CENTER FOR EQUITY, GENDER & LEADERSHIP

# EQUITABLE LANGUAGE CERTIFICATE: HATEFUL SPEECH BRIEFING

*By the Center for Equity, Gender & Leadership (EGAL) at UC Berkeley Haas School of Business*

*April 2023*

This briefing explores what hateful speech is, where it comes from, and what its impacts are on individuals and society. It includes a deeper dive into the proliferation of hateful speech online, particularly on social media platforms. Finally, it outlines research and strategies aimed at tackling hate speech, both on- and offline.

This briefing was written by Ishita Rustagi, Genevieve Smith, and Ornella Tchoumie.

# Table of Contents

# A. WHAT IS HATEFUL SPEECH?

Hateful speech – or “Hate speech” – is an umbrella term that encompasses “toxic” or “dangerous” speech. There is no universal definition of this term because hate speech is fluid and context dependent. That is, a single definition may not reflect the lived experience of all those encountering hateful language (see Box 1). As a result, researchers and organizations / policy makers contending with hate speech utilize definitions most suited for the context(s) in which they operate, rather than seeking out a universally applicable definition. However, hateful speech generally refers to **forms of expression that incite violence, instill fear, insult a person or a group’s dignity, and/or threaten the wellbeing of an individual or community.**[1]

### Box 1: Hate speech is fluid and context-dependent

Hate speech is fluid and heavily context dependent. A phrase or symbol may be considered hate speech in one geographical or social context, but not in another. For example, before the swastika was appropriated as a Nazi symbol, coming to symbolize fascism, it originated in India, where it is still used as a symbol of good fortune[2]. Whether something is considered hateful or not also depends on the speaker and the receiver. Many groups and people have reclaimed slurs or insults as terms of empowerment – for example, the word “B*tch,” which has been used as an insult towards women, has been reclaimed by some women who may use it as a term of empowerment or endearment towards other women.

## B. WHAT ARE THE IMPACTS OF HATEFUL SPEECH – BOTH ON AND OFFLINE?

While defining and identifying hate speech may be context dependent and not always clear cut, the harms caused are evident, and can range from emotional and mental to physical harm. Recipients or targets of hate speech may experience harm instantaneously, or may suffer longer-term impacts.

For *individuals, communities, and societies*, hate speech can...

- Cause emotional and mental distress, negatively impacting the mental health and wellbeing of the specific individuals or communities experiencing hate speech;
- Amplify harmful stereotypes and bias;
- Enhance divisiveness and polarization;
- Be used as a tool of oppression by those in power; and
- In more extreme cases, support or encourage acts of violence towards the target groups[4]

Organizations are not exempt from the impacts of hate speech. For the *businesses* or platforms on which hate speech occurs or is allowed to proliferate, this can...

- Negatively impact brand reputation;[5, 6]
- Enhance risk (brand, financial and regulatory); and
- Be at odds with their purpose, vision and principles.

It is important to note that hate speech occurs on a spectrum. Its impacts – particularly online – can range from subtle microaggressions to calls for violence against individuals or groups, and can be expressed through text, audio, visuals, or any combination of these forms of media on various platforms. Misinformation may also be leveraged to generate and propagate harmful language. See Box 2 for fictional case studies of what hate speech can look like on social media platforms. See Box 3 for a series of real world examples demonstrating some of the more extreme impacts that hate speech can have.

### Box 2. Fictional case studies (derived from real examples) showing how harmful language can impact individuals and communities through social media

1. **Harmful language in online groups**
    a. Laila, a Black woman who enjoys gardening, joined a Gardening group on social media. She encountered racist content in the comments on her post. Feeling unwelcome, she left the group, deciding to search for a group that focuses on Black people interested in gardening.
2. **Harmful language in response to personal posts**
    a. Charlie is a member of the LGBTQ+ community. During Pride month, they uploaded a short video to their social media feed with snippets of celebratory activities, and a caption and hashtags about the importance of Pride. However, they soon encountered a barrage of hateful comments on their post and discriminatory memes and gifs being shared in response to their post. Charlie no longer felt like they belonged on this social media platform, and deleted their profile.
3. **Harmful language on marketplace platforms**
    a. Neeraj decided not to buy the plant he had agreed to purchase from an online seller on an app that connected people in his neighborhood. He communicated this to the seller through the chat function, and the seller, enraged, responded with slurs and other harmful language. Neeraj felt unsafe given the proximity of the seller to his living space, and decided against using this platform for future community building or inquiries.

### Box 3. Examples of more extreme harmful impacts of hate speech

****Trigger warning** This box contains descriptions of real harms committed against certain communities. Only continue reading if you are in the right headspace to do so. Otherwise, continue with section C.**

1. In Rwanda in the 1990s, a radio station allied with leaders of the government and with a wide listenership of Hutus, repeatedly used language describing the Tutsi community as “inyenzi” (cockroaches), and “inzoka” (snakes). This hateful language eventually incited the Hutus to commit mass genocide against the Tutsis – Hutu neighbors took arms and moved house to house hunting their Tutsi neighbors.[7]
2. Per an Amnesty International Report, Facebook’s algorithms ‘proactively amplified’ hateful speech targeting Myanmar’s Rohingya ethnic minority. The Rohingya have long been persecuted by Myanmar’s Buddhist majority, but the report found that Myanmar’s armed forces actively used Facebook’s platform to boost anti-Rohingya propaganda, a move which allowed them to garner support for a campaign of rampant violence against the Rohingya minority in August 2017.[8] In December 2021, Rohingya refugees filed lawsuits seeking $150 billion in compensation for Meta’s role in amplifying hate speech.[9]
3. In November 2021, Abrham Mearag’s father, a Tigrayan chemistry professor, was shot and killed outside his home, in the wake of a series of hateful posts on Facebook targeting him for attack. In December 2022, Ethiopian researchers Abrham Meareg and Fisseha Tekle, along with Kenyan human rights group Katiba Institute, filed a lawsuit against Meta for hateful speech on the platform fueling ethnic violence in Ethiopia.[10]

# C. WHERE DOES HATEFUL SPEECH COME FROM?

Hate speech can reiterate and amplify messages of bias and discrimination. In particular, the following factors are of note:

- *Limiting stereotypes and bias:* Hate speech is a manifestation of the limiting stereotypes and discrimination that exist in society. The use of a swastika in contexts where it represents white supremacy and anti-Semitism is an example of this, as the symbol can reflect and reinforce white supremacy and discrimination against Jewish communities.

- *"Us vs. them" mentality:*[11] When a group or community in power (the "in-group") sees themselves as superior, they may attempt to establish this superiority or create othering towards "out-groups" by using hateful speech to amplify differences, negatively judge, and/or express condemnation towards members of the out-groups. The example of Hutus and Tutsis in Rwanda (see Box 3) demonstrates this – hate speech was used to incite othering and violence towards Tutsis, a group that had previously been in power, in an attempt to shift the social dynamics.

- *Power dynamics:* When a group, community, or individual already in power has a fear of losing power, or a desire to reinforce existing power dynamics, they can use hateful speech to negatively impact those with less power. The example from Myanmar (see Box 3) demonstrates this.

# D. WHY DOES HATEFUL SPEECH PROLIFERATE ONLINE?

With frictionless technologies such as social media, which aim to connect people from all over the world and provide users with platforms on which they can express themselves, hate speech can proliferate quickly and easily. There are several factors that make this possible, including the following:

- The lack of in-person / face-to-face communication can create distance and minimize space for empathy.
- People are afforded anonymity by social media platforms, and are able to hide or disguise their identities.[12] This further results in a lack of accountability for users creating or propagating hateful content.
- Social media allows people to easily broadcast their messages to millions of people across the world, making it easier to amplify hateful messages.
- Algorithms on social media platforms are designed to maximize engagement. Research finds that polarizing content usually results in higher levels of engagement, which means that such algorithms can often inadvertently promote hateful speech.[13]
- At the same time, current content monitoring / moderating capabilities (both led by human moderators and/or hate speech detection algorithms) have several limitations:
  - Content moderation approaches tend to be reactive rather than proactive,[14] which means that even when harmful content is removed, it may have already resulted in harm.
  - Algorithms built to flag harmful content may fail to grasp underlying meanings or contexts. For example, they often fail to recognize whether a term is being used as a reclaimed means of empowerment or a slur, based on who is using it. Research finds that these tools risk filtering out voices of marginalized groups.[15]

- In general, Natural Language Processing (NLP) tools tend to underperform for demographic groups whose language varieties are not well represented in datasets. E.g., a study analyzing five widely-used speech recognition tools found that they misunderstood words spoken by Black people nearly twice as often as they misunderstood words spoken by White users. This can be traced back to underrepresentation of African American English in the language datasets these AI systems learn from.[16]
- Additionally, AI models that are built for harmful speech detection are primarily text-based, but in reality hate speech manifests in different types of mediums - text, image, video, sound, etc.

# E. HOW DO WE TACKLE HATEFUL SPEECH? (ON- AND OFFLINE)

Existing methods to tackle hate speech tend to be reactive (responding / flagging hateful speech), versus considering proactive methods to mitigate hate speech in social media products. These reactive approaches include content moderation algorithms (that utilize NLP) and human content moderators flagging and addressing harmful content. Developers and researchers are also implementing design tweaks, such as having users consider their posts before posting.

We outline below some innovative frameworks being researched and applied in the tech industry to tackle hate speech:

- *Quarantining hate speech*:[17] When a user posts hate speech targeting another user, an algorithm – upon detecting the hateful content – 'quarantines' the message until the receiver opts in to see it. The receiver is initially sent a warning alert with a hate speech severity score and the sender's name. If the receiver declines to see the content, it is deleted entirely. The parameters for defining whether or not content should be quarantined are not explicitly outlined in the research, and there are multiple ways to apply this framework. Reddit uses a version of this approach, sometimes choosing to quarantine an entire community such that it isn't included in search or recommendation functions, and is unable to crosspost or share messages.[18]
- Better Design:[19] This approach is more proactive, and advocates for social media algorithms to prioritize community building over engagement. Researchers offer several ways to operationalize this, including:
    - Taking a *hyperlocal* approach: prioritizing posts from community members, friends, and family members – within a 10-mile radius, for instance. The goal here is to offer a more intimate relational sphere on a user's newsfeed, to encourage more empathy and positive engagement.

  - *Centering mindfulness: combatting the frictionless features of social media platforms by offering 'empathetic prompts' for users to reflect on whether or not they want to post hateful content, or 'ideological prompts' reminding users their post may not be seen by folks with differing views.*
- *Restorative justice*:[20] This method focuses more on the needs of the victims of hate speech, with the goal of rehabilitating the offender, giving them the tools to make amends with the victim and community. See here for an example of how researchers applied this (at a small scale) to an online community in Canada.
- *Proactive Product Lifecycle Framework*: It is critical to consider how to promote inclusive language and tackle harmful language in products *proactively*. This can be done by asking key questions at every stage of the product lifecycle. In the Equitable Language Certificate program we share and apply an Inclusive Language Product Lifecycle Framework including sample questions that teams can ask themselves and/or other relevant stakeholders to proactively research and develop products that promote inclusive language.

# F. CONCLUSION AND CALL TO ACTION

It is evident that hateful speech is prevalent both on- and offline, and can have a wide range of impacts on individuals, organizations, and communities more broadly. In order to build social media products that bring people together, it is important for organizations and teams to understand how hateful speech may manifest or be amplified through these platforms and take a stance on mitigating it – reactively, but also importantly, proactively.

# REFERENCES

[1] Anderson, L., & Barnes, M. (2022, January 25). Hate speech. Stanford Encyclopedia of Philosophy. Retrieved from https://plato.stanford.edu/entries/hate-speech/

[2] Campion, M. J. (2014, October 23). How the world loved the swastika - until Hitler stole it. BBC News. Retrieved January 18, 2023, from https://www.bbc.com/news/magazine-29644591

[3] Hate speech. Transparency Center. (n.d.). Retrieved January 18, 2023, from https://transparency.fb.com/policies/community-standards/hate-speech/#policy-details

[4] Council on Foreign Relations. (n.d.). Hate speech on Social Media: Global Comparisons. Council on Foreign Relations. Retrieved January 18, 2023, from https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

[5] Nicas, J. (2021, February 3). How YouTube drives people to the internet's darkest corners. The Wall Street Journal. Retrieved from https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478

[6] Frenkel, S., Isaac, M., & Conger, K. (2018, October 29). On Instagram, 11,696 examples of how hate thrives on social media. The New York Times. Retrieved from https://www.nytimes.com/2018/10/29/technology/hate-on-social-media.html

[7] Ndahiro, K. (2019, October 24). In Rwanda, we know all about dehumanizing language. The Atlantic. Retrieved from https://www.theatlantic.com/ideas/archive/2019/04/rwanda-shows-how-hateful-speech-leads-violence/587041/

[8] Guzman, C. de. (2022, September 29). Report: Facebook algorithms promoted Anti-Rohingya Violence. Time. Retrieved from https://time.com/6217730/myanmar-meta-rohingya-facebook/

[9] Nix, Naomi. “Facebook Facing Mounting Legal Fights over Myanmar Genocide.” Bloomberg.com, Bloomberg, 6 Dec. 2021 https://www.bloomberg.com/news/articles/2021-12-06/facebook-facing-mounting-legal-fights-over-myanmar-genocide.

[10] Solon, O., Prinsloo, L., Genga, B., & Bloomberg (2022, December 14). Facebook parent meta sued for amplifying hate speech and incitement to violence in Ethiopia's Civil War. Fortune. Retrieved from https://fortune.com/2022/12/14/facebook-parent-meta-sued-amplifying-hate-speech-incitement-violence-ethiopia-civil-war/

[11] Davies, Janey B. “Us vs Them Mentality: How This Thinking Trap Divides Society.” *Learning Mind*, 15 Apr. 2022, www.learning-mind.com/us-vs-them-mentality.

[12] Matamoros-Fernández, Ariadna, and Johan Farkas. "Racism, Hate Speech, And Social Media: A Systematic Review And Critique". Television &Amp; New Media, vol 22, no. 2, 2021, pp. 205-224. SAGE Publications, https://doi.org/10.1177/1527476420982230. Accessed 22 May 2022.

[13] Nicas, J. (2021, February 3). How YouTube drives people to the internet's darkest corners. The Wall Street Journal. Retrieved from https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478

[14] Ullmann, Stefanie. “Quarantining Online Hate Speech: Technical and Ethical Perspectives.” SpringerLink, 14 Oct. 2019. Retrieved from link.springer.com/article/10.1007/s10676-019-09516-z?error=cookies_not_supported&code=50ea9df6-6db0-46b5-b388-a9b6b5718aea.

[15] Smith, G., Rustagi, I., Sheares, A., & Nee, J. (2021, October). Responsible Language in Artificial Intelligence & Machine Learning: An Equity Fluent Leadership Playbook. Berkeley Haas Center for Equity, Gender & Leadership.

[16] Smith, G., Nee, J., Rustagi, I., & Sheares, A. (2021, October 27). Advancing Language for Racial Equity and Inclusion: An Equity Fluent Leadership Playbook. Berkeley Haas Center for Equity, Gender & Leadership.

[17] Ullmann, Stefanie. “Quarantining Online Hate Speech: Technical and Ethical Perspectives.” SpringerLink, 14 Oct. 2019, link.springer.com/article/10.1007/s10676-019-09516-z?error=cookies_not_supported&code=50ea9df6-6db0-46b5-b388-a9b6b5718aea.

[18] R/announcements - revamping the quarantine function. reddit. (n.d.). Retrieved January 18, 2023, from https://www.reddit.com/r/announcements/comments/9jf8nh/revamping_the_quarantine_function/

[19] Munn, Luke. “Angry by Design: Toxic Communication and Technical Architectures.” Humanities and Social Sciences Communications, vol. 7, no. 1, 2020. Crossref, https://doi.org/10.1057/s41599-020-00550-7.

[20] Hasinoff, A. A., Gibson, A. D., & Salehi, N. (2020, August 14). The promise of restorative justice in addressing online harm. Brookings. Retrieved May 22, 2022, from https://www.brookings.edu/techstream/the-promise-of-restorative-justice-in-addressing-online-harm/