulations and

policy proposals including:

- the principle of proportionality to assign greater obligations to large social media platforms for timely removal of illegal content, proactive risk mitigation and transparency measures
- regulations for online advertising and recommender systems
- robust systems for notices and trusted flaggers
- provision for independent auditing, code of conduct and independent dispute settlement bodies
- access to data for bona fide research
- online interface and interactive architectures that have the potential to change user behavior
- fair corporate practices for content moderation by removing opaque and exploitative arrangements around outsourced labor

Connection

Another key intervention on a global scale (replicable at the national and community levels) is to connect existing critical stakeholder communities to leverage, incubate, curate, and expand on best practices including coping strategies and countertactics to online extreme speech. Connection is an important action frame because there are already a large number of grassroots initiatives and organizations that are active in countering online hate. Connecting them for specific goals around different aspects of online hate can significantly increase their effectiveness and scalability and can also provide ways to address systemic issues such as racial bias. One example is the AI4Dignity project, which is developing a replicable process model to create collaborative spaces of coding by connecting artificial intelligence (AI) developers, fact checkers and academic ethnographers from different countries to detect and label extreme speech. Such activities will not only help in expanding technology access for the fact checking communities but also in addressing critical systemic issues such as bias and lack of transparency in AI-assisted content moderation by bringing more inclusive and culturally sensitive datasets. AI4Dignity's key focus is on fact checkers because they comprise a critical community with vast contextual knowledge about extreme speech. They are also often the targets of online extreme speech. UN entities can facilitate and fund collaborative coding spaces by involving other critical communities with similar contextual knowledge such as online comedians, anti-hate advocacy groups, grassroots digital influencers and independent journalists advocating for social justice.

Country Level Interventions

UN entities should support long-term studies for nuanced measures based on distinct media ecosystems that have evolved within different countries. However, three areas of intervention require urgent attention across the board:

Repressive and authoritarian assaults on online speech

- In cases of state aligned coordinated attacks or authoritarian controls over platform regulations, and dramatic turmoil when companies feel the pressure to take swift actions at the cost of due diligence processes, UN entities should apply pressure to, and if necessary, support social media companies to comply with global standards of content moderation and human rights protection by offering procedural clarity around escalation protocols and decision making. An institutionalized global structure to regularly convene social media companies to address repressive assaults on online speech will be a significant step towards addressing upheavals that unfold at the national levels. Convening different social media companies is especially critical during elections since disinformation campaigns funded by resource rich political parties have begun to increasingly adopt cross-platform manipulation tactics.
- When ruling governments are directly involved in digital disinformation and hate campaigns, it is critical to partner with civil society monitoring groups and global digital rights organizations for awareness raising and capacity building of key State actors such as judges and the judicial personnel. UN entities can also provide context sensitive positive narratives to social media companies to engage hate influencers online for user education and sensitization.
- Repressive regimes have begun to copycat stricter regulatory models adopted in developed economies with stable democratic systems for authoritarian controls over speech in their own countries. Monitoring how governments emulate stricter regulatory models and deploy AI-assisted technologies for repressive purposes is a critical area of intervention.

Gray zones, fringe actors and smaller/domestic platforms

- Engaging the "Big Tech" is crucial but policy measures should recognize that regulatory control over large transnational social media companies would not fully solve a complex social and economic problem. Political manipulation of online discourse through algorithmic and computational affordances has become the new face of electoral propaganda globally, but especially in the global South, partisan politics has spawned a breeding shadow industry that operates through gray practices of clickbait operators, hired influencers, and loosely knit networks of dispersed amplifiers who are drawn into precarious and informal labor arrangements crafted by ambitious mediators.

- Monitoring and supporting compliance to global standards among smaller, homegrown platforms and data influence services is important because they are increasingly implicated in shadow practices of extreme speech. Hate speakers have also often migrated to smaller platforms to avoid the regulatory gaze.

Gender based abuses

Gender-based abusive trolling is a particularly virulent form of online extreme speech and a disturbing trend that cuts across diverse cultures with vastly different levels of protection and opportunities for women and sexual minorities.

- Connection is a key action frame in this area. UN entities should connect anti-harassment campaigns crafted in different parts of the world for specific programs including digital safety trainings, free legal counsel, cyber harassment helpline, capacity building for reporting abusive content on social media platforms, partnerships with social media companies for quick response/redressal, and proposals for legislative reforms.
- UN entities should also help forge connections between anti-harassment campaigns and creative online feminist projects such as video tutorials and multimodal first-person narratives of women to push back against online trolling.
- Social media advertisers should be targeted to demote gender abusive content.
- In contexts where gender-based abuse is entangled with partisan or repressive politics targeting minorities and political opponents, the problem must be tackled as part of a broader set of tactics aimed at engaging repressive regimes and gray zones.

Bilateral and Geopolitical Interventions

Extreme speech is also a weaponized tool in bilateral and geopolitical conflicts to create and reinforce sentiments of mistrust, exclusion, fear, and anger toward perceived external enemies, and to also unite allies. In relation to this specific variant of online extreme speech, UN entities should evolve strategies in conjunction with diplomatic tools for intermediation and de-escalation, foremost by engaging key actors in Member States, introducing independent mediation and expertise, and combining these interventions with awareness raising activities among common online users.

Community Level Interventions and Deep Extreme Speech

If one part of extreme speech circulation relates to technology specific features of virality and algorithmic mediation, a significant part of it operates by tapping social trust and cultural capital at community levels, often making deep inroads into the "intimate sphere" of families, kin networks, neighbors, caste-based groups, ethnic groups, and other socially rooted formations as well as by building on historical structures of privilege. Mobilizing community

level awareness programs and rapid response systems that are sensitive to diverse social conditions of digital hate cultures is critical in addressing what might be described as "deep extreme speech". Key fields of action include:

- Partnering with local cultural influencers for organic influence in social media networks (such as WhatsApp groups) to promote positive narratives
- Mobilizing inclusive narratives and awareness raising by extending the network of partners to include not only conventional beneficiaries such as NGOs but also online comedians, poets, musicians, cinema celebrities, online meme creators, and online game developers
- Developing counterspeech and positive campaigns by using memes, GIFs, humorous posts and coordinated 'likes' for promoting the posts so that push back responses are culturally appropriate and digitally contemporary
- Convening self-styled political trolls, local politicians, and commercial digital influencers for awareness raising activities, and sensitizing them about global human rights standards and the dangers of digital campaign manipulations
- Strengthening grassroots anti-racist and anti-hate communities to report online extreme speech to social media companies and monitor progress once complaints are raised
- Strengthening local communities to petition lawmakers to support victims of online harassment and raise resources for legal help
- Offering technical support to local groups to develop hate monitoring dashboards
- Empowering local groups to mobilize community 'bystander support' when victims of online hate choose to make their complaints public
- Partnering with existing anti-hate media programs (radio, television, and print) to evolve integrated polymedia responses against online hate
- Developing innovative means of sensitizing hate speakers by channelizing donations to antihate groups for every instance of offensive and hateful speech act spotted online (i.e., hate speakers would be funding anti-hate initiatives each time they post a hateful message and thus undermining their own agendas).

INTRODUCTION

This strategy paper outlines critical and potential areas of engagement for UN entities to foster an open, safe, and accountable internet by addressing the spread and severity of online extreme speech, and the mammoth challenge they have posed to human rights protection. Moving beyond technological solutionism, crisis driven actions and moral panics about digital communication, it recommends investing in processes that can build sustainable communities of practice in a decentralized and connected ecosystem at the global, national, bilateral and community levels. This involves identifying key risk areas that require urgent and concerted top-level action focusing on governments and online social media platforms, and a longer-term engagement to combat hate within everyday cultures of online exchange. It calls for keen attention toward dynamic scenarios where online extreme speech expressions, actors, practices, networks, and technologies are in a state of constant flux and evolution, and therefore, evasive and slippery for regulatory action and policy making.

Recognizing the vast array of actions already activated by governments, multilateral agencies, local communities, internet intermediary service providers, academia, internet watchdogs and other stakeholders, as well as ongoing charged debates around what approaches are appropriate and what are inept, this paper proposes four priority areas for UN entities that can support, expand, and leverage existing efforts; set benchmarks for critical response; and intervene where other efforts have so far remained inadequate. It urges UN entities to focus their resources and expertise on:

- tackling *global unevenness* in platform governance
- *connecting* critical communities
- monitoring *'gray' zones, fringe actors*, and *smaller/domestic platforms*
- engaging *repressive states* to tackle coordinated disinformation and hate campaigns.

EXTREME SPEECH FRAMEWORK

To explore the dynamism of online vitriol and develop agile and context-sensitive responses in these priority areas, this paper builds on the framework of "extreme speech" rather than the more commonly invoked term, "hate speech". This shift is for several reasons:

- **Focus on practice:** Extreme speech framework emphasizes the need to focus on media practice, i.e., what people do that is related to media and how they reconfigure and reproduce broader structures of power within which such practices are embedded. The media practice perspective signals the importance of people's agency set within structural conditions of power and resulting dynamism in online ecosystems. Rather than focusing only on online content and data forensics, a keen understanding of online practices and online users' lifeworlds is needed to understand emergent forms of online hate and various networks of circulation that intricately intermingle to perpetuate them.
- **Ambiguity of speech:** Extreme speech framework emphasizes the situated nature of speech cultures. The same expression could work as subversive speech in some contexts, and hateful speech in others. The implications of incivility or extremeness of speech cannot be understood without analyzing particular forms of recognition and responsiveness to people's demands that exist in societies. In some contexts, expectations of civil language are an expression of power and subversive politics engages in extreme forms of speech to challenge the status quo. In other words, extreme speech can be a way to speak back to authorities, and policy measures should be sensitive to which groups engage in these speech forms, and what their relative power position is within particular social and political contexts.
- **Limits of hate speech:** Distinct from the normative emphasis of hate speech which comes with a heavy evaluative load, extreme speech stresses the importance of comprehension over classification and proposes to develop measures by understanding (if not condoning) actors, practices, and networks that constitute vitriolic cultures online.
- **Dangers of the "hate speech discourse":** Multifarious and often manipulative political agendas have grown around the regulatory discourse of hate speech. Examples abound where regimes have misused the hate speech discourse to squash dissent or target vulnerable groups. Repressive states have (mis)used the concepts of hate speech and lately disinformation by conjoining them with sedition, threat to national security, blasphemy, defamation, and other legislations. 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.[1]

1 Habashi, B. (2013). *Speaking Hatefully: Culture, Communication and Political Action in Hungary.* University

- **Lived concepts:** Extreme speech perspective calls for working with lived concepts and emic categories of communities for developing policy measures rather than the normative language imposed from the outside.
- **Epistemic parity and historical awareness:** Extreme speech research calls for deep contextualization that can account for grave historical continuities of racialization and dispossession instead of framing ongoing digital turbulences as a sudden crisis caused by digital communication. This entails systematic inquiries into longer histories of racial construction and hierarchies shaped by colonialism that have been revived and weaponized by current regimes, including those aimed against people within one's own national communities. The normative emphasis of the hate speech discourse hinges on the imperative for immediate action, and hence raises the risk of glossing over historical trajectories. Following this point of departure, extreme speech research has stressed for epistemic parity, and the call to depart from the self-righteous schema of the rational-liberal center (the self-understanding of the West) and the extreme periphery (the rendering of the non-West). The logic of the rational West versus the extreme other has long informed media development and media policy traditions engaged in tailoring solutions for hate speech. The extreme speech framework stresses that there is no center and periphery when it comes to violent emotionality of words. This critical perspective offers ways to identify global patterns, styles, and tropes of hateful speech that circulate between different repressive and exclusionary scenarios within and between the global North and the global South in the current digital age.

In terms of its definitional scope, extreme speech analysis draws a distinction between "derogatory extreme speech" aimed at any group (including those holding power) and "exclusionary extreme speech" that implicitly or explicitly excludes or causes harm to a person or a group on the basis of their group belonging.[2] In terms of exclusionary extreme speech, the analysis builds on existing definitional standards around hate speech set up by the United Nations[3] and the distinction that Wardle and Derakhshan draw between disinformation ("when false information is knowingly shared to cause harm") and malinformation ("when genuine information is shared to cause harm").[4] Extreme speech analysis covers misinformation

Park, PA: Pennsylvania State University Press.

2 See the section on AI4Dignity project in this paper (pp 22–23) for full definitions of these terms.

3 United Nations. (2020). *United Nations Strategy and Plan of Action on Hate Speech: A Detailed Guidance on Implementation for United Nations Field Presences.* The UN definition identifies hate speech as "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".

4 See Wardle, C., and Derakhshan, H. (2017). I*nformation Disorder: Towards an Interdisciplinary Framework for Research and Policy Making.* Council of Europe, https://edoc.coe.int/en/

(spreading false information without the intention to cause harm) so far as it is part of the social fields where deliberate efforts to spread hate activate a variety of actors and networks that end up spreading hateful language that could cause harm to vulnerable groups. The purpose of extreme speech analysis is therefore to exceed the legal focus on culpability and to instead analyze—with ethnographic and historical depth— the actual ways in which different actors and actions come to animate one another, and how new interventions need to be crafted to address not only those who deliberately "engineer" hateful language and disinformation but also those who are "taken by it" or do it to earn a livelihood. While recognizing the importance of crafting specific actions against actors and entities that deliberately spread hate, extreme speech analysis nonetheless widens the scope of culpability centric legal-normative analysis to a broader social-cultural analysis. This approach allows us to chart new analytical pathways and fields of action beyond intentionality-based investigations. Some of these fields of action in terms of connection, collaboration, and culturally appropriate trust-based interventions are highlighted throughout this paper.

In terms of research, using this framework and gleaning from cases around the world, extreme speech analysis has highlighted that in the last two decades, online vitriol and hateful cantankerous cultures have precipitated a condition of violent exclusion[5] based on "exacerbated fracture lines of difference that include race, gender, sexuality, religion, nation and class" in a context where "computational capital has built itself and its machines out of those capitalized and technologized social differentiations".[6]

On the level of digital practice, ethnographic studies in extreme speech research have shown that emerging vocabularies of hateful speech blend with idioms and humor genres that have cultural approval in local or national contexts.[7] Online hateful speech often comes dressed as jokes, "funny" memes, witty name calling, sobriquets, wordplays, and coded language. Furthermore, digital technologies have provided ways to develop new forms of extreme speech

media/7495-information- disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html accessed 15 March 2020. Now available at https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77

5 Udupa, S. (2017). Gaali Cultures: The Politics of Abusive Exchange on Social Media. *New Media & Society,* 20 (4), 1506–22. Udupa, S, Gagliardone, I., & Hervik, P. (eds.) (2021). *Digital Hate: The Global Conjuncture of Extreme Speech.* Bloomington: Indiana University Press. Udupa, S., and Pohjonen, M. (2019). Extreme Speech and Global Digital Cultures. *International Journal of Communication,* 13, 3049–67.

6 Beller, J. (2003). Numismatics of the Sensual, Calculus of the Image: The Pyrotechnics of Control. *Image & Narrative,* http://www.imageandnarrative.be/inarchive/mediumtheory/jonathanlbeller.htm

7 For instance, see Haynes, N. (2019). Writing on the Walls: Discourses on Bolivia Immigrants in Chilean Meme Humor. *International Journal of Communication,* 13, 3122–42 for a discussion on internet memes that target Bolivian immigrants in northern Chile. See also Hervik, P. (2019). Ritualized Opposition in Danish Online Practices of Extremist Langauge and Thought. *International Journal of Communication,* 13, 3104–21.

that come in the guise of "facts" and "evidence-based" untruths targeting specific groups rather than employing explicitly derogatory and dehumanizing language. "Deep fakes" are a good example for how digital mediation allows ways to present discriminatory and hateful language as trustworthy information that appeals to visceral apprehension, often confusing the senses. These trends suggest that there is an emerging overlap between digital disinformation and hateful speech, although each cannot be reduced to the other. Even more gravely, repressive and authoritarian regimes have weaponized online extreme speech, subjecting their own citizens to violent surveillance, and violating human rights norms in their policies towards refugees, immigrants, minorities, and historically disadvantaged communities. There are thus global (technologized) patterns to exclusions as well as national and local manifestations that are often culturally sanctioned and regime backed.

How can UN entities devise novel and effective ways to combat this complex scenario?

Without doubt, UN actors and entities should work with regional and national level legislations around hateful speech, but they need to simultaneously address broad ranging developments that can impact foundational definitions, especially where human rights obligations of States are waning and protections to vulnerable communities are under direct attack. Importantly, they should support and evolve mechanisms that can embed extreme speech moderation and mitigation efforts within democratic processes, however messy and prolonged these processes might be.

Drawing attention to the dynamic flows of online extreme speech that are simultaneously global, national, and local, this paper urges UN actors to develop a multiprong, multi-layered approach that can build robust flexibility and context sensitivity into response and mitigation strategies.

The rest of the paper will present a schema to situate the importance of the four priority areas flagged at the beginning, elaborating on other areas of action that are related to them, and how they can be developed both as a specific mix of measures and holistically, to address the problem of online extreme speech.

Towards developing a systematic approach, the paper will first offer definitions of a set of high-level action frames for UN entities and the rationale underpinning them. Following this, it proposes strategies for four distinct yet overlapping domains of intervention (global, national, bilateral, and local) by highlighting a mix of action frames relevant for each domain. The paper will unpack the action frames by offering a set of concrete measures that can be further developed and combined based on a grounded understanding of the specific challenges relevant for different domains of intervention. Each action frame involves engagement with one or

more stakeholders: governments, social media platforms (as part of the broader set of internet intermediary service providers), civil society organizations (CSOs)/NGOs/communities and academia/researchers. The paper will discuss illustrative cases to demonstrate the relevance and benefits of different measures as well as possible challenges and further work.

ACTION FRAMES

- **Intermediation** – involves strategic mediations between governments and social media platforms, especially in evolving sound country level regulatory practices; between social media companies and researchers, as part of the transparency agenda and research access; and between governments and researchers when critical research on online extreme speech is threatened by repressive states.

- **Policy pressure** – involves specific strategic action points, including addressing global unevenness in platform governance; placing pressure on social media companies to fund grassroots organizations and research activities aimed at tackling extreme speech; and addressing threats to activists and political misuse of legal provisions. A global institutional structure to regularly convene social media companies (both large and small), state regulators, and CSOs at the UNHQ is critical.

- **Connection** – involves connecting, curating, and scaling up already existing critical communities in a multilateral way (rather than in the hub-spoke model)[8] that can leverage the UN's vast global reach and community level organization and offer scalability to related initiatives in the area of extreme speech mitigation. These critical communities include fact checkers, anti-hate groups, online comedians, AI developers, and independent journalists. Such connections are important also because they can bring people centric perspectives to machine learning models to operationalize the 'human-in-the-loop' principle and inclusive AI, in the current context where AI systems are playing an increasingly important role in extreme speech amplification and moderation.

- **Monitoring** – involves strengthening efforts to monitor online extreme speech patterns by integrating ongoing initiatives and new commissioned research to develop online hate dashboards and shared database.

- **Training and awareness** – include consultations and awareness raising about global standards around hateful speech and international human rights norms. These activities also include resource sharing for best practices and creative interventions that are in sync with digitally native habits, styles, and jargons. This could be achieved by expanding the ambit of participants beyond conventional beneficiaries such as local NGOs into a broader range of actors—politicians, self-identifying trolls, online game developers, online meme generators, entrepreneurial digital influencers, and victims of online extreme speech.

A combination of action frames can be mobilized for different levels of interventions.

8 This builds on Goldman and Chen's (2010) application of a layered model of regulation for public service broadcasting. They suggest that public service media should encompass a wider range of content providers and information activists through a decentralized mechanism that addresses all the four layers—physical infrastructure, connection (between various platforms engaged in public service media), curation (supporting content and services of public value) and creation (creating content that the market insufficiently or erroneously addresses). See Goldman, E., & Chen, A. H. (2010). Modelling policy for new public service media networks. *Harvard Journal of Law and Technology,* 24(1), 111–170.

GLOBAL INTERVENTIONS

Key action frames: Policy pressure and Connection

The boundary defying tendency of digital technologies and the global reach of transnational tech companies have led to an upsurge of information flows that crisscross conventional territorial borders in unforeseen ways. On the one hand, digital participatory cultures have facilitated anti-authoritarian uprisings and multifarious social movements around climate justice, anti-racism (#BlackLivesMatter), decolonization, anti-harassment (#MeToo) and other pertinent issues, infusing these struggles with planetary resonance. On the other hand, the same infrastructural possibilities have enabled hateful speech to augment and gain virality on a global scale. Although there are still large gaps in global research on the topic and there is no consensus on whether online hate speech is on the rise, existing studies have shown that there are globally circulating tropes and resources that shape and ramp up online hateful expressions. Anti-legacy media criticism and skepticism, for instance, is a trope that is common across online right-wing supporters in Germany, the US, India, Denmark, Turkey, Hungary and other countries.[9] Beyond thematic patterns, there are globally shared cultures of online use and extreme speech formats that variously enable people to say things that they would not say in "real life" interactional situations. Internet memes and trolling are illustrative examples for how globally shared digital formats and practices can provide the means for exclusionary discourses of different kinds to manifest and amplify within distinct national or regional contexts. The format-inducing effects of the global internet are strikingly evident in digital fun cultures that embed distance and deniability in hateful exchange.[10] Trolls are able to participate in collective celebration of aggression—cheering each other and jeering at opponents—and simultaneously distancing themselves from the consequences of what they say and do online.

Perhaps the most important aspect of the global dimension of online extreme speech is the influence of transnational tech companies and the diverse policies they have evolved around hate speech moderation. UN entities have a particularly important role to play in engaging transnational internet intermediary service providers and social media platforms and apply pressure for legal compliance and social responsibility. Rather than overemphasizing the aspect of media literacy that implicitly places the burden upon ordinary online users to report

9 For instance, a study based on online content analysis found that there are common vocabularies of hateful speech between "international" alt-right groups and parts of the Irish digital sphere. See Siapera, E., Moreo, E., & Zhou, J. (2018). *Hate Track: Tracking and Monitoring Racist Speech Online*. Dublin: Irish Human Rights and Equality Commission.

10 Udupa, S. (2019). Nationalism in the Digital Age: Fun as a Metapractice of Extreme Speech. *International Journal of Communication,* 13, 3143–63.

extreme speech and detect disinformation, a significant part of critical measures should be concentrated on social media companies and their obligations toward upholding democratic values. The foremost of these measures, this paper suggests, is to address global unevenness in platform governance.

Policy Pressure: Platform Governance

Existing platform governance practices around online extreme speech globally range from very strict regulations to very lax scenarios. There are high penalties on social media companies for non-compliance and failure to respond within tighter timeframes in countries like Germany[11] and France.[12] Similarly, federal laws in Ethiopia require social media companies to remove hate speech or disinformation in one day.[13] If Singapore's latest regulation has raised an outcry around what is dubbed as the "Orwellian fake news law",[14] China's restrictive censorship laws are long known for raising the indomitable "firewall". At the other end of the spectrum, there are countries where social media companies do not even meet the formal requirements of appointing a legal representative to address the concerns that users and authorities raise.[15] In Brazil, Facebook has exempted politicians from fact checking, despite fact checkers complaining with evidence that elected representatives regularly relay hateful speech and disinformation.[16] Following the dramatic Capitol Riot in 2021, Twitter suspended the accounts

11 https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html accessed19 February 2021. The Network Enforcement Act, NetzDG 2017, requires social networks to remove "manifestly illegal content" within 24 hours or face heavy fines up to "5 million euros against the person responsible for the complaints management system" and "the fine against the company itself can be up to 50 million euros." The law has a broad definition of punishable content. Aside from "punishable fake news and other unlawful content, it includes "insult, malicious gossip, defamation, public incitement to crime, incitement to hatred, disseminating portrayals of violence and threatening the commission of a felony."

12 In May 2020, France adopted a bill to counter online hate (Projet de loi Avia, the Avia Law). Digital rights advocates have criticized the regulations in France and Germany for abdicating the due diligence process of appropriate judicial review before content removal. See https://www.article19.org/resources/france-the-online-hate-speech-law-is-a-serious-setback-for-freedom-of-expression/, accessed 19 March 2021. In June 2020, the French Constitutional Council declared that the main provisions of the "Avia law" unconstitutional https://edri.org/our-work/french-avia-law-declared-unconstitutional-what-does-this-teach-us-at-eu-level/ accessed 19 March 2021.

13 https://www.accessnow.org/cms/assets/uploads/2020/05/Hate-Speech-and-Disinformation-Prevention-and-Suppression-Proclamation.pdf accessed 18 March 2021.

14 https://rsf.org/en/2020-world-press-freedom-index-entering-decisive-decade-journalism-exacerbated-coronavirus accessed 15 February 2021.

15 The flip side of this regulatory requirement will be discussed under the section on repressive and authoritarian states.

16 https://www.poynter.org/fact-checking/2021/facebook-has-an-apparent-double-standard-over-covid-19-misinformation-in-brazil-researchers-say/ accessed 19 March 2021; https://www1.folha.uol.com.br/equilibrioesaude/2021/03/bolsonaro-violou-regras-do-facebook-para-covid-ao-menos-29-vezes-em-2021-mas-nao-foi-punido.shtml accessed 19 March 2021.

of Donald Trump and many of his supporters, sparking a global debate over "deplatforming".[17] However, similar actions are lacking in countries like India where political parties have engaged in coordinated campaign manipulations. Taking note of this vast unevenness, UN entities should mobilize political and diplomatic tools to facilitate the implementation of best platform governance practices across Member States, drawing lessons from some of the latest regulations and policy proposals.

The two landmark digital legislations proposed by the European Union in December 2020—the Digital Services Act (DSA) and Digital markets Act (DMA)—are noteworthy for some of the far-reaching policy directions they have enunciated in relation to "illegal content" (defined according to Union or Member State legislations). Also hailed as the "new constitution for the internet",[18] the proposed legislations have touched up several critical areas for platform governance, highlighting the need for institutional safeguards, due diligence, and procedural clarity for content moderation and free speech. While continuing to exempt platforms from liabilities for illegal content posted by users (similar to the US legislations), the proposed legislations nonetheless hold the platforms liable if they do not act with a specified timeframe to remove access to such content once they receive notices or complaints. Beyond content takedowns, the regulation allows for other substantive policy measures. Some of these proposed measures are directly relevant in addressing the problem of global unevenness in platform governance and anchoring such efforts to legitimate public interest objectives:

- **Principle of proportionality:** Implementing the principle of proportionality, the DSA has proposed more severe obligations on "very large platforms" defined as "systemic platforms"[19] that have more than 45 million users in the EU region. The obligations are proportionately distributed based on the size and nature of services.[20] The obligations on systemic platforms include strict transparency standards that require them to publish reports to inform policymakers, users, regulators, and researchers about how they curate, moderate, categorize and remove online content. UN entities should lobby for similar

17 Guo, E. 2021. Deplatforming Trump will work, even if it won't solve everything. MIT Technology Review, https://www.technologyreview.com/2021/01/08/1015956/twitter-bans-trump-deplatforming/ accessed 19 March 2021.

18 https://www.golem.de/news/digitale-dienste-gesetz-das-neue-grundgesetz-fuer-die-internetwirtschaft-2012-152892.html accessed 13 February 2021.

19 https://ec.europa.eu/digital-single-market/en/digital-services-act-package accessed 15 February 2021.

20 The regulation highlights a nested structure—"internet intermediary service providers" refer to the whole range of intermediary services while hosting services are part of this larger set, and online networks a subset of hosting services. The regulation "sets out basic obligations applicable to all providers of intermediary services, as well as additional obligations for providers of hosting services, and more specifically, online platforms and very large online platforms". Additional obligations are not applied to what the Union has defined as "micro or small enterprises" to "avoid disproportionate burdens" https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

additional obligations on global social media corporations that operate in Member States although policy measures should not ignore smaller, domestic players.[21]

- **Removal of illegal content:** the DSA requires internet service providers to act "expeditiously to remove or to disable access to that content…upon obtaining actual knowledge or awareness of illegal content". This knowledge could come from its own investigations or notices submitted by individuals or entities recognized by the regulation.
- **Independent regulator and fines:** the DSA's proposal to establish a strong and autonomous European regulator to oversee and implement the legislations comes armed with heavy dissuasive fines (as high as 6 per cent of the annual income or turnover of the internet service provider).
- **Online advertising:** Recognizing that online advertising can further amplify illegal and harmful content, the DSA regulation has mandated online platforms to maintain archives of advertisements they publish and "ensure that the recipients of the service have certain individualized information necessary for them to understand when and on whose behalf the advertisement is displayed.[22] In addition, companies are obliged to provide recipients of the service with information on the "main parameters used for determining that specific advertising is to be displayed to them, providing meaningful explanations of the logic used to that end, including when this is based on profiling."[23] This policy measure can be another rallying point for UN entities to tackle targeted and computational political campaigns and "dark ads"[24] that run on divisive and hate-filled agendas—a scenario that has gained growing salience on a global scale.[25]Although companies like Facebook are publishing data on advertisements in publicly accessible formats, there is still vast unevenness both geographically and among social media platforms.
- **Recommender systems:** the DSA places additional regulatory controls on algorithmically mediated recommender systems of platforms. Article 29 of the regulation states that the platforms "should clearly present the main parameters for such recommender systems in an

21 See the discussion under "Gray zones, fringe actors and smaller platforms".

22 The US has proposed the Honest Ads Act (2017). https://www.congress.gov/bill/115th-congress/senate-bill/1989 accessed 15 February 2021.

23 Public policy commentators have noted that the proposed EU legislations need to place more restrictions on micro-targeted "hyper-invasive surveillance advertising", urging for stricter regulations with e-privacy mandates on online advertising. https://edri.org/our-work/eu-attempt-to-regulate-big-tech/ accessed 10 February, 2021.

24 In the Brexit referendum and 2016 U.S. election, for instance, reports have revealed the circulation of "dark ads" on social media platforms. These advertisements had no "accompanying information about their funding or why they were targeted at users." https://time.com/5921760/europe-digital-services-act-big-tech/ accessed 13 February 2021.

25 Bradshaw, S., & Howard, P. N. (2017). *Troops, trolls, and troublemakers: A global inventory of organized social media manipulation* (Working Paper 2017.12). Project on Computational Propaganda, University of Oxford.

easily comprehensible manner to ensure that the recipients understand how information is prioritized for them. They should also ensure that the recipients enjoy alternative options for the main parameters, including options that are not based on profiling of the recipient." This is yet another important policy measure to tackle algorithmically mediated radicalization and its potential to create aggressive ideological hate groups primed with regular exposure to information that deepens their bias.

- **Notices and trusted flaggers:** the DSA regulation requires providers of internet hosting services to implement "user-friendly notice and action mechanisms" and internal complaint handling systems through which users can report violations. Platforms in turn are obligated to inform the user (whose content has been removed) of its decision, the reasons for its decision, and available redressal possibilities to contest the decision. To protect against misuse of this provision, the regulation (Article 20) allows online platforms to "suspend, for a reasonable period of time and after having issued a prior warning, the processing of notices and complaints…by individuals or entities or by complainants that frequently submit notices or complaints that are manifestly unfounded." Through the category of "trusted flaggers", the regulation proposes to expedite this process for greater public good. Online platforms are obligated to process and decide on the notices submitted by trusted flaggers "on priority and without delay". Trusted flagger status is "awarded to entities and not individuals that have demonstrated…that they have particular expertise and competence in tackling illegal content, that they represent collective interests and that they work in a diligent and objective manner." UN entities should not only advocate for this policy proposal globally but can also mediate in identifying trusted flaggers within Member States.
- **Independent dispute settlement bodies:** To protect against indiscriminate takedowns and infringement of freedom of expression, the DSA (Article 18) has proposed to set up certified dispute settlement bodies to which online users can lodge complaints and seek redressal, after failing to find redressal through platforms' internal complaint procedures.
- **Risk mitigation:** Placing further expectations on very large platforms, the DSA (Article 26) requires them to implement risk mitigation measures following an assessment of systemic risks that arise from coordinated manipulation of the platform's service and intentional sabotage. Platforms are expected to "enhance or…adapt.. the design and functioning of their content moderation, algorithmic recommender systems and online interfaces. They may also include corrective measures, such as discontinuing advertising revenue for specific content, or other actions, such as improving the visibility of authoritative information sources." Platforms are also encouraged to "initiate or increase cooperation with trusted flaggers, organize training sessions and exchanges with trusted flagger organizations, and cooperate with other service providers, including by initiating or joining existing codes of conduct or other self-regulatory measures." Proactive measures envisaged by the DSA

provide yet another substantive area for policy advocacy for UN entities.

- **Transparency reports:** the DSA (Articles 13, 23) requires internet service providers (except micro-or small enterprises) to annually report on "the content moderation they engage in, including the measures taken as a result of the application and enforcement of their terms of conditions". It is worth noting that companies have been responding to similar requirements in the United States. In November 2020, Facebook disclosed the extent of hate speech that is shared on its platform, revealing that "out of every 10,000 content views in the third quarter, 10 to 11 included hate speech."[26] However, civil rights organization Anti-Defamation League, one of the groups that led the advertisement boycott against Facebook in the summer of 2020, argued that the report did not provide information on the total number of hate speech instances that users had flagged and to what extent the company had taken action on these reports.[27] UN entities should address uneven application of this requirement globally, placing pressure on global social media corporations to maintain similar standards across locations where they operate.
- **Independent auditing:** the DSA requires platforms to ensure independent expert verification and provide "all relevant data necessary to perform the audit properly." Although companies like Facebook have implemented such measures in the US context,[28] there are huge gaps in the global wide application of similar practices.
- **Code of conduct:** There have been several efforts at drawing a voluntary code of conduct for social media companies, but the DSA has tightened the regulatory norm by placing obligations on very large online platforms to comply with a code of conduct and any refusal to participate in the application of the code of conduct invites scrutiny for possible infringement of obligations laid down by the regulation.[29] This is yet another area for policy advocacy for UN entities.
- **Research and access:** Highlighting the value of research, the DSA has proposed a framework that compels large online platforms to provide data access to vetted researchers. Considering the opaque operations of social media companies and their heavy-handed approach to research requests and blocking of API access,[30] UN entities should advance this

26 https://www.reuters.com/article/uk-facebook-content/facebook-offers-up-first-ever-estimate-of-hate-speech-prevalence-on-its-platform-idINKBN27Z2QY accessed 16 February 2021.

27 https://www.reuters.com/article/uk-facebook-content/facebook-offers-up-first-ever-estimate-of-hate-speech-prevalence-on-its-platform-idINKBN27Z2QY accessed 16 February 2021.

28 See Facebook's Civil Rights Audit 2020, https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf accessed 17 February 2021.

29 https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN. These stricter measures have followed from criticism of non-compliance and lack of standardized and transparent procedures in the implementation of the Code of Conduct.

30 Freelon, D., & Wells, C. (2020). Disinformation as political communication. Political Communication, 37(2), 145–156.

policy objective and facilitate critical research through greater data access.[31]

In addition, towards evolving more even and equitable platform governance practices, two more areas require urgent action:

- **Corporate practices for content moderation:** Social media companies—especially the workforce that directly deal with online speech moderation—should be inducted into and socialized within the discursive institution of journalism. A necessary step is to urge social media companies to equip content moderators with literacy around the "cognitive toolkit"[32] and conventions of journalism that place normative emphasis on truth telling for public good. Social media companies should be pressed to position content moderators as meaningful agents performing a publicly relevant role rather than treating them as "low-status" workers in their organizational hierarchies.[33] Companies should organize or enhance existing training programs for content moderators by inviting critical scholars and CSOs/NGOs so that moderators develop a keen understanding of the various political, cultural and social issues that shape online extreme speech within national or local contexts. Large social media companies have engaged academic researchers in an ad hoc manner, and sometimes, such engagements hinge upon the commitment and enterprise of individual executives stationed at different country level offices rather than a rigorous company-wide policy applied across all the locations. Importantly, companies should also put in place sufficient measures to mitigate psychological and emotional stress associated with content moderation work by providing on-the-floor counseling support and fair working conditions.[34] Although social media giants such as Facebook and YouTube have regularly issued public statements to affirm their commitment towards protecting content moderators from psychological stress, these measures have not been uniform, while smaller platforms, on the other hand, have maneuvered local connections to evade the regulatory gaze on this issue. Indian media reports, for instance, have highlighted that content moderators working for smaller platforms like TikTok (banned in India since 2020), LIKEE

31 The DSA's proposed framework requires that "access to data…should be proportionate and appropriately protect the rights and legitimate interests, including trade secrets and other confidential information, of the platform and any other parties concerned, including the recipients of the service."

32 Hanitzch, T. et al. (2019). Journalistic culture in a global context: A conceptual map. In Worlds of Journalism: Journalistic Cultures Around the Globe (pp. 23–45). New York: Columbia University Press. p. 33.

33 Roberts, S. T. (2019). *Behind the Screen: Content Modreation in the Shadows of Social Media.* New Haven, CT: Yale University Press.

34 A report published by Forbes notes that, "Facebook employs about 15,000 content moderators directly or indirectly. If they have three million posts to moderate each day, that's 200 per person: 25 each and every hour in an eight-hour shift. That's under 150 seconds to decide if a post meets or violates community standards". https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=37b95ce54d03 accessed 24 November 2020.

and Bigo Live, which are newly popular for their short and live video sharing, do not have the "luxuries like counsellors".[35] In addition to addressing grossly inadequate support systems available for content moderating labor, a further step would be to urge companies to end opaque and exploitative contractual arrangements with third party vendors, and instead recruit content moderators as regular employees with protections and perks comparable to those extended to any other technology service.[36] These recommended measures recognize that robust content moderation practices can follow only when organizational processes and structures behind them are fair and robust.

- **Targeting hate influencers online with positive narratives:** UN entities, with their vast networks of local units, should provide context sensitive positive narratives to global social media companies directly to target hate influencers. Such measures can complement and balance other restrictive, securitized approaches such as data forensics for content takedowns and blocking problematic social media accounts.

Simultaneously, UN entities should be alert on the potentialities of emerging technological developments such as open-source protocols of "alternative social media" that have promised greater community autonomy in contrast to corporate social media's "layers of abstraction and centralization that eliminate users from decision-making processes".[37] Mastodon, a decentralized microblogging system, for instance, has developed a social media architecture that offers more user control over content and data, but the actual political implications of their development remain to be seen.

35 https://www.livemint.com/news/india/inside-the-world-of-india-s-content-mods-11584543074609.html accessed 24 November 2020.

36 A study published by the NYU Stern Centre for Business and Human Rights calls for ending outsourcing in content moderation activities, urging Facebook to provide secure employment to content moderators and bring content moderation practices under the oversight of experienced executives. See https://www.stern.nyu.edu/experience-stern/faculty-research/who-moderates-social-media-giants-call-end-outsourcing Following mounting public pressure, Facebook agreed in principle in May 2020 to "pay US$52 million to compensate current and former content moderators who developed mental health issues on the job" (https://techcrunch.com/2020/05/12/facebook-moderators-ptsd-settlement/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAA6XmkmzvdmES5PBDEp3dmyEjW8_EA_6vga59LlTP25GesJd0LbqWBuKry1vaAq45GF4qez6mTFTlU8rmw99IdiIDEpwEJe3TKJElp9b1F_-vLbcoc0qtiEl9sSudoP6Wq9aI3ApogQ7PmrqLCYG55fMA5eTTckJljgJoyczATpS accessed 24 November 2020). However, media reports in countries like India that houses several outsourcing centers for the global tech companies pointed out that the lawsuit covered only people who have worked for Facebook through third-party vendors in the US, leaving out vendors spread around the globe (estimated to be 11,250 people). https://www.deccanchronicle.com/technology/in-other-news/160520/why-india-needs-to-be-the-centre-for-content-moderation-reform.html accessed 24 November 2020.

37 Zulli, D., Liu, M., and Gehl, R. (2020). Rethinking the 'social' in social media: Insights into topology, abstraction, and scale on the Mastodon social network. *New Media & Society,* 22(7): 1188–1205.

Connection

Another key intervention on a global scale (and moving down, on the national and community levels) would be to connect existing critical stakeholder communities to leverage, incubate, curate and expand on best practices, coping strategies, technologies, and countertactics to online extreme speech. Some examples are the shared repositories of fact checking tools that different organizations have created to help each other. These initiatives are laudable but access to such repositories is constrained by language (since these tend to be largely in English) as well as expected technological knowledge and internet access.

Addressing language and other constraints, the action frame of "connection" should be strengthened across diverse projects and involved groups. This action frame can not only offer ways to build and share a repository of successful initiatives, tools, and experiences, but they also help in addressing critical systemic issues such as bias and lack of transparency in AI-assisted content moderation and filtering, and corporate control over determining hateful language. Below is a long description of the AI4Dignity[38] project as an illustrative case for designing and implementing measures that can operationalize the human-in-the-loop principle by facilitating connections between academic researchers from different disciplines and critical communities such as fact checkers on a global level.

AI4Dignity: Collaborative AI Counters Hate

Responding to the challenge of combating online hate, governments and companies have increasingly turned to AI as a tool that can detect, decelerate, and remove online extreme speech. AI deployment is explored in several areas of content moderation: detecting content (flagging, tagging, and labeling); evaluating content (blocking and takedowns); curating content (recommending, promoting or downranking content); and responding to content (automated messages and responses to detected content). Deployment of AI is assumed to bring scalability, reduce costs, and decrease human discretion and emotional labor. However, mounting empirical evidence attests that such efforts face many challenges.

One of the key challenges is the quality, scope and inclusivity of training data sets.
Several studies have shown that classification algorithms are limited by the homogenous work forces of technology companies that employ disproportionately fewer women and

38 AI4Dignity project (2021–2022) is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 957442). This author is the principal investigator of the project. Other members include Hinrich Scheutze, Elonnai Hickok, Antonis Maronikolakis, Axel Wisiorek, Laura Csuka and Leah Nann.

people of color.[39] Language-based asymmetries and uneven allocation of corporate and state resources for extreme speech moderation that affects different communities globally and within the nation-state are other reasons for quality issues in training datasets.

The second challenge is the lack of procedural guidelines and frameworks that can bring cultural contextualization to these systems. There is obviously no catch-all algorithm that can work for different contexts. Lack of cultural contextualization has resulted in false positives and overapplication. In addition, hate groups have managed to escape keyword-based machine detection through clever combinations of words, misspellings, satire, and coded language. For instance, a UN sponsored Independent International Fact-Finding Mission in Myanmar found that, "subtleties in the Myanmar language and the use of fables and allegories make some potentially dangerous posts difficult to detect."[40] The dynamic nature of online hate speech—where hateful expressions keep changing—adds to the complexity.

The difficulty of deploying AI-assisted systems for content moderation in diverse national, linguistic and cultural contexts is further compounded by the fact that groups that are directly involved in flagging online hate-speech content at the local and regional levels lack the technological tools that can expedite and scale up their work, although they come with cultural knowledge about what constitutes hateful speech within specific contexts.

Although these challenges are widely acknowledged, corporate imaginations still position AI as a tool that can generate best decisions, because the inherently "neutral" machine-learning model is presumed to only become more robust, training itself with more and more data, in the onward march towards perfecting what Silicon Valley "high priests" ambitiously define as "social physics."[41] This utopian vision of AI conceals the materialities and politics behind AI technologies. As Ed Finn writes, "there is no such thing as 'just code'."[42] Algorithms are always the "product of social, technical, and political decisions and negotiations" that occur throughout their development and implementation.[43]

The process of choosing and labeling information that feeds the "machine" is never neutral. Offering one way to address this

39 Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York: New York University Press.

40 Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar, U.N. Doc.A/HRC/39/CRP.2, September 17, 2018, para 1311.

41 Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* New York: Public Affairs.

42 Ed, F. (2017). *What Algorithms Want: Imagination in the Age of Computing.* Cambridge, MA: MIT Press.

43 Forlano, L. (2018). Invisible Algorithms, Invisible Politics. Public Books blog, February 2, 2018. https://www.publicbooks.org/invisible-algorithms-invisible-politics/ accessed 12 January 2020.

challenge, AI4Dignity is implementing a community-based approach by involving fact checkers as critical intermediaries.

Without doubt, fact checkers are already overburdened with verification related tasks, but flagging extreme speech could be 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. Our interactions with independent fact checkers have shown how their inboxes are filled with hateful messages because their public role in verification invariably upsets groups that seek to (re)shape public discourse for exclusionary ideologies. By involving fact checkers, AI4Dignity aims 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. In this way, it is creating a mechanism where the "close cousin" of disinformation, namely extreme speech, is spotted during the course of fact checkers' daily routines, without interrupting their everyday activities as much as possible.

Building spaces of direct dialogue and collaboration between AI developers and relatively independent fact checkers who are not part of large media corporations, political party machineris or social media companies is a key component of AI4Dignity. Furthermore, this dialogue has involved ethnographers specialized in particular regions in developing the labels and verifying the datasets.

A key activity of AI4Dignity is "Counterathon": a marathon of coding to counter online extreme speech. During the event, AI/NLP (natural language processing) developers and independent fact checkers will work in small national teams overseen by ethnographers. A series of interactions between the research team and fact checkers and annotated passages that fact checkers will upload before the event have provided the preliminary groundwork for the NLP models to undergo further iterations and modifications during Counterathon.[44]

44 Building on existing definitions, this author has developed a model for three categories of extreme speech for annotation: derogatory extreme speech, exclusionary extreme speech, and dangerous speech. Fact checkers are requested to label the passages (ranging from the minimum string of words that comprises a meaningful unit in a particular language) to about six to seven sentences under these three categories. Derogatory extreme speech refers to expressions that do not conform to accepted norms of civility within specific regional/local/national contexts and targets people/groups based on racialized categories or protected characteristics (ethnicity, national origin, caste, religious affiliation, sexual orientation, gender, language group) or others (state, media, politicians). It includes derogatory expressions not only about people but also about abstract categories/concepts 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 did not tend to provoke violence" or amounted to implied or direct call for exclusion of target groups. The cited passage comes from Redmond Bate vs Director of Public Prosecutions before the Lord Justice Sedley and Justice Collins on July 23, 1999; The Times, July 28, 1999. Exclusionary extreme speech Expressions that call for or implies excluding historically disadvantaged and vulnerable people/groups from the "in-group" based on national origin, gender, sexual orientation, ethnicity, caste, racialized

Through a facilitated triangulation between fact checkers, AI developers and ethnographers, the project is developing a replicable process model that can create collaborative spaces beyond the purview of global corporations. This process model aims to stabilize a more encompassing collaborative structure in which the "hybrid" models of human-machine filters are able to incorporate dynamic reciprocity between critical communities. The AI4Dignity toolbox will provide guidelines to organize similar events and replicate the model at different locations and on different scales (local, national and subnational/regional).

Extreme speech databases that are generated during the pilot Counterathon event and further contributions will contribute towards research analysis of extreme speech patterns and targets.[45]

Figure 1 provides a schematic diagram for the collaborative community-based classification approach, outlining the process model for advocacy and implementation at various levels, and the resulting decentralized extreme speech datasets as resources for evidence-based policymaking.

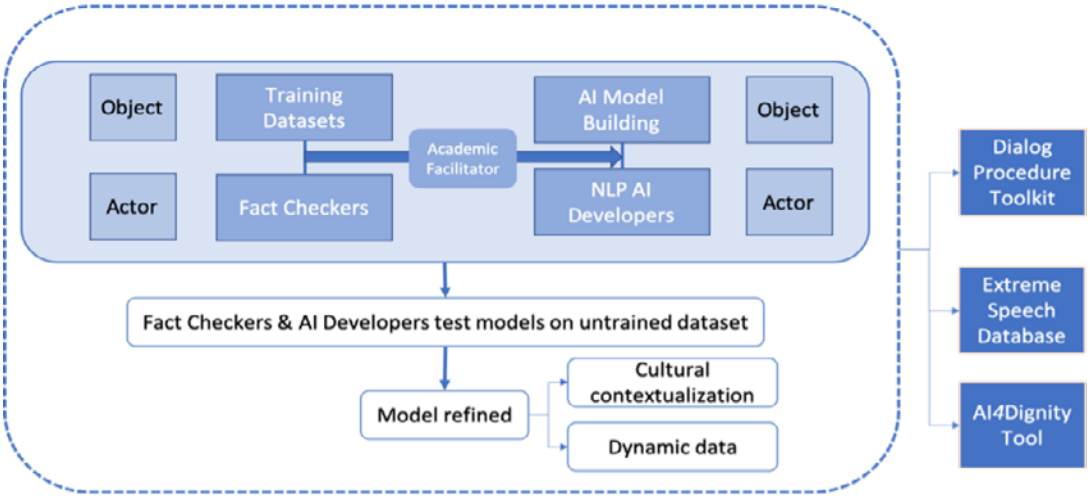

Figure 1: AI4Dignity process model

categories, language or religious affiliation. 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 abstract ideas/ideologies/institutions amount to a call for/imply exclusion of vulnerable groups associated with these categories. Dangerous speech, developed by Benesch (2012), refers to expressions that have reasonable chances to trigger /catalyze harm and violence against target groups (including ostracism, segregation, deportation, and genocide). Benesch, S. (2012). Dangerous speech: A proposal to prevent group violence. New York: World Policy Institute.

45 A policy brief from the AI4Dignity project is available at https://epub.ub.uni-muenchen.de/76087/1/AI4Dignity-AI_Extreme_Speech_Policy_Brief.pdf

This strategy paper urges UN entities to facilitate and fund similar collaborative and curated dialogue spaces with the twin objectives of bringing inclusive datasets to AI models and developing context sensitive human-machine hybrid models. Alongside fact checkers, such initiatives could include other critical communities such as

- online comedians
- anti-hate advocacy groups
- grassroots digital influencers and
- independent journalists

who are active in promoting social justice, resisting repressive states and advocating for inclusive societies. Like fact checkers, these communities are also often the targets of hateful messages in exclusionary populist milieus[46] as well as criticalstakeholders for the cultural knowledge they possess about extreme speech.

In selecting participating organizations, a basic benchmark for "relative autonomy"—in terms of ensuring that they are not part of any political party apparatus or full-time contractors/ employees of social media companies—is important because there are growing trends to politicize fact checking initiatives by forcing fact checkers to fall in line or funding them to toe the party line, or by hijacking and appropriating the very word "fact check" for partisan gains.[47]

Further benchmarks in selecting participants would include representation of diverse linguistic communities, including low resource languages. UN reports have noted that "… in contexts where multiple local languages are used, the UN may not have sufficient capacities to monitor hate speech practices comprehensively. This will add difficulties to UN's work if it does not have the right language skillsets. The use of technology tools such as Natural Language Processing for low resource languages may help mitigate this challenge."[48] The

46 See https://rsf.org/en/news/fact checkers-harassed-social-networks for the Reporters without Borders report on harassment of fact checkers in Brazil, accessed 19 February 2021.

47 In India, fact checking as a growing civil society and business enterprise is showing susceptibility to political and ideological pressure while independent groups continue to assert their autonomy. A large fact checking group, OpIndia, for instance, has declared openly that they do not claim to be "ideologically neutral", and that they will "continue to be right-leaning" (Sharma, 2018). In the UK, media have reported the controversies surrounding the conservative party renaming their Twitter account as "Factcheck UK". In Nigeria, online digital influencers working for political parties describe themselves as "fact checking" opponents and not fake news peddlers. See https://mg.co.za/article/2019-04-18-00-nigerias-propaganda-secretaries/ accessed 19 March 2021. As opposed to heavily funded fact checking initiatives, grounded, community level interventions are critical to fend off ideological heavyweights backed with financial power.

48 Joint Strategy and Plan of Action on Hate Speech, Department of Peace Operations and Department of Political and Peacebuilding Affairs, United Nations.

problem of multiple languages should be addressed on priority by expanding the reach of NLP expertise to low resource languages.

Certification from independent professional associations, where possible (for eg., the International Network of Fact Checkers), is yet another safeguard in the selection process. Procedural guidelines delineated by the DSA in terms of vetting "trusted flaggers" are relevant resources in identifying credible partners (see the section on platform governance). These efforts would be a step towards bringing transparency and social accountability to address algorithmic bias, "black box" issues and lack of traceability in AI decision making, as well as technology gaps in extreme speech detection on a global scale.[49]

COUNTRY LEVEL INTERVENTIONS

Key action frames: specific mix based on the national media ecosystem

Arguably, country level interventions present the most vexing challenge since any list of recommended measures, however dynamic and evolving, cannot be applied uniformly across different countries. Despite the global flow of online content, the influence of regulatory frameworks, legislations, technological infrastructures, media systems, political cultures, linguistic worlds and historical patterns of power precipitate significantly at the national level, defying assumptions that 21st century globalization has eroded the power of the nation-state. Tailoring a mix of measures for online extreme speech requires keen attention to the vast variation in political, cultural and media systems that are institutionalized or consolidated at the national level. This challenge could be tackled on two levels. As a longer-term strategy, a mix of action frames can be delineated after mapping the countries in terms of their "media ecosystems" characterized by structures of power that have stabilized over time and shifting processes of change. This analysis should draw upon comparative research on media systems[50] and journalism cultures[51] as well as "media development indicators" developed by UNESCO.[52]

49 Admittedly, there are more issues with AI-assisted models such as function creep that results in overreach and violation of digital privacy, and growing challenges with multimodal content (audiovisual, mashups, memes, coded expressions) that AI systems are less equipped to detect. These challenges should be addressed by developing due process guidelines in AI deployment for content moderation.

50 Hallin, D. C., & Mancini, P. (2012). *Comparing Media Systems Beyond the Western World.* New York: Cambridge University Press.

51 Hanitzsch et al., 2019.

52 https://en.unesco.org/programme/ipdc/initiatives/mdis accessed 14 February 2021.

As a near term strategy, the World Press Freedom Index by Reporters without Borders (available since 2002)[53] could be used as a broad indicator to assess the impacts of media ecosystems on online extreme speech in different countries. The World Press Freedom Index's questionnaire that now covers 180 countries includes a detailed section on internet enabled media and "cyber-harassment", and is therefore directly relevant for measures around online extreme speech.[54] These indicators should be weighted in relation to internet freedom reports published by independent watchdog groups.

Internet and press freedom indicators and variations in media ecosystem are critical parameters in developing context sensitive platform governance practices and anti-hate initiatives. Based on Hallin and Mancini's model and McCargo's theory, it is possible to consider at least four distinct media ecosystems based on media-politics relationship:[55]

- **Liberal commercialism:** This is characterized by commercial media systems with greater number of profit-oriented media enterprises and higher degree of professionalization of journalism with its own codes of practice.
- **Democratic corporatism:** This is defined by higher degree of professionalization of journalism but also historically strong patterns of party affiliated media. The presence of strong public service broadcasting is an important feature.
- **Polarized pluralism:** Characterized by a "strong prevalence of partisan media, a tendency to instrumentalization of media by political and economic elites, frequent state intervention and involvement in the media system, lesser development of journalistic professionalism and prevalence of clientelism."[56]
- **Pluralist polyvalence:** The media is "situational, with media actors and organizations shifting roles from situation to situation, borrowing from different models, and adapting to changing [political] conditions". Informal links between political elites and media elites are common.[57]

53 https://rsf.org/en/detailed-methodology#:~:text=How%20the%20index%20is%20compiled,journalists%20during%20the%20period%20evaluated accessed 14 February 2021

54 There are drawbacks in these surveys, and therefore, media freedom indices serve only as a starting point to a thorough analysis of the national media system needed for longer-term engagements. For a critical assessment of different media freedom indices, see Schneider, L. (2020) *Measuring Global Media Freedom: The Media Freedom Analyzer as a New Assessment Tool.* Wiesbaden: Springer.

55 This typology needs be finetuned with further research, regularly revised to account for changes, and updated with the latest media freedom indices. The typology draws from the comparative media systems analysis developed by Hallin and Mancini (2012) based on the structure of media markets, the degree and form of political parallelism, journalistic professionalism, and the role of the state. Rather than considering these formations as systems with stable structures, it is more fruitful to approach them as ecosystems with shifting practices and agency of diverse actors yet embedded within distinctive patterns of media-politics relationships that have precipitated over time.

56 Hallin & Mancini 2012, p.279.

57 Hallin & Mancini, 2012, p. 301. See McCargo, D. (2012). Partisan polyvalance: Characterizing the political

It is only by examining distinct dynamics of power in these media ecosystems and their shifting patterns over time that most effective policy measures for extreme speech can be developed. The above longer-term strategy requires research and phased roll out. This paper urges UN entities to support such long-term studies for nuanced measures. At the same time, it is important to take note of the three urgent areas for intervention listed below since these cross-cutting tendencies are seen, with varying degrees of severity, across diverse media ecosystems.

Urgent Priorities

Repressive and authoritarian assaults on online speech

The new millennium decades represent a tumultuous period in history, as the political stunts of populist leaders and everyday activities of millions of online users have repowered small and spectacular spaces of exclusion. The global wide resurgence of right-wing movements, anti-minority and anti-migrant politics in this period reveal a particular precipitation—a political formation that has relied predominantly, if not exclusively, on digital channels. The liberal approach to hate speech builds on the premise that the state and autonomous civil society will uphold democratic values and condemn people who indulge in hateful speech through established processes and institutions of democracy. In the last two decades, the self-declared illiberal and populist authoritarian regimes have unabashedly challenged this logic by unleashing a wave of repressive attacks against advocates for inclusive and just societies. Such regimes have offered a sense of impunity to dispersed online users spewing hatred online. They have also directly or through opaque arrangements enlisted hate speakers and deployed bots to peddle exclusionary narratives, toward a promised secure future for those they consider as proper citizens.

Even more, repressive regimes have increasingly resorted to silencing and intimidating legitimate voices that have raised demands for justice and dignity by channelizing internet enabled media. They have turned regulatory provisions that require social media platforms to share their data, allow inspections and cooperate with authorities for independent audits[58] and similar measures into a weapon to gain control over social media discourses. Turkey's new social media law is one of the latest instances where the regime has imposed advertising

role of Asian media. In *Comparing Media Systems Beyond the Western World* (pp. 201–223). New York: Cambridge University Press.

58 For instance, the DSA proposes that the European Commission and the Digital Services Coordinator (a newly instituted regulatory authority) "may require access to or reporting of specific data. Such a requirement may include, for example, the data necessary to assess the risks and possible harms brought about by the platform's systems, data on the accuracy, functioning and testing of algorithmic systems for content moderation, recommender systems or advertising systems, or data on processes and outputs of content moderation or of internal complaint-handling systems within the meaning of this Regulation."

bans on social media companies such as Twitter, Pinterest and Periscope for failing to appoint a local representative to take down contentious content.[59] The regulatory measure to oblige social media companies to appoint "compliance officers"—a measure that appears benign and progressive in established democracies—has served as a means to clampdown on dissenters in these contexts. Scholars have shown that the "[COVID 19] pandemic has only accelerated and amplified the effect of the so-called 'anti-fake-news' laws on a global level as governments from Romania to Botswana emulate scare tactics seen in Singapore and Malaysia."[60] Not only do these sweeping rules lack due diligence processes, but they have also widened the ambit of control by including online news services and video streaming. Such measures have raised concerns about over-censorship and the undermining of end-to-end encryption, reminding yet again that EU and other regulations in the West that are cited as "global models" set a precedence for stricter controls, but they serve as a double-edged sword and a direct weapon of repression in contexts where democratic rights are under attack.

Internet shutdowns are another commonly used repressive tool, often putting entire populations under complete online lockdown. For instance, digital rights organizations have reported the consequences of severe forms of internet censorship in Iran.[61] A study conducted by a team of computer scientists at the University of Michigan in 2020 noted "increasing censorship activity in more than 100 countries" spanning Sudan, Sri Lanka, Norway, Zimbabwe and other countries.[62] The study has identified eleven categories of websites facing increasing censorship, including websites that contained content related to human rights issues and news media.

State complicity in weaponizing digital technologies against vulnerable and minoritized populations has emerged as another grave concern. If facial recognition and "emotion recognition" have become the new tools in the far-reaching infrastructures of state surveillance and repressive clampdown on resistance in countries like China,[63] the state

59 https://www.aljazeera.com/news/2021/1/19/turkey-slaps-advertising-ban-on-twitter-with-new-social-media-law accessed 16 February 2021.

60 Ong, J. (2021). Southeast Asia's information crisis: Where the state is the biggest bad actor and regulation is a bad word. SSRC Items. 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 20 February 2021.

61 https://www.article19.org/data/files/medialibrary/38315/The-National-internet-AR-KA-final.pdf accessed 4 March 2021.

62 https://dl.acm.org/doi/10.1145/3372297.3417883 accessed 5 April 2021

63 In a detailed study on China's application of behavior recognition technologies, British digital rights watchdog Article 19 has shown that behavior recognition applications are deployed to cover a wide breadth of governance including credit worthiness, criminal behavior as well as student attentiveness inside the classroom. https://www.article19.org/wp-content/uploads/2021/01/ER-Tech-China-Report.pdf

has been conducting a long-standing campaign of violence and extreme speech against the Rohingya minorities in Myanmar, using state owned media and internet platforms to further perpetrate this aggression.[64]

In such situations, UN entities should partner with civil society monitoring groups and global digital rights organizations to increase policy pressure and raise awareness, aside from engaging with governments directly. In the EU context, scholars have argued that European institutions and Member States "should be obliged by EU law to withdraw funding provided to and prohibit political coalition with political parties and other organisations whose members repeatedly represent views that are irreconcilable with the values of the European Union, provided that the party or other entity fails to sanction this."[65] UN entities should strengthen similar policy and diplomatic tools for global monitoring and regulatory action in Member States, including capacity building of key State actors,[66] monitoring the deployment of AI-assisted technologies, and empowering CSOs and research networks to strengthen "lobby initiatives that can maneuver around repressive regimes". [67]

A significant step is to enlist the support of social media companies in scenarios of state repression, extralegal intimidation, and political misuse of legal provisions. Interestingly, in Southeast Asia, studies have found that a few global social media companies were active in multi-stakeholder initiatives for election integrity at a time when state led targeted assaults against regime critical voices continued apace.[68] In countries like India, corporate actions against extreme speech have vacillated between platforms' efforts to hold on to their "global standards" in hate speech policies and direct complicity in the ideological politics of the ruling regime.[69] Often shaped by the commitments, prudence, and political leanings of local level executives (as pointed out earlier), corporate responsiveness has been ambivalent and

accessed 28 January 2021.

64 Lee, R. (2019). Extreme speech in Myanmar: The role of state media in teh Rohingya forced migration crisis. *International Journal of Communication,* 13, 3203–3224.

65 Bayer, J., & Bárd, P. (2020). *Hate speech and hate crime in the EU and the evalution of online content regulation approaches.* Brussels: Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, p. 13

66 The UN Strategy and Plan of Action on Hate Speech (September 2020) has highlighted capacity building for key State actors, "notably judges, judicial personnel (such as prosecutors and court officials), law enforcement agents and members of the security forces on international human rights norms and standards relating to hate speech, especially the standard of incitement to discrimination, hostility or violence that amounts to a criminal offence (as indicated in the Rabat Plan of Action)."

67 Ong, 2021.

68 Ong, 2021.

69 https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346 accessed 16 February 2021.

uncertain. Recent news reports in India have cited internet whistleblowers who have exposed Facebook's "double standard" in enforcing content takedown policies, and how the company has been lenient towards fake accounts and fake engagement that are backed by powerful politicians.[70] Following international outcry, Facebook banned Myanmar's commander-in-chief and military officials from its platform after admitting it was "too slow" to respond to the concerns of UN officials and human rights advocates.[71] Twitter took a similar action against Donald Trump in the US, while also raising questions about corporate veto power.

In cases of coordinated attacks or authoritarian controls over platform regulations, and dramatic turmoil when companies feel the pressure to take swift actions at the cost of due diligence processes, UN entities should apply pressure on, and if necessary, guide social media companies to comply with global standards of content moderation and human rights protection by offering procedural clarity around escalation protocols and decision making. A significant step is to recognize that "individualized and user-focused enforcement models" that social media companies have adopted are not sufficient under repressive conditions.[72] In Myanmar, for instance, government sponsored genocide campaign made hateful speech against the Rohingya Muslims so widespread and pervasive that taking down content based on "individual instances of hate speech" would be neither feasible nor effective.[73] In cases of "pervasive hate directed to vulnerable populations", policies should draw guidance from "both international human rights law and international law on remedies" and encourage companies to "proactively police their networks for coordinated speech campaigns against vulnerable groups, in conditions that might indicate such speech could contribute to impunity for violence"[74] (see also the point on risk mitigation under "platform governance"). In addition, design features of social media companies that encourage polarized content through algorithmic mediations should be brought into periodic scrutiny. An institutionalized global structure to regularly convene social media companies to address repressive assaults on online speech will be a significant step towards addressing upheavals that unfold at the national levels. Convening different social media companies is especially critical during the elections since disinformation campaigns funded by resource rich political parties have begun to increasingly adopt cross-platform manipulation tactics. These measures should go hand in

70 https://www.theguardian.com/technology/2021/apr/15/facebook-india-bjp-fake-accounts?CMP=Share_iOSApp_Other accessed 22 April 2021.

71 Land, M. K. and Hamilton, R. L. (2020). Beyond Takedown: Expanding the toolkit for responding to online hate. In Dojcinovic, P. (ed.) *Propaganda, War Crimes Trials and International Law: From Cognition to Criminality.* London: Routledge. Also available as Research Paper No. 2020-11 at http://dx.doi.org/10.2139/ssrn.3514234.

72 Land & Hamilton, 2020, p. 2.

73 Land & Hamilton, 2020, p. 5.

74 Land & Hamilton, 2020, p. 3

hand with community level and creative actions against online hate listed in the next section.

In summary

Measures to address repressive attacks against online speech and coordinated hate campaigns:

- Applying pressure on, and if necessary, guiding social media companies to comply with global standards of content moderation and human rights protection by offering procedural clarity around escalation protocols and decision making.
- An institutionalized global structure to regularly convene social media companies to address repressive assaults on online speech.
- Partnering with civil society monitoring groups and global digital rights organizations for awareness raising and to increase policy pressure for platform governance
- Capacity building of key State actors such as judges and the judicial personnel to sensitize them about sound practices of platform governance
- Monitoring the emulation of stricter regulatory models implemented in developed economies with stable democracies for repressive purposes in other national contexts
- Monitoring the deployment of AI-assisted technologies by state actors

Gray zones, fringe actors and smaller/domestic platforms

Political manipulation of online discourse through algorithmic and computational affordances has become the new face of electoral propaganda globally, but especially in the global South, partisan politics has spawned a breeding shadow industry that operates through gray practices of clickbait operators, hired influencers, and loosely knit networks of dispersed amplifiers who are drawn into precarious and informal labor arrangements crafted by ambitious mediators. In several countries, political actors and ruling governments are directly engaging and sponsoring such practices.

In India, a significant part of such arrangements is is carried out by the "unofficial" wings of the political party campaign systems, which in turn attach to the "official" party structure through "third party pages" and opaque arrangements for "service delivery" (euphemism for manipulating online discourses). Among other things, these "unofficial" and dispersed networks are encouraged to "innovate" on campaign content both in terms of divisive messaging and disinformation. Lately, these gray zones have gone mainstream by synchronizing messaging across YouTube, Twitter, Facebook, and other social media platforms, and relaying open threats

to independent media and civil society activists. Heavily funded campaigns have engaged transnational data analytics expertise as well as a motley mix of domestic players who present themselves as "politically agnostic" digital consultants and promise "data-tested solutions" of tracking and calibrating voter sentiments for electoral success. The landscape of political extreme speech therefore has a wide range of actors—precarious business entrepreneurs who strive to make a livelihood with petty promotional work to ideologically motivated volunteers who expect no monetary compensation to self-styled professional entities that have branched out from business process outsourcing (BPO) and the Information Technology industry by spotting "data tested" digital influence as the next big business opportunity.[75] These arrangements have augmented the conditions for political actors to peddle politically expedient and ideologically driven online vitriol. Studies have gathered evidence for the ways in which the ruling right-wing political party exploited social media narratives by using thousands of WhatsApp groups with dispersed volunteers and "loosely affiliated online supporters" to engage in 'trending' campaign friendly hashtags on Twitter.[76] Through this "cross-platform media manipulation" tactic, "hundreds of trends were fabricated" during the elections.[77] These trends were later picked by other media outlets leading to amplification of the ruling party's campaign line. It is important to stress that such cross-platform manipulations necessitate simultaneous policy actions across different social media companies, and hence convening different social media platforms for regular discussions especially during important events such as elections (as pointed out in the section) is critical.

Similar to the variegated scenario of precarious labor, electoral campaign manipulations, ideology-led aggression, political opportunism, and commercialization of divisive digital content seen in India, studies have exposed "networked disinformation" that functions "as a distributed labor of political deception to a hierarchy of digital workers" in the Southeast Asian context.[78] These studies have revealed that the precarious labor conditions of disinformation workers who serve political clients by engaging in project based digital work are characterized by "race-to-the-bottom" work arrangements. These "casual workers" are forced to cope with stressful work on their own, "in the absence of clear guidelines, psychosocial support systems,

75 Udupa, S. (2019). India needs a fresh strategy to tackle online extreme speech. *Economic and Political Weekly Engage.* https://www.epw.in/engage/article/election-2019-india-needs-fresh-strategy-to-tackle-new-digital-tools

76 Jakesch, M., Garimella, K., Eckles, D. and Naaman, M. (2021). #Trend Alert: How a Cross-Platform Organization Manipulated Twitter Trends in the Indian General Election. J. ACM XX, XX, Article XXX (April 2021), https://doi.org/TBD, p. xxx

77 Jakesch, M., Garimella, K., Eckles, D. and Naaman, M. 2021, p. xxx:2

78 Ong, J. C., & Cabanes, J. V. (2018). *Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines.* www.newtontechfordev.com

or remuneration".[79] Therefore, the "chief architects of networked disinformation who intimidate dissenting voices and craft new prospects for political clients via digital influence are themselves precarious architects of precarious labor arrangements in the creative industries that make workers vulnerable to slipping to the underground".[80] Precarious and opportunistic arrangements that characterize a large number of Rodrigo Duterte's "trolls" are similar to political rivalries that have led to an online army of "AKTrolls" who marshal support for Recep Tayyip Erdogan in Turkey by flooding online discussions with allegations, counter-allegations, rumors, and lies.[81]

Similar practices of online propaganda and threats to election integrity are widespread in Africa. In the Zimbabwean elections in 2018, the two major political parties hired "online warriors"—a combination of bots and actual people (paid or volunteering youths)—to manufacture and disseminate party propaganda on Twitter, Facebook, and WhatsApp."[82] Sobriquets and name calling were rampant on social networking sites. Key presidential contender Nelson Chamisa's followers nicknamed as "Nerorists" and the other contender Mnangagwa's followers nicknamed as "Varakashi" acted as "cyber storm troopers" to push their respective leader's propaganda.[83] Such tactics included spreading false news and rumor, the most dramatic of which was to raise a suspicion that the Zimbabwe Electoral Commission was biased and lacked the credibility as a neutral arbiter. For example, an app bearing the logo of the Commission that invited the users to "click to vote" spread rapidly on WhatsApp. "But, responding to the prompt led to a message congratulating the user on voting for Mnangagwa, suggesting that the supposedly independent electoral body had endorsed the Zanu-PF leader".[84] Such messages that enticed the users to click innocent looking buttons on handheld gadgets delivered disinformation-based propaganda by making use of the mundane and compulsive habits of digital communication.
In Nigeria, studies have shown that online hate speech is a "major driver of election violence", revealing that social media platforms such as WhatsApp have become the new battlegrounds for online hate campaigns.[85] Especially during election times, false rumors have reached

79 Ong & Cabanes, 2018, p. 29

80 Ong & Cabanes, 2018.

81 Saka, E. (2018). Social media in turkey as a psace for political battles: AKTrolls and other politically motivated trolling. *Middle East Critique,* 27(2), 161–177.

82 https://theconversation.com/a-vicious-online-propaganda-war-that-includes-fake-news-is-being-waged-in-zimbabwe-99402 accessed 3 March 2021.

83 Chibuwe, A. (2020). Social Media and Elections in Zimbabwe: Twitter War between Pro-ZANU-PF and Pro-MDC-A Netizens. *Communicatio: South African Journal for Communication Theory and Research* 46(4): 7–30.

84 https://theconversation.com/a-vicious-online-propaganda-war-that-includes-fake-news-is-being-waged-in-zimbabwe-99402 accessed 3 March 2021.

85 Ezeibe, C. C., & Ikeanyibe, O. M. (2017). Ethnic politics, hate speech and access to political power in

the threshold of dangerous speech by "terrifying Nigerian Christians with predictions that Muslims plan to kill, rape, and subjugate them".[86] Availability of affordable smart phones and messenger applications with light data usage has facilitated the spread of manipulated content. For instance, studies have shown how "…in early January, a message entitled 'Fulani War Threat' circulated on WhatsApp. It was falsely presented as an English translation of a pamphlet in Arabic disseminated to mosques in northern Nigeria by a group called FUNAM, or the Fulani Nationalist Movement. Staff at the outstanding civil society groups…investigated and found that the group doesn't exist. The fake "pamphlet" that circulated online seemed to call on Muslims to boycott the elections, and instead to prepare for holy war."[87] These patterns resemble trends in South Asia where religious identity is misused as the key driver of politically manipulated digital disinformation.

Burundi's political crisis witnessed an increase in the radicalization of the regime since the second post-conflict elections in 2010, which escalated especially in 2015 following the late President Pierre Nkurunziza's bid for a third presidential term. This led to the growing influence of hardline leaders of the ruling party who sought to undermine the Arusha accord— "an agreement between Hutu and Tutsi elites in 2000 that put in place an ethnic quota system for state institutions, including the army, and established a two-term presidential limit."[88] Social media provided a means for communication for citizens and journalists to "to coordinate and inform the international community about the conflict"[89] in the wake of repressive attacks on radios, which were the most common source of information for Burundians prior to social media expansion. However, the Nkurunziza regime also sought to gain control over social media discourse through orchestrated hateful messaging. Studies have shown that the Twitter account of the presidency's spokesperson sent out highly charged messages against political rivals, dubbing them as "terrorists", as well as attacking journalists and sending out hateful messages against Rwanda.[90] At the same time, the presidency's Twitter account sought to maintain the decorum, displaying stability and legitimacy to an international audience.[91]

Nigeria. *Africa Today,* 63(4), 65–83; also https://mg.co.za/article/2019-04-18-00-nigerias-propaganda-secretaries/ accessed 3 March 2021.

86 Benesch, S. (2021). Nigerian politician's dangerous lives at risk on the eve of 2019 elections. https://dangerousspeech.org/nigerian-politicians-dangerous-speech-puts-lives-at-risk-on-the-eve-of-2019-elections/ accessed 3 March 2021

87 Benesch, 2021.

88 https://www.crisisgroup.org/africa/central-africa/burundi/burundi-dangerous-third-term accessed 5 March 2021

89 Dimitrakopoulu, D., & Boukala, S. (2018). Exploring democracy and violence in Burundi: A multi-methodical analysis of hegemonic discourses on Twitter. *Media, War & Conflict,* 11(1), 125–148, p. 126.

90 Dimitrakopoulou & Boukala, 2018.

91 Dimitrakopoulou & Boukala, 2018.

This kind of "forked tongue" extreme speech—where the country's leaders offer an image of stability by maintaining expected decorum to appease the international community, while simultaneously allowing and encouraging their subordinates to spew hatred online—is not uncommon in other national scenarios. The 2020 UN report on the strategic assessment mission for United Nations engagement in Burundi has observed that the 2020 elections in Burundi that led to the change in the leadership were peaceful, but has also cautioned that the "prevailing context remains fragile" because of "concerns over the overwhelming control of the ruling party…in state institutions" while also noting the "spirit of openness" displayed by the Ndayishimiye presidency.[92] The report has also noted concerns expressed by the opposition and civil society organizations about the "marginalization and silencing of independent media"[93] and therefore, the evolving role of social media in this changed scenario remains to be watched.

Similar to the weaponization of "fake news" as a tool to discredit media and political opponents in other parts of Africa, political actors and government officials in South Africa have raked up the trope of disinformation for political gains. Worryingly, such manipulations have revived the term "Stratcom"—the disbanded propaganda disseminator of the apartheid regimes—as ways to discredit political opposition. In conjunction, heavily funded clandestine social media propaganda activities have emerged as the new face of the corrupt nexus between the political class and crony capitalism. A family of wealthy entrepreneurs who were closely linked to the scandal ridden leadership of South African President Jacob Zuma hired the infamous UK PR firm Bell Pottinger to spread a racially divisive narrative on social media through bots and hired amplifiers. Online troopers flooded social media discussions with distorted news sites, shared stories, networks of websites, and retweets to pump up the momentum and tilt the discourse in favor of the regime.[94] Social media manipulations "worked in tandem" with TV and print media owned by the same family. In a bizarre turn of discourse, the paid social media campaign of this camp appropriated the left progressive term, "white monopoly capital", to hide its corrupt deals by whipping up a distorted story of racial injustice.[95] Although Zuma's regime ended with his resignation in February 2018, following which manipulated social media and other propaganda machineries became public, and Bell Pottinger went bankrupt

92 https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/287/21/PDF/N2028721.pdf?OpenElement, pp.4–5 accessed 10 August 2021.

93 https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/287/21/PDF/N2028721.pdf?OpenElement, p. 10.

94 https://techcentral.co.za/go-inside-guptabot-fake-news-network/76767/ ; https://www.timeslive.co.za/news/south-africa/2017-09-04-how-the-gupta-campaign-weaponised-social-media/ https://www.bellingcat.com/news/africa/2017/08/04/guptaleaks-google-analytics/ accessed 2 March 2021.

95 https://www.nytimes.com/2018/02/04/business/bell-pottinger-guptas-zuma-south-africa.html accessed 2 March 2021.

soon after, disinformation campaigns among different political rivals still pose a challenge to electoral integrity and democratic systems in the country.

In Saudi Arabia, media reports have revealed that regime supported "troll farms" are tasked with targeting dissenting voices. US based media reported that troll farms consist of "'social media specialists' who operate via group chats in apps like WhatsApp and Telegram, sending them lists of people to threaten, insult and intimate; daily tweet quotas to fill; and pro-government messages to augment."[96] Similar to disinformation campaigns in Africa, South Asia and East Asia, other tactics have involved creating derogatory memes, publishing pornographic images to distract from contentious issues and coordinated messaging online. Media accounts have also reported that disinformation campaign agents regularly hijack deceased people's social media accounts,[97] aside from implanting regime supported spies inside social media companies.[98] In media interviews, political opponents have alleged that the Saudi regime recruited highly reputed global corporate consultancy firms to identify Twitter influencers, and later intimidated these influencers to submission or arrested them, forcing some to go into exile.[99]

Across diverse political scenarios, especially in the global South, an anatomy of vicious campaigns that mobilize offensive, hateful, and even dangerous speech to suppress dissent or gain voter loyalties reveals that top-down propaganda machineries of political parties or ruling regimes ramp up and coordinate closely with bottom-up enthusiasm among volunteers, leading to blurred online arenas of sponsored, volunteered and manipulated content.

The intricate networks and gray zones through which online extreme speech circulates stress the need for policy measures that go beyond the focus on social media companies and the underlying assumption that regulatory control over big social media companies would solve a complex social problem. Community level awareness programs and rapid response systems that are sensitive to diverse social conditions of digital hate cultures are critical in addressing this challenge (see the next section on community level interventions).

At the same time, complex realities on the ground require fine tuning platform governance policies. The principle of proportionality enunciated by the EU is especially inadequate in this

96 https://www.nytimes.com/2018/10/20/us/politics/saudi-image-campaign-twitter.html accessed 2 March 2021.

97 https://english.alaraby.co.uk/english/indepth/2019/2/25/saudi-trolls-hacking-dead-peoples-twitter-to-spread-propaganda accessed 4 March 2021.

98 https://www.nytimes.com/2019/11/06/technology/twitter-saudi-arabia-spies.html accessed 4 March 2021.

99 https://www.washingtonpost.com/gdpr-consent/?next_url=https%3a%2f%2fwww.washingtonpost.com%2fopinions%2f2019%2f11%2f14%2fsaudi-spies-hacked-my-phone-tried-stop-my-activism-i-wont-stop-fighting%2f accessed 4 March 2021.

regard. the DSA's (2020) reasoning that fostering the growth of smaller players will provide a "level-playing field against providers of illegal content"[100] is based on a limited understanding of how hateful speech is shared online. Without doubt, large multinational social media companies play a major role in amplification, and therefore stricter regulations are warranted. However, this should not turn attention away from smaller and niche platforms that have also emerged as a breeding ground for hateful subcultures. Ethnographic work has shown that hate speakers have used or repurposed smaller platforms by hopping between them to avoid the regulatory gaze. Online games, 4Chan, reddit communities and other niche spaces have spawned toxic subcultures of hate through digitally native formats of mashups, memes, and playful interactional frames. The exodus of right-wing conservatives and Trump supporters to Parler, a self-styled free speech platform, after facing the ban on Twitter, offers yet another example for the susceptibility or collusion of smaller platforms in exclusionary extreme speech. Factcheckers in Brazil have noted a similar migration of right-wing supporters to Parler since July 2020, but after Apple and Google blocked the app on their app stores, access to the platform has been restricted, forcing several users to retreat to big platforms or search for newer ones. Faced with regulatory actions, violent Jihadi groups similarly moved to encrypted channels such as Telegram or file-sharing sites such as Pastebin, and the extreme right migrated to platforms such as VK.com or Gab.ai.[101] A recent study in Europe has shown that anti-establishment right-wing celebrities migrated to Telegram and to a "larger alternative social media ecology" after being "deplatformed" by major social media companies such as Facebook, Twitter and YouTube for "offenses such as organized hate".[102] The network graphs that mapped the connections between right-wing celebrities and platforms revealed the prominence of "BitChute (alternative to YouTube), Minds (alternative to Facebook), Gab (alternative to Twitter) as well as Telegram (hybrid messaging and broadcasting platform).[103]

These realities expose the limits of the proposed EU legislations (DSA and DMA) and the premise that breaking open the "centralized platform economy" would solve the problem of virality and amplification of online illegal content. The focus therefore should be extended to smaller social media platforms that offer niche spaces for online hatred to proliferate through regional language friendly applications, platform specific jargons and practices, and

100 https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

101 It is argued that three formal features of digital hate cultures make them ungovernable: swarm structure characterized by decentralized networks; exploitation of inconsistencies in web governance between different social media companies as well as between private and government actors that allows hate content to migrate when detected; and the use of coded language to evade content moderation Ganesh, B. (2018). The ungovernability of digital hate culture. *Journal of International Affairs,* 71(2), 30–49.

102 Rogers, R. (2020). Deplatforming: Following extreme internet celebrities to Telegram and alternative social media. *European Journal of Communication,* 35(3): 213–229.

103 Rogers, 2020, p. 219.

lax regulatory attention. This is especially pertinent in the context of the global South. Media reports and ethnographic studies carried out by this author have documented that executive of smaller companies tend to develop cozy alliances with resource rich political parties or act as their proxy representatives, or otherwise, for commercial interests, parade their technologies before prospective clients regardless of their political ideologies to win more clients. They also seek to profit from platform migration when users, faced with blocking and other content moderation actions, leave large social media companies in search of unmoderated platforms. While not all of them directly monetize extreme speech, it is important to monitor how they are evolving.

Several smaller startups in India provide regional language messaging services, allowing their platforms during the election time for partisan messaging and manipulations through intricate networks of lobbying and buyouts. For instance, in early 2021, after Twitter took action against right-wing hateful speech on its platforms by suspending and blocking many handles, and resisted government's requests to stop publishing posts that were critical of its policies, religious majoritarian voices rushed to secure their spot on "Koo App", a homegrown platform founded in March 2020.[104] Koo App offers services not only in English but also in five Indian languages. Pro-government television channels promoted it as "the best Twitter alternative for Indians" and publicized the so-called "trending hashtags" on the new App.[105] French ethical hacker Robert Baptiste raised concerns that the App exposed users to high vulnerabilities, as he was able to easily access personal data of users.[106] ShareChat is another Indian social media startup that currently has 160 million monthly active users and operates in 15 Indian languages and not in English. Its opaque content moderation policies and lack of regular transparency reports (except during the 2019 general elections in India)[107] have raised concerns about the risk of unregulated circulation of problematic speech on its platforms. In 2020, ShareChat launched the short-video platform "Moj" only days after TikTok was banned in India.[108] Moj's terms of use mention that the company "may share [user] information with appropriate law enforcement authorities if [they] have good-faith belief that it is

104 https://www.aljazeera.com/economy/2021/2/17/angry-bird-twitters-india-troubles-give-local-rival-koo-a-lift; https://time.com/5935003/india-farmers-protests-twitter/; https://www.nytimes.com/2021/02/10/technology/india-twitter.html; https://www.bbc.com/news/world-asia-india-56007451 accessed 3 March 2021.

105 https://www.bbc.co.uk/news/world-asia-india-56037901 accessed 3 March 2021.

106 https://www.livemint.com/news/india/french-security-researcher-accuses-koo-of-leaking-user-data-11613022668933.html accessed 3 March 2021.

107 https://medium.com/sharechat/our-efforts-ahead-of-the-general-elections-2019-sharechat-76aa393a4b9a accessed 2 March 2021.

108 https://www.cnbc.com/2020/09/28/indian-start-up-sharechat-is-one-of-many-looking-to-fill-the-vacuum-left-by-tiktok-ban.html accessed 2 March 2021.

reasonably necessary to share your personal data or information in order to comply with any legal obligation or any government request."[109] In the current context of tighter regulatory controls over online discourse, such explicit terms of use ingrained in homegrown startups could contribute to further restrictions on open discussions. In Africa, smaller platforms such as IMO, Likee and Vskit are gaining popularity, while more research is awaited to map their political impact.[110] In Southeast Asia, popular platforms such as Line (in Thailand) and Viber (in the Philippines) are shown to be the "cesspools of dis-and misinformation" but they have consistently avoided participating in regulatory discussions.[111]

Although the EU proposal to require very large platforms (Facebook, Twitter and YouTube) to open up to competitors with mandatory interoperability is appropriate in the pursuit of anti-competition policy objectives, this would not be an obvious solution to fix the problem of hateful content. This proposal assumes that regulating large platforms and fostering smaller players would lead to a scenario where "users could freely choose which social media community they would like to be part of—for example depending on their content moderation preferences and privacy needs—while still being able to connect with and talk to all of their social friends and contacts". [112] This approach comes with a liberal baggage, and more gravely, the dangers that this very marketplace for ideas could provide easy ways for hate mongers to hop between the platforms. Even more, political vested interests are likely to invest and drive the market of multiple smaller players toward partisan messaging. The proposal for interoperability, in other words, should be combined with a host of other measures to address proxy campaigning. In several cases of repressive and authoritarian conditions, global corporations with some stable moderation practices and resources have been more responsive to implementing safeguards against online hate and illegal content than the sundry mix of unregulated platforms that operate in the gray zones. Although this strategy raises the risk of framing "corporate capital as an explicit ally in the struggle for a just world",[113] the point emphasized here is that the dangers of small platforms that are equally as capitalized require some hardheaded policy actions. Such measures should also not ignore the potentiality of smaller platforms to enable progressive alternative discourses and user autonomy through open-source protocols and low capital ventures.

Especially where ruling regimes are directly involved in intimidation and disinformation campaign,

109 https://help.mojapp.in/policies/terms accessed on 4 June 2021.

110 https://www.usiu.ac.ke/assets/image/Kenya_Social_Media_Lanscape_Report_2020.pdf accessed 5 March 2021

111 Ong, 2021.

112 EDRi. (2020). *Platform regulation done right: EDRi position paper on the EU Digital Services Act.* Brussels: European Digital Rights, p.4.

113 Flood, D. (2019). Responding to 'Fake News' in an Era of Hashtag Leftism. *Anthropology News,* 29 January.

engaging with large social media platforms would be necessary to implement some pushback mechanisms. For instance, in late December 2019, Facebook/Instagram and Twitter announced that they suspended accounts for their "coordinated inauthentic behaviour" and "state-backed information operations" of Saudi Arabia (Facebook detected a second campaign from UAE and Egypt too).[114] As a preventive strategy, platforms should also create interfaces that "nudge users toward responsible speech choices" and tweak their algorithms to "deprioritize particularly extreme or virulent content".[115]

In summary

Measures to address repressive attacks against online speech and coordinated hate campaigns:

- Engaging the "Big Tech" is critical but policy measures should recognize that regulatory control over big social media companies would not fully solve a complex social problem
- Monitoring and supporting compliance to global standards among smaller platforms and niche internet spaces is important because hate speakers have used or repurposed smaller platforms by hopping between them to avoid the regulatory gaze
- Mobilizing community level awareness programs and rapid response systems that are sensitive to diverse, country specific social conditions of digital hate cultures is critical (see more under 'community level interventions')
- Convening self-styled political trolls, local level politicians, and commercial digital influencers for awareness raising activities, and sensitizing them about global human rights standards and the dangers of digital campaign manipulations by sharing the narratives of those who have been severely affected by digital hate both locally and in other parts of the world is another area of intervention
- Where ruling regimes are directly involved in intimidation and disinformation campaigns, engaging with large social media platforms for necessary regulatory actions against coordinated attacks is necessary. In some cases, big social media platforms have participated in responsible governance efforts

114 https://about.fb.com/news/2019/08/cib-uae-egypt-saudi-arabia/; https://blog.twitter.com/en_us/topics/company/2019/new-disclosures-to-our-archive-of-state-backed-information-operations.html accessed 4 March 2021.

115 Land, M. K. (2018). Speech duties. *The American Journal of International Law Unbound,* 112, p. 329.

more enthusiastically than ruling governments that are eager to seize online discourse to advance their own goals. In cases where regimes are adamant, social media platforms provide an important gateway for intervention. Platforms should undertake measures not only in the liberal democracies of the West where policy pressure on content moderation is high but also in other parts of the world to act against illegal content and encourage users to adopt responsible speech practices.

Gender-based abuses

Gender-based abusive trolling is a particularly virulent form of online extreme speech and a disturbing trend that cuts across diverse cultures with vastly different levels of protection and opportunities for women and sexual minorities. Publicized in some cases and insidious in others, trolling attacks have involved targeting women and LGBTQI+ public figures active in politics, social voluntary work or journalism, especially those who are pushing back against regressive right-wing discourses. Seeking to delegitimize daring female voices, such abuses typically invoke the image of illicit sex and prostitution in proses and sexist epithets that sometimes reveal their preset formats. Far from the grounds of veracity, such abuses gain valence through repetition and reverberation. Allegations against women public figures for sleeping with male politicians are not meant to compete on grounds of truth, but as reverberations that could exhaust the targets. This form of abuse can be recognized as "evaluative talk" that invokes practices of "verbal obscenity" and emerges from "specific cultural systems of moral judgment". [116]

Digital anonymity is often cited as the reason for such abuses since it is assumed to catalyze "deindividuation" with reduced terms of self-evaluation. However, a closer focus on local abuse cultures reveals that abusers operate upon local knowledge and it is quite common for victims to second guess who the abuser is. Under conditions where abusers feel they have political immunity and the backing of the regime, vitriol turns into direct threats that attack social and personal security with precise knowledge of the target's life routines and lifestyles. These online messages, for example, would name the child of the woman online commentator and the time her child would go to the school on a particular route. Attackers have also threatened independent women journalists and political commentators with gangrape and acid attacks, often in collusion with political forces on the ground.

116 Butler, J. (1997). *Excitable Speech: A Politics of the Performative.* New York: Routledge, p.109.

In various ways, political extreme speech that comes in the form of gender-based abuse draws from a broader online culture of misogyny rampant in online game cultures[117] and darker niches of internet channels spanning image boards, closed online communities as well as publicly available social media accounts run by individuals and groups. It is often identified as the "Manosphere", a vast online space consisting of inceldom ("involuntary celibates"); the RedPill community that believes in the conspiracy that men are trapped in an illusion that women have created to perpetuate "male servitude"; MGTOW ("Men going their own way") who reject engaging with women; and Men's Rights Activists who vow to protect their freedoms against the "feminist enemy".[118] Although such phenomena are seen as West-centric, they have reverberated across the globe with various mutations through online games, chatrooms, imageboards, online pornography and image/video sharing social media apps. In some instances, broad anti-women ideologies have dovetailed with more personalized forms manifest as "revenge porn" often crafted by dejected/rejected lovers who make intimate scenes public, as well as setting up fake profiles and doxxing that ride on online voyeurism and sexual innuendo. In southern India, a case was registered against online trolls who shared videos of a schoolteacher with derogatory and sexist epithets when the school released a video of her teaching a class to facilitate online learning during the COVID-19 pandemic. Such brazen misuse of videos available easily on the internet has raised concerns about privacy and safety of women, as materials that float on the net have become an easy resource for stalking, eavesdropping, rumor mongering, threats, and extortion.

In other instances, gender-based abuses do not always contain explicit derogatory content but they seek to assert control over political discourse through masculinist interpretations of female modesty. In Indonesia, partisan politics has instrumentalized discourses around "womanhood", enabling religious actors and political groups to fabricate instances of violation of religious laws and ethical norms.[119] These groups in turn have raised swarming online armies

117 For more on the Gamergate controversy, see Massanari, A. (2015). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media and Society,* 19(3), 329–346.

118 Csuka, L. (2020). "The ideological structure and persuasive force of r/TheRedPill", Masters' Thesis, LMU Munich. Csuka cites instances where the conceptual apparatuses of these online communities have translated into physical attacks in the actual world: "Alek Minassian conducted a terror attack in Toronto by driving a van into a crowd, killing ten people and wounding several more in April 2018. Prior to his actions, he announced on 4chan his admiration for Elliot Rodger who in 2014 killed six people and injured thirteen in Isla Vista. In November 2018, Scott Beierle shot two women in a yoga studio in Tallahassee. All three of them identified with and credited their actions explicitly to the conceptual origin of inceldom, a community from the Manosphere."

119 Pratidina, I. (2021). "Motherhood" revisited": Pushing boundaries in Indonesia's online discourse. In S. Udupa, Gagliardone, I., & Hervik, P. (Eds.), *Digital Hate: The Global Conjuncture of Extreme Speech.* Bloomington: Indiana University Press.

against women public figures and minorities. Similarly, anti-immigrant discourses in the UK have resorted to gender-based attacks against hijab wearing Muslim women commentators online, citing the public goal of women empowerment in twisted ways to further a racist, Islamophobic agenda. In India, perpetrators of gender-based abuse have claimed legitimacy by invoking the high ideals of nationhood and how their contestations with "pseudoliberal" women journalists, activists, and politicians are meant to serve the high ideal of cleaning the nation from corrupt ideas. Such practices draw on a longer history of articulating patriotism in relation to "moral restraint", which partly unfolds through a conservative politics focused exclusively on regulating sexuality.[120] In some cases, abuse escalates to a full-blown shaming punishment, where online networks of swears and accusations create a bounded arena for shaming sanctions that fall "most heavily on women in terms of governance of sexuality". In all these cases, gender-based abuses serve as a ground to advance political vigilantism thick with moral language.

Social consequences of shaming punishments have been evident in several high-profile Twitter wars in recent years involving politicians, cinema stars, activists, and sports celebrities. Their private details have been leaked, and false accusations made on their "moral" behavior. Ethnographic work bears evidence to how shaming has been emotionally taxing and tactically demanding for many "ordinary" online users active in political debating.[121] For some users, especially those with the privilege of political party protection or the support of news organizations, these attacks have deepened their resolve to continue debating political issues online. However, the moral injuries of gender-based abuse have impacted everyday interactions among an increasing number of young women entering political debates on social media—forcing some to go mild, "neutral" in their opinion, or completely silent. Such practices amount to censoring voices by mobilizing "the [online] qualities of viral outrage to impose a disproportionate cost on the very act of speaking out".[122]

In a comparative study covering Kenya, Senegal, Ethiopia, Uganda and South Africa, researchers found that 39% of women who responded to the questions said they were very concerned about their safety online.[123] 28% said they have become more concerned about digital safety over the

120 Baxi, U. (2009). Humiliation and Justice. In G. Guru (Ed.), *Humiliation: Claims and Contexts* (pp. 58–78). New Delhi: Oxford University Press, p. 72.

121 Udupa, 2017.

122 Tufekci, Z. (2018). *Twitter and Tear Gas: The Power and Fragility of Networked Protest.* Yale: Yale University Press.

123 Iyer, N., Nyamwire, B. & Nabulega, S. (2020). *Alternate Realities, Alternate internets. African Feminist Research for a Feminist internet.* Pollicy. https://ogbv.pollicy.org/report.pdf accessed 2 March 2021.

past five years because of having experienced or witnessed accounts of online violence and attacks. 36% reported sexual harassment online, 33% reported offensive name calling and 27% said they were stalked (repeated contact and doxxing). Participants in Kenya highlighted that the internet had "promoted novel methods of control through surveillance and tracking."[124] Several respondents complained about coordinated trolling that threatened to physically hurt them, some going so far as to call for the murder of the targeted women.[125] Women have largely responded to this phenomenon by blocking or deleting perpetrators (66%), ignoring them (14%) and deleting their own social media accounts (14%). Very few have used the option of reporting to the website or the social media platform (12%). These variegated responses suggest that risks of digital communication are still not addressed adequately. On the other hand, the effects of gender-based harassment have been very severe in some cases. The same study reported that online harassment and attacks on high school girls are on the rise. In countries such as Ethiopia and South Africa, traumatic experiences of online harassment have driven young women to end their lives.[126]

In a different study, researchers identified growing trends of online violence against women in politics during the COVID-19 pandemic in Kenya.[127] Trolls have especially targeted outspoken women politicians fighting for gender equality and social justice, as power contestations and political rivalries have exposed them to a deluge of offensive and dangerous messaging online. Trolls who attacked a popular woman politician, for instance, created fake and parody accounts, demanding that she should share her nude pictures to prove that she was not a man. Other violent attacks on women politicians have involved misogynic attention to their physical bodies, insults and name calling.

A common feature of gender-based attacks on women in politics or journalism is that they are deeply entwined with conflicts over power and political wrangling. In South Africa, social media manipulations of a wealthy business family discussed in the previous section also included targeted attacks against outspoken women journalists. Sponsored malign campaigns sent out sexist, photoshopped and derogatory images of women journalists, calling them "presstitutes" (press as prostitute)—the exact term hurled against women journalists critical

124 Iyer, Nyamwire & Nabulega 2020, p. 18.

125 Iyer, Nyamwire & Nabulega 2020, p 26.

126 Iyer, Nyamwire & Nabulega 2020, p 43.

127 Kenya ICT Action Network. (2020). *Trends of Online Violence against Women in Politics During the COVID19 pandemic in Kenya.* https://africaninternetrights.org/sites/default/files/Trends-of-Online-Violence-against-Women-in-Politics-During-the-COVID19-pandemic-in-Kenya.pdf accessed 2 March 2021.

of religious majoritarian politics in India.[128] In Afghanistan, women journalists voicing their opinion against conservative politics have faced similar accusations of prostitution as well as death threats online.[129] In Uganda, women journalists critical of the government have not only received online death threats but have also suffered actual incidents of kidnapping.[130] Same patterns of harassment are seen in Lebanon where women journalists supporting anti-government protests have seen a flood of fake photos that present them as sexual objects in "compromising positions".[131] Their phones have been hacked, photos have been leaked without consent, and photoshopped images have been doctored. In these cases, religious authority has authenticated conservative politics, urging followers to banish such journalists for disrespecting the Islamic law.

In Saudi Arabia, women commentators and journalists critical of the regime have faced harsh consequences online and beyond. In a case that drew media attention, prominent Al Jazeera anchor and journalist Ghada Oueiss was the target of pro-Saudi social media accounts in a smear campaign after she reported new findings concerning the murder of Saudi journalist Jamal Kashoggi in April 2020. In an interview to the International Press Institute, she recounted that her phone was hacked and her photos in a Jacuzzi were leaked, and the photoshopped images and caricatures showed her in "compromising positions" to suggest that she "obtained her position through sexual favors". [132]

A key aspect of gender-based abuse—online and otherwise—is that its severity is linked to other structures of disprivilege, especially when women belong to minoritized or historically disadvantaged groups. Women's rights activists in the global South have noted that, "Women who belong to, or are identified as belonging to, religious, racial, or ethnic minority groups, Dalit and Bahujan women, the LGBTQI+ community, and women with disabilities face disproportionate abuse, misogyny, and violence online."[133] This observation echoes the views of the UN Special Rapporteur on Minority Issues: "In the case of online violence targeting

128 UNESCO. (2018). *Journalism `Fake News´ & Disinformation. Handbook for Journalism and Training.* https://unesdoc.unesco.org/ark:/48223/pf0000265552 accessed 3 March 2021.

129 https://rsf.org/en/news/afghan-cleric-who-defied-covid-19-lockdown-threatens-woman-journalist accessed 2 March 2021.

130 https://rsf.org/en/news/tv-reporter-kidnapped-and-beaten-over-post-about-first-lady accessed 2 March 2021

131 https://english.alaraby.co.uk/english/comment/2019/11/8/nobody-knows-lebanons-problems-better-than-its-women accessed 2 March 2021.

132 https://ipi.media/middle-eastern-journalists-targeted-by-misogynistic-smear-campaigns/ https://cpj.org/2021/02/ghada-oueiss-hacking-harassment-jamal-khashoggi/ accessed 4 March 2021.

133 Salim, M. (2021). How women from marginalized communities navigate online gendered hate and violence. IT for Change, February.

minority women such as Muslim and Dalit women, and in countries such as India, Pakistan, and Nepal, it's important for governments to be aware and acknowledge that it is not only a gendered issue. These women are doubly targeted and disadvantaged—as women and as members of minorities groups who still face abuse, prejudice, and even persecution because of their religion or caste."[134] Several accounts of gender-based harassment in South Asia and Africa attest to this observation.

Set both within the technocultures of misogyny and specific constellation of power and privilege within local political contexts, gender-based abuse has thus emerged as a weaponized repertoire to squash dissent, contestation, or the sheer aspiration for political participation.

A key action frame for UN entities in this area would be "connection". This suggestion comes from a recognition that there are already a large number of creative grassroots initiatives against online gender-based abuse that could immensely benefit from the financial, technological and tactical support of the UN to increase their effectiveness and scalability. UN entities should connect various initiatives that are spread out around the globe, especially by

i) creating interfaces to link anti-harassment campaigns crafted in different parts of the world, and
ii) between groups that have evolved tactics to respond to harassment and those that are proactively enunciating feminist politics with the creative use of online channels for video based and multimodal narratives.

For instance, the "Girls at Dhabas" project in Pakistan invites women in South Asia to "reclaim public spaces on their own terms" and publishes the stories of women taking a stroll on the streets in the night, or hanging out at "male-only" venues, and celebrating these moments of "transgression" and "occupation" through visual and textual narratives online.[135] Drawing on the benefits of affordable video production and circulation online, the #DigitalHifazat campaign in India similarly launched a series of videos to document the narratives of how women use the internet.[136] These videos provided a platform for women with disabilities, Dalit women, queer women, and queer Dalit women to share their experiences. Such multimodal, personal narratives have a much stronger potential for change and sensitization than the conventional text-based formats of awareness raising.

134 Mariya Salim's interview with Fernand de Verennes, the United Nations Special Rapporteur on Minority Issues, cited in Salim (2021, p. 4).

135 https://girlsatdhabas.wordpress.com accessed 24 Feb 2021.

136 https://feminisminindia.com/2016/11/16/digitalhifazat-campaign-cyber-violence-women-india/ accessed 3 March 2021.

The Digital Rights Foundation in Pakistan offers another example for actions such as digital safety trainings for women, a cyber harassment helpline (Ab Aur Nahin, not anymore) that provides free legal counsel, and campaigns such as "Hamara internet, our internet" to map and raise awareness about online harassment using interactive online formats like quizzes.[137]

Connecting anti-harassment initiatives in different parts of the world can facilitate interactions that will not only provide ways to share resources but also anchor the lessons gained from different activities in culturally appropriate frames, since gender rights are not the same in different social worlds. Some projects however have a potential to travel across contexts. For example, the Peng! Collective's "Zero Trollerance" campaign in Germany[138] in 2015 located 5000 users who were tweeting abusive content to "harass and incite violence" against women and transgender people. These users were identified through a "simple language analysis of Twitter data". Once identified, they were involuntarily enrolled in a six-step self-help program. Each day, the "trolls" received a tweet from a "troll coach bot" with a video link to the "day's step" and motivational content to stay away from such behavior. The six-day program came with six video tutorials. The organization described this as way to troll the trolls: "We trolled the trolls. Except we only used kindness, not hate." Without doubt, the actual content of such videos should stay close to the cultural specificities of different regions, and it is imperative to obtain users' consent, but the strategy to directly address online trolls with well-tailored video tutorials holds the potential for similar experiments in other parts of the world.

Similarly, in a major campaign called #FBrape in 2013, the Women, Action and the Media (WAM!) group in the USA and the Everyday Sexism Project[139] targeted Facebook advertisers, highlighting the consequences of promoting abusive content. A deluge of tweets criticizing Facebook's policies forced 15 major companies to withdraw their advertisements and suspend their marketing campaigns on Facebook. Consequently, Facebook agreed to revise its policies around moderating gender-based abusive content on its platform. In a related project, WAM! and Twitter collaborated for a joint pilot project to create a reporting platform that helped to take action against reported content within 24 hours, as well as fine-tuning the company's content moderation standards based on a joint investigation of abusive content.[140] Such initiatives should be extended to different linguistic communities, beyond the English-speaking world.

137 https://digitalrightsfoundation.pk/ accessed 24 Feb 2021.

138 https://pen.gg/campaign/zerotrollerance/ accessed 24 Feb 2021.

139 https://everydaysexism.com/ accessed 24 Feb 2021.

140 https://www.genderit.org/sites/default/files/wam-twitter-abuse-report_1.pdf accessed 15 March 2021.

Connecting ongoing debates about legislative reforms is yet another area for intervention. Digital rights advocates have called for legislative reforms by developing clear thresholds to bring targeted gender-based harassment within the definition of criminal speech when it has the potential to instigate harm. These have ranged from proposals to enforce stricter self-regulation regimes for internet platforms to holding them accountable for disseminating explicitly violent content such as death and rape threats by invoking the principle of "absolute liability". Legal practitioners examining gender-based abuse in India, for instance, have argued that large internet platforms that "enable toxic masculinity, permit the issuance and wide dissemination of death and rape threats, and thus have a chilling effect on the participation of women in society on account of fear of abuse both online and offline," should be treated as enterprises engaged in "hazardous or inherently dangerous activity" and are therefore "absolutely liable to compensate those who are affected by their operations [even if they are unintended or accidental]".[141] Such calls for stricter actions against perpetrators of gender-based abuse and internet platforms hosting such content have increasingly expressed their skepticism about counterspeech as an effective strategy. While all caution should be taken to avoid regulatory overreach in terms of criminalizing cyberbullying, in the context where gender-based attacks can threaten the safety of women and the LGBTQI+ communities, advocacy for stricter regulatory frameworks such as bringing harmful gender-based attacks within the purview of criminal speech and timely content takedowns is an important step.

Aside from legislative and regulatory reforms, UN entities should connect digital rights groups that are involved in capacity building for key actors in the judiciary about the unique challenges and effects of online gender-based harassment. Along these lines, digital rights groups in India have carried out training sessions for judges to sensitize them about issues of online privacy and safety, and the entrenched patriarchy of the law enforcement systems. Such measures are important because, "institutions of law and justice carry deep prejudices that not only delegitimize the rights of women belonging to minority social groups, but also penalize them for their very aspiration and agency to seek justice."[142]

In contexts where gender-based abuse is entangled with partisan or socially conservative majoritarian politics targeting minorities and political opponents, the problem has to be tackled as part of a broader set of tactics aimed at engaging repressive regimes and gray zones (described earlier). In repressive contexts, one of the key activities at the community level is

141 Raghavan, A. (2021). Legislating an absolute liability standard for intermediaries for gendered cyber abuse. IT for Change, February 2021. https://itforchange.net/sites/default/files/1883/Arti-Raghvan-Rethinking-Legal-Institutional- Approaches-to-Sexist-Hate-Speech-ITfC-IT-for-Change_0.pdf

142 Salim, M. (2021), p. 3.

to connect gender rights advocacy groups that have developed digital safety toolkits across locations, so that victims are able to directly access these resources regardless of whether these advocacy groups have their presence in the countries where victims reside. For instance, a victim of gender-based harassment in Lebanon should be able to connect with anti-online harassment groups in Pakistan or Germany. Such connections are possible if there is a curated space of shared resources that UN entities can host and support. Connecting the victims of gender-based abuse is also important in evolving coping strategies, resource sharing, and technological competence (blocking, reporting, password protection etc.).

In summary

- Connecting creative grassroots initiatives against online gender-based abuse across diverse locations to increase their effectiveness and scalability
- Connection can be achieved by creating interfaces
  i) to link anti-harassment campaigns crafted in different parts of the world in areas including digital safety trainings for women, free legal counsel, cyber harassment helpline, capacity building for reporting abusive content to social media platforms, partnerships with social media companies for quick response/redressal and proposals for legislative reforms
  ii) between groups that have evolved tactics to respond to harassment and those that proactively enunciate feminist politics with the creative use of online channels for video tutorials and multimodal first-person narratives
- Targeting social media advertisers to demote abusive content
- In contexts where gender-based abuse is entangled with partisan or repressive politics targeting minorities and political opponents, the problem has to be tackled as part of a broader set of tactics aimed at engaging repressive regimes and gray zones
- In repressive contexts, one of the key activities at the community level is to connect gender rights advocacy groups that have developed digital safety toolkits across locations, so that victims are able to directly access these resources regardless of whether these advocacy groups have their presence in the countries where victims reside.

BILATERAL AND GEOPOLITICAL INTERVENTIONS

Key action frames: Intermediation and awareness

Extreme speech is also a weaponized tool in bilateral and geopolitical conflicts to create and reinforce sentiments of mistrust, exclusion, fear, and anger toward perceived external enemies, and simultaneously to unite allies.[143] Their instrumental use and impact—under the labels of propaganda and psychological warfare—have been widely documented and researched, as this phenomenon predates digital communication. In the latest digital manifestations, propagandists acting on behalf of nation states have sought to combine hateful content with organized disinformation attacks. Security and defense studies have framed the emerging trends of digital information disorder as "information warfare," arguing that imagination has become the primary target of manipulation in the information era.[144] The impact of manipulative actions is based on stimulating emotions such as enthusiasm or fear. In the context of modern hybrid warfare, disinformation and manipulation blur the terms of war and make it imprecise in the field of international law. Studies have identified a malign chain of cause and effect between disinformation campaigns of ISIS, Russia, and the Trump establishment, who all used strategies of weaponizing the grievances of those who felt left out by economic globalization. These trends are complex because of the involvement of nonstate actors who use information technologies to support asymmetric tactics that spark conflict. Tactics adopted by state apparatuses have resulted in attempts to exploit fragilities and polarizations within specific national polities, targeting ordinary users as active participants in the spread of hate and disinformation.

During the MH17 plane crash in Ukraine, citizen users acted as curators of pro-Kremlin disinformation by producing, selecting and spreading the most popular content about the event on Twitter.[145] It is not only the state-supported media monopoly that produces and disseminates propaganda in the context of Russia-Ukraine, but citizens themselves who

143 Udupa, S., Gagliardone, I., Deem, A., & Csuka, L. (2020). Field of disinformation, democratic processes and conflict prevention. Social Science Research Council, February. https://www.ssrc.org/publications/view/the-field-of-disinformation-democratic-processes-and-conflict-prevention-a-scan-of-the-literature/

144 Arazna, M. (2015). Conflicts in the 21st century based on multidimensional warfare: "Hybrid warfare", disinformation and manipulation. Security and Defence Quarterly, 8(3), 103–129; See also Lewandowsky, S., Stritzke, W. G. K., Freund, A. M., Oberauer, K., & Krueger, J. (2013). Misinformation, Disinformation, and Violent Conflict: From Iraq and the 'War on Terror' to Future Threats to Peace. American Psychologist, 68(7), 487–501; Richey, M. (2017). Contemporary Russian revisionism: Understanding the Kremlin's hybrid warfare and strategic and tactical deployment of disinformation. Asia Europe Journal, 16(1), 101–113.

145 Golovchenko, Y., Hartmann, M., & Adler-Nissen, R. (2018). State, media and civil society in the information warfare over Ukraine: Citizen curators of digital disinformation. *International Affairs,* 94(5), 975–994.

further their own disenfranchisement by using social media to generate, consume or distribute disinformation.[146] Studies have argued that these developments have undermined the autonomy and agency of civil society in the region.

Similarly, in the Arab world, online extreme speech should be understood as a factor in the ongoing configurations of rivalries and alliances, which go beyond the national frameworks. Scholars have shown how, for instance, media-politics relationship in Lebanon and Saudi Arabia "cannot be understood without reference to the relations between them and to the wider transnational context of media and politics in the Arab world".[147] Online extreme speech flows between Pakistan and India, the two South Asian rivals, reveal orchestrated messaging that is picked up and augmented by ordinary users, amounting to reifying the external enemy as well as building internal solidarities, often with interests to consolidate power domestically.

Iran's internet censorship of domestic voices (mentioned under the 'repressive regimes' section) is bolstered by a parallel effort to disseminate a pro-Iran propaganda internationally. News reports have uncovered a network of propagandistic news websites that operate internationally and in multiple languages, to amplify the Supreme Leader's speeches and push for narratives that justify its position vis-à-vis its rivals, namely, Israel, Saudi Arabia and the United States.[148]

In relation to this specific variant of online extreme speech, UN entities should evolve strategies in conjunction with diplomatic tools for intermediation and de-escalation, foremost by engaging key actors in Member States, introducing independent mediation and expertise, and combining these interventions with community level awareness raising activities among ordinary online users.

COMMUNITY LEVEL INTERVENTIONS AND DEEP EXTREME SPEECH

Key action frames: Connection, monitoring, awareness

If one part of online extreme speech circulation relates to technology specific features of virality, algorithmic mediation and so on, a significant part of it operates by tapping social trust and cultural capital at community levels, often making deep inroads into the "intimate sphere" of

146 Mejias, U., & Vokuev, N. (2017). Disinformation and the Media: The Case of Russia and Ukraine. *Media, Culture & Society,* 29(7), 1027–1042.

147 Kraidy, M. M. (2012). The rise of transnational media systems Implications of pan-Arab media for comparative research. In Hallin, D. & Mancini, P. (Eds.), *Comparing Media Systems Beyond the Western World* (pp. 177–200). New York: Cambridge University Press; Hallin and Mancini, 2012, p. 300.

148 https://www.reuters.com/article/us-cyber-iran-specialreport-idUSKCN1NZ1FT accessed 4 March 2021.

families, kin networks, neighbors, caste-based groups, ethnic groups and other socially rooted formations. As a result, hateful language has developed a panoply of banal expressions laced with humor, sarcasm, and popular cultural idioms. At the same time, remixing, mashups, and narratives presented as "facts" have led to rumors on social networking sites and messaging platforms, provoking disturbing incidents of physical violence.

Patterns of distributing political extreme speech in several countries in the global South require a special mention.[149] These patterns highlight the limits of Eurocentric regulatory models, and how the DSA (2020), for instance, defines "public information". With an objective to extend the scope of messages that fall within the regulatory ambit, the DSA has defined the category of "Dissemination to the public" to include all information that is made available to a "potentially unlimited number of persons" and that the "mere possibility to create groups of users of a given service should not, in itself, be understood to mean that the information disseminated in that manner is not disseminated to the public".[150] To balance this against data privacy, it excludes the "dissemination of information within closed groups consisting of a finite number of pre-determined persons", and instead categorizes such type of exchange in the realm of "interpersonal communication". Although these measures appear to be reasonable in terms of protecting digital privacy, such understandings belie vastly "creative" practices of distributing extreme speech in the global South.

In India, WhatsApp groups have been remodeled for political "broadcasts'"and "organic bottom-up messaging" by installing "party men" within WhatsApp groups of family members, friends, colleagues, neighbors and other trusted communities. Typically, a party moderator would find his way into these WhatsApp groups through local connections or by leveraging "community work" such as local brokerage to help people access state benefits and so on, and once admitted, he would relay party messages in unobtrusive ways, often embellished with jokes, "good morning" greetings, religious hymns and iconographies, microlocal development work such as water and electricity supply, and other forms of socially vetted and existentially relevant content. "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 also been recruited to develop and expand such "organic" social media networks for party propaganda. Online extreme speech circulates through networks that build on the charisma of local celebrities, social trust, and everyday habits of exchange. At the same time, the gray zones of digital manipulation mentioned in the previous section

149 More research is needed to examine if similar patterns exist in the global North.

150 https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN accessed 4 March 2021.

operate at the local levels through an elaborate network of digital influence service providers who are ready with a panoply of services—clickbait, GIFs, Facebook likes, Twitter followers and so on—that come on a platter with a price tag. It is common for politicians in India, for instance, to have their own "social media consultant" whose job is to accompany their boss during all the public visits, instantly relay their public engagements by posting images and videos online, organize live chats and "watch parties" to ramp up audience, and promote the posts to gain traction through coordinated "likes", which are sometimes purchased from "third parties". The real challenge is then to redefine such intricate networks for positive and just narratives. UN missions should strive toward building (relatively) autonomous community spaces for countering this kind of "deep extreme speech"—the social variant of technologized "deep fakes"—so that social trust as the key engine of online extreme speech is repurposed for positive narratives.

Some examples have included grassroots community level engagements to sensitize school children and teachers and preparing them to detect hate and disinformation. UN missions should put pressure on social media companies to fund such activities beyond the purview of corporate social responsibility and media awareness programs they have initiated. It is important that a significant part of community level interventions remain independent from direct corporate or political party influence.

UN entities also have the opportunity to mobilize their grassroots connections and support communities that are spearheading efforts to ground digital discourse in democratic values. In this direction, encouraging local partners to form community WhatsApp groups to counter hateful speech and recruiting local cultural influencers would be critical. These influencers should include comedians, poets, musicians, cinema celebrities, online meme creators, digital influence service providers, and online game developers. Involving these different actors will help to ensure that positive narratives are culturally resonant in local contexts as well as digitally contemporary so that such narratives adopt the formats and logics of how discourses actually circulate in online networks. During the 2013 elections in Kenya, anti-hate advocates teamed up with a popular comedy series on television to embed awareness raising information about online propaganda and hateful speech in the comedy narrative. The result was encouraging. An impact assessment study showed that viewers who watched these episodes had a keener grasp of how leaders manipulate their followers with polarizing language and better understanding of the legal consequences of engaging in such speech. [151]

Counter speech campaigns should similarly work with digitally native genres of memes, GIFs,

151 Kogen, L. (2013). Testing a media intervention in Kenya: Vioja Mahakamani, dangerous Speech and the Benesch guidelines. Center for Global Communication Studies, p.3. accessed 22 February 2021.

and humor, and where appropriate, hip-hop, rap, photography and local popular cultures (cinema and music). Several ongoing experiments have demonstrated the value of such creative interventions. Some of the examples cited under the section on gender-based abuses, including multimodal first-person narratives and video tutorials, could be utilized to address other forms of content manipulations. In Myanmar, in the midst of violent attacks against the Rohingya Muslims both online and offline, Facebook stickers featuring a flower in an animated character's mouth were created as a symbol of peace with the message, "End hate speech with flower speech".[152]

Not only the content but the strategies of circulation should also be digitally aware, and recognize that the volume of likes and the relative influence of followers (i.e., how central they are to the networks) are significant determinants of the campaign's virality and visibility.[153] An interesting experiment is the #jagärhär (in Sweden) and #ichbinhier (in Germany) counterspeech movements that mobilize the reassuring phrase, "I am here" to counter hate with polite pushbacks, and back it with coordinated "likes" from others in the movement. Once a positive message is sent out to a hate speaker, others in the movement receive notifications to like this content, thereby amplifying the circulatory force of counterspeech.

Similarly, in an effort to "fight fire with water", iheartmob has created a unique project to mobilize bystander support in addressing online hate.[154] It provides a way for victims of online hate to reach out to them and seek immediate online support to resist and cope with the trauma of hate speech. Aside from receiving vital information on digital safety and ways to discern the severity of threats, victims have the option of making their report public and get community support—the "heartmob"—organized by the group. "Bystanders looking to provide support will receive public requests, along with chosen actions of support."[155] Such interactive community building can provide strong mechanisms of protection and reassurance for online hate victims at the local level. Enlisting the support of local NGOs, journalists, students, and community workers would be important in building such bystander support.

An interesting effort that is worth exploring in other contexts is the "Hass hilft [literally, hate helps]" project implemented in Germany.[156] This organization vows to donate one Euro for every "inhuman" comment they encounter online to anti-hate and refugee support groups.

152 https://www.facebook.com/supportflowerspeech?fref=ts

153 A simple evaluation of indegree, outdegree and Eigenvector centrality measures provides an indication of the influence of online users in the public web.

154 HTTPS://IHEARTMOB.ORG/PAGES/FAQS#HEARTMOB accessed 22 February 2021.

155 HTTPS://IHEARTMOB.ORG/PAGES/FAQS#HEARTMOB accessed 24 February 2021.

156 www.hasshilft.de accessed 24 February 2021.

The premise is that online users who post inhuman comments end up donating for initiatives that have taken up the very cause of eliminating such comments. As the organizers of the movement describe, "…all haters and agitators practically donate against themselves", injecting a "dilemma" into the communicational space of hate mongering.

Another important area for concerted action is to strengthen the community of "trusted flaggers"[157] and equip grassroots organizations with necessary technological resources to report offensive and harmful content to social media companies in a timely manner and monitor the progress of their complaints. One of the key activist engagements of Reconquista internet group[158] in Germany has been to report racist messages and content that glorifies violence to Facebook and evaluate and archive the outcome of their reports by monitoring if the company took down the content in a timely manner.

Equally, efforts to strengthen local communities to petition lawmakers to support victims of online harassment and raise resources for legal help would be critical. Anti-Defamation League's pioneering work in this area is worth replicating in different parts of the world, especially at the national and community levels. This group has also pioneered innovative use of AI-assisted models for content flagging and user education.

Importantly, community level efforts should work with local concepts, local cultural idioms and regional linguistic repertoire rather than the "suspiciously foreign" semiotic baggage of "peace operations". Utilizing the local cultural repertoires for peace, unity, harmony, and cognate concepts is critical for "carving out spaces" that can strengthen and renew the legitimacy of the UN.[159] Cultural language and narratives adopted by grassroots organizations in their diverse engagements around online hate and harassment provide a significant pool of resources in this effort. For example, the "Feminism in India" project used hifazat (protection in Urdu) for its campaigns against gender-based harassment in India; the anti-hate movement in Myanmar was led by a coalition of civil society activists called Panzagar that translates to "flower speech" in English; and the community led movement against electoral violence in Kenya was named Sisi ni Amani ("we are peace" in Swahili).

Such contextualized interventions should also recognize the effectiveness of different kinds of media in a "polymedia environment" where media forms exist in close relation to one another,

157 See the section on "Platform governance" in this paper.

158 https://t.me/ReconquistaNetz accessed 24 February 2021.

159 https://theglobalobservatory.org/2020/10/reflections-on-the-future-of-peacekeeping-operations/ accessed 24 February 2021. Miyashita stresses that the success of UN peacekeeping operations will depend on the legitimacy the UN itself is seen to have.

and people navigate a variety of media forms in their everyday lives. This means tackling online hate cannot be done entirely within the domain of internet-based media but requires a broader and integrated approach based on a thorough mapping of the popularity of different kinds of media. To evolve integrated polymedia responses at the community level, UN entities should leverage ongoing grassroots media initiatives active in sensitizing people about hateful message such as La Benevolencija's radio programs in Rwanda, Burundi and DRC.[160]

Initiatives such as the AI4Dignity project (described under 'global interventions') that aim to both leverage and bring contextual knowledge to AI-assisted content moderation should be replicated at the local and country levels to achieve further granularity and linguistic diversity in the datasets. Therefore, "Counterathon" events should be organized to create collaborative coding spaces for AI/NLP developers, critical intermediaries like fact checkers and academic interlocutors at the local and national levels, based on lessons gained from international events. The connective network can be fine-tuned further at the community levels by creating mechanisms where fact checkers and other anti-hate interlocutors who are already in correspondence with specific hate groups are able to connect with them for precise interventions. This "matching" can also occur based on the shared ethnic/religious/caste/linguistic background of anti-hate teams and hate speakers. An interesting study has shown that online haters are more likely to accept fact checks published by people who have already been interacting with them on social networks.[161] This fine-tuning is important because hateful expressions and disinformation proliferate when they flow within the networks of trust and familiarity, and any efforts at combating them should also operate through such trust-based, interactive networks.

The interactive element in the connective network model can be enhanced further by directly engaging with online hate mongers. A key point of departure here is to design interactions that have the potential to change user behavior. Defined as "participatory enforcement"[162], these efforts insert dynamic interactional frames into hateful exchange online. A good example is the #WeCounterHate project that uses AI-identified content to prioritize messages posted on Twitter for human review. Once human moderators confirm the content as problematic, the system generates a response to "add friction to the user experience" as well as to engage in user education.[163] Typically, the response would be: "This hate tweet is now being countered. Think twice before retweeting. For every retweet, a donation will be committed to a non-profit

160 http://www.labenevolencija.org/ accessed 24 February 2021.

161 Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact checking on Twitter: When do corrections have an effect? *Political Communication,* 35, 196–219.

162 Land & Hamilton, 2020, p. 12.

163 Land & Hamilton, 2020, p. 13.

fighting for equality, inclusion and diversity." The expected outcome is to at least unsettle the atmosphere of normality around hateful exchange and encourage users to reflect on their practice. Although social media companies such as Facebook and YouTube have introduced some of these features in their user interface, engaging civil society organizations and critical community intermediaries will provide further cultural nuance and authenticity to this exercise.

In summary

- Partnering with local cultural influencers for organic influence in social media networks (such as WhatsApp groups) to promote positive narratives
- Developing counterspeech and positive campaigns by using memes, GIFs, and humorous posts so that they are culturally resonant and digitally contemporary
- Strengthening counterspeech and positive campaigns by utilizing the unique features of digital circulation such as coordinated "likes" to promote such posts
- Mobilizing positive narratives and awareness raising by extending the network of partners to include not only conventional beneficiaries such as NGOs but also online comedians, poets, musicians, cinema celebrities, online meme creators, and online game developers
- Strengthening grassroots communities to report online extreme speech to social media companies and monitor progress once complaints are raised
- Strengthening local communities to petition lawmakers to support victims of online harassment and raise resources for legal help
- Offering technical support to local groups to develop hate monitoring dashboards
- Empowering local groups to mobilize community "bystander support" when victims of online hate choose to make their complaints public
- Partnering with existing anti-hate media programs (radio, television, and print) to evolve integrated polymedia responses against online hate
- Developing innovative means of sensitizing hate speakers by channelizing donations to antihate groups for every instance of offensive and exclusionary extreme speech spotted online (in other words, hate speakers would be funding anti-hate initiatives each time they post a hateful message)
- Organizing initiatives such as the AI4Dignity project that aim to both leverage and bring contextual knowledge to AI-assisted content moderation at the local and country levels to achieve further granularity, linguistic diversity, and interactivity.

Across all these levels of intervention and action frames, it is important to recognize that extreme speech is not merely a problem of digital communication but of deeper histories of racialization, of coloniality of power now manifest as exploitative and racialized data relations, and of repressive states that have turned against their own citizens. A context sensitive approach calls for a multiprong, rapidly evolving, and robustly flexible mechanism that can simultaneously counter existing patterns of digital harms and anticipate trends that are simmering and emergent.

Udupa, Sahana. 2021. Digital Technology and Extreme Speech: Approaches to Counter Online Hate. Research Paper for the United Nations Peacekeeping Technology Strategy.
New York: United Nations Peace Keeping.

AUTHOR INFO

Sahana Udupa is professor of media anthropology at LMU Munich, Germany, where she leads two European Research Council funded projects on digital politics.
https://orcid.org/0000-0003-3647-9570

ACKNOWLEDGEMENTS

The author thanks Laura Csuka and Miriam Homer for their excellent research assistance. She also thanks the reviewers and Naomi Miyashita at the UN Department of Peace Operations for initiating and supporting this paper through different stages of its development. This paper has built upon the insights gained from a longer study on digital politics funded by the European Research Council under the Horizon 2020 program (grant agreement number 714285 and 957442).