Policy Proposal: Digital Anti-Hate Immersion

Dr. Saadia Gabriel, UCLA

Hate speech, both online and in K-12 schools, is linked with dehumanization. By its nature, it is the reduction of an entire identity group to a caricature that can be portrayed as inferior, other, even evil. It has long lasting effects beyond direct victims - as found by a recent Nature study [1], "*immersion in a hateful environment leads to empathic numbing: people exposed to hate speech have limited ability to attribute the psychological perspective of others, regardless of their group membership*." The patterns of language weaponization are taught just like grammar or mathematics. By Fall 2020, it was found that 82% of households with K-12 students had reliable digital device access [2]. This exposure creates an online environment where students are constantly absorbing toxic and hateful content. When Black female journalists and politicians are attacked in one out of every ten tweets targeted at them [3], it gives this behavior the noxious semblance of "normality." When immigrants are falsely accused of eating pets [4], that begins a pattern that can escalate into violent behavior. One-off sensitivity training, community dialogue and disciplinary actions like suspensions are important, but not enough to combat environmental factors. Neither is it sufficient to block harmful websites or remove hateful comments online. A state-level policy proposal to defend future generations against a cycle of hateful language is to proactively and consistently present an opposing view to online dehumanization. I suggest that users of school-issued devices be required to interact with a browser extension that provides explanations of why comments they or others post on social media may be harmful. Student perpetrators of hate violence should be required to only use devices equipped with this browser extension while on school premises for a period of time (e.g. 6 months). Such a setup could even be gamified, so students are rewarded for helping to flag harmful content they witness. For practicality of implementation, these explanations could be partially AI-based but grounded in historical and cultural roots of hate speech. This is an initial step towards creating a digital environment that is as immersive in its efforts to oppose hate as it has been in sustaining it.

References

[1] Pluta, Agnieszka, Joanna Mazurek, Jakub Wojciechowski, Tomasz Wolak, Wiktor Soral and Michał Bilewicz. “Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others’ pain.” *Scientific Reports* 13 (2023).

[2] Niu Gao. “Testimony: California’s K-12 Digital Divide Has Narrowed, but Access Gaps Persist.” Public Policy Institute of California Blog (February 21, 2024). https://www.ppic.org/blog/testimony-californias-k-12-digital-divide-has-narrowed-but-access-gaps-persist/

[3] Tyler Musgrave, Alia Cummings and Sarita Schoenebeck. “Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media.” CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, April 2022.

[4] Hannah Knowles and Sarah Ellison. “Trump, GOP fuel conspiracy theories: Eating pets, a rigged debate and QAnon.” The Washington Post (September 15, 2024). https://www.washingtonpost.com/politics/2024/09/15/trump-republicans-baseless-conspiracy-theories/