Improving AI-mediated Hate Speech Detection: A Genuine Ethical Dilemma

Maren Behrensen, Philosophy, University of Twente (Netherlands)

Paper for CEPE 2023, May 16-18, Chicago, Ill.

Abstract:

AI-mediated hate speech detection is indispensable for contemporary communication platforms. But it has known deficiencies in terms of bias and context-awareness. I argue that improving on these known deficiencies leads into a genuine ethical dilemma: It will increase the epistemic and social utility of these platforms, while also helping bad faith political and corporate actors to suppress unwelcome speech more swiftly and efficiently.

Introduction

After Elon Musk acquired Twitter, he quickly fired many employees that worked in key areas of content moderation, and scrapped entire teams providing input for this task, including the Ethical AI team (Knight 2022; Stolton et al. 2022). Within days, the platform was overwhelmed by functionality issues and a deluge of slurs, misinformation, and paid "verified" accounts impersonating companies and public figures, including Musk himself (Mac et al. 2022). This prompted many users to leave the platform, and drastically reduced the platform's appeal to advertisers, endangering its financial viability. While Twitter has not collapsed, and many disgruntled users have since returned, these events illustrate the relevance of content moderation and user management for social media platforms. Elon Musk has claimed to be a free speech absolutist (Glozer at al. 2022), believing that any and all restrictions on speech are more harmful (in the long run) than harms that might result from the restricted speech. (Betraying his own professed principles, he has sought to restrict speech on Twitter that criticized or mocked him; Milmo and Hern 2022). For the governance of social media platforms, free speech absolutism implies anarchism: there is no central authority to legislate speech; users regulate themselves and each other.

Musk (or perhaps: his advisers) quickly found out that this anarchism was not economically sustainable. It threatened to turn away advertisers, and cut into the company's budget. Jiménez-Durán (2022) even suggests that the positive effects of content moderation do not apply to users, but manifest as a slight increase in advertising revenue. Furthermore, there is increasing political and legal pressure toward content moderation. Gorwa et al. (2020) detail how the livestreaming of the terrorist attack in Christchurch on Facebook Live in March 2019 (and subsequent sharing of the footage) led to mounting government demands for the swift blocking and removal of such content. The authors also note:

> "Under recent regulatory measures like the German NetzDG [Network Enforcement Act] or the EU Code of Conduct on Hate Speech, platforms are increasingly being bound to a very short time window for content takedowns that effectively necessitates their use of automated systems to detect illegal or otherwise problematic material proactively and at scale." (Gorwa et al. 2020: 2)

Such considerations about a changing legal landscape that pushes against free speech anarchism will also affect corporate decision-making.

Aside from financial and political considerations, content moderation can also be defended on moral grounds. Unmoderated platforms are rife with all forms of abuse: trolling, bullying, slurs, threats. As Gillespie (2018: 21) points out: content moderation constitutes a social media platforms as platforms, it shapes their users' experiences and determines their quality. Langvardt (2018: 1358) describes the "broad dilemma" of content moderation as follows: "The internet makes it easy for bad actors, ranging from trolls to spammers to malicious hackers, to deter or frustrate speech within online channels." So if these online channels are to remain usable for anyone who is not a troll, spammer, or hacker, some form of content moderation is needed. Grimmelmann (2015: 50-51) identifies three major benefits of content moderation: productivity, openness, and lower costs. Well-moderated online platforms are more productive in the sense that more information of higher quality can be exchanged between users, increasing the platform's epistemic value. It will also be more inviting and easier to use for a broader range of users than a platform beset by trolling and abuse, thus increasing its social value. Lastly, well-moderated platforms will require fewer resources in terms of maintaining its infrastructure, and user will expend less time and energy in sifting through unwanted or offensive content and policing each other's behavior.

This last point adds to legislative pressure on social media corporations to invest in content moderation; in addition to facing pressure from lawmakers, they will recognize it as an investment that yields dividends. However, as Grimmelmann notes (2015: 52), the three benefits are not necessarily commensurable, as "the most natural way to protect infrastructure is to discourage use by limiting access, while the most natural way to promote the sharing of information is to encourage extensive use by opening up access." In other words: the epistemic and social value of content moderation can be at odds with its financial value. Twitter's recent troubles seem to have been caused by the incommensurability: desperately wanting to cut costs, while also aiming to expand the platform's user base and actual usage, Musk demolished much of what made it attractive to users. (As he is struggling to recoup his losses, he is now focusing on limiting access by essentially turning Twitter into a paid service.)

Despite this incommensurability, Musk's blunder illustrates how moderation pays off in more than one way. For users interested in exchanging information, it safeguards the epistemic value of the platform; this safeguarding is a direct investment into the social and economic sustainability of the platforms; and it increasingly turns into a legal requirement in some jurisdictions, especially the European Union with its comparably strict hate speech and privacy laws, platforms are under increasing legal pressure to step up their moderation efforts (Gorwa et al. 2020; Land and Helfer 2022).

In short, the social media industry has a real need for powerful and efficient content moderation. This need has sparked a thriving field for AI research and development: automated hate speech detection. The sheer amount of content, and the frequency with which new content is generated on large social media platforms makes it impossible to tackle this challenge with human labor alone. (Although it must be noted that the exploitation of human laborers, and harms resulting from exposure to representations of extreme violence, still pose serious ethical concerns about commercial content moderation in its current form; Arsht and Etcovitch 2018; Barrett 2020; Steiger et al. 2021). So using AI to solve the problem of sifting through extremely large data sets to find and flag the most toxic content seems like a promising solution; in fact, it might seem like the only solution.

At the same time, we know that AI-mediated hate speech detection has deficiencies, and I argue here that improving it in light of these known deficiencies, specifically bias and lack of context, will exacerbate ethical concerns about nefarious dual use. It is not just possible, but likely that this improved AI will be used to suppress speech that is valuable, but politically unwelcome. I

will not engage with or contribute to the debate about what constitutes hate speech (for an overview, see Sellars 2016). For my argument here, such a definition is not necessary; and as will be shown, the dynamic nature of the concept of 'hate speech' is central to the ethical dilemma I will sketch. In a nutshell, the problem is that the same tools that target incontrovertibly racist or sexist speech can be used to target anti-racist or feminist counterspeech. What they are being used for depends on human knowledge and human intentions: annotators with limited knowledge of the context of potentially hateful content, or owners of social media platforms acting in bad faith can easily turn content moderation into censorship.

Known Problems

As with other applications of AI, such as crime prediction or health care (Ugwudike 2021; Celi et al. 2022), AI-mediated hate speech detection has been documented to perpetuate existing social, cultural, and political biases. Examples include discrimination against speakers of African American English (AAE), when they reclaim words (such as *n***a*) that would be considered slurs coming from non-AAE speakers (Sap et al. 2019) and against queer persons when they reclaim words that would indicate hostility when uttered by speakers who do not consider themselves queer (such as *f*g* or *t****y*; Dias Oliva et al. 2020). Analogous observations can be made about the use of terms that are generally considered sexist (such as *b***h*) as an affectionate form of address within an in-group. A related problem, albeit less relevant here, are false positives: obviously harmless content being flagged as harmful or pornographic (there are many examples for this from image recognition techniques, such as Facebook flagging photos of the Little Mermaid statue in Copenhagen as a violation of community guidelines; Langvardt 2018: 1355).

The semantic difference in language use between in-group and out-group is central to the problem of bias. It makes the difference between an innocent form of address and a slur, and AI-mediated systems are poorly equipped to pick up on these differences between groups of language users. It is likely to miss information that is vital for successful content moderation, and in the worst case, it might target the wrong group of language users. While the industry is aware of this problem, and researchers and developers are continually proposing new technical solutions to address it, it clearly appears to persist (at least for the average user of one of the large social media platforms).

Another well-documented concern is the disproportionate dominance of English in online content moderation efforts (Udupa et al. 2022). This results in a lack of context and knowledge about hate speech and toxic propaganda in online communities where English is not the main language, or not used at all. This lack of context and knowledge affects both human laborers and AI-mediated systems. Like bias, it creates both false positives and false negatives that could be avoided with proper knowledge of the community in question.

One aspect of the lack of content is a simple quantitative issue. As a global *lingua franca*, the most significant investments in the area of AI and language will focus on English. This is a relative position of dominance: other languages with hundreds of millions of speakers will also receive sustained attention, while languages with mere thousands of speakers will not. Some of this attention arguably mirrors colonial structures, that is, the field of AI and language may be more developed for Spanish and French than for Bengali or Swahili. But there is nothing *per se* that would stand in the way of developing AI-mediated hate speech detection for all languages above a certain threshold of speakers; indeed, such efforts are already under way (see, for instance, Das et al. 2022; Ibrahim et al. 2022). But training an AI for this task goes beyond language skills.

In their report on a transnational project on AI-mediated hate speech detection, Udupa et al. (2022) discuss how Kenya's political landscape, in which political allegiances are strongly tied to ethnic belonging, makes it very difficult to separate ethnicity (which would be a "protected category" in terms of hate speech) from political affiliation (which would not be a protected category). The authors report on a research project which sought to develop standards for content moderation in four countries (Brazil, Germany, India, Kenya) via what they call "ethical scaling", that is, the involvement of fact-checkers and annotators who are from these countries and knowledgeable about the local political context. They note significant political and cultural differences between these contexts; and so, while "ethical scaling" seems like a promising response to lack of context in AI training, it would require enormous amounts of human labor to expand it beyond a comparative context including just a handful of countries. The context that is lacking in the training of AI must ultimately come from humans. To overcome the current bias toward the context of North America and a few Western European countries, massive input from other regions of the world is needed, and efforts will need to be made to ensure that this input is not itself affected by human bias (the recommendations in Barrett 2020 lead to the same conclusion).

Human ignorance and human bias are the main causes for the known problems sketched here, as they feed into the sampling, annotation, and interpretation of the data (Cortiz and Zubiaga 2021). What makes these problems even more intractable is the fact that online speech is quick to adapt to filtering efforts that focus on specific keywords. Recent examples include the "Let's Go, Brandon" memes and the "Dark Brandon" responses, the first being a disguised insult of President Biden, the second being a reclaiming of right-wing efforts to paint Biden as a geriatric autocrat. Common tactics are the use of words or names with an alternate spelling (some Twitter users took to calling Elon Musk "Elmo" after he appeared to censor content critical of him) and the use of words with a generally positive meaning in a disparaging manner (Udupa et al., at 2022: 22, note the use of the term *Goldstück* – golden nugget – to refer to refugees in German far-right circles).

Still more challenging for AI-mediated hate speech detection is the use of implication rather than direct attack (Wiegand et al. 2021; Yin and Zubiaga 2022), such as masking homophobic or transphobic messages as "mere questions" or "concern for women and children", the use of "dogwhistles" (Saul 2018) that conceal the appeal of message to an audience's prejudices, and the use of "racial figleaves" (Saul 2017), that is, utterances designed to provide plausible deniability for racist content (the simplest one of which is "I am not racist, but…"). Implication, dogwhistles, and figleaves work in different ways to obscure context; context that would also throw off AI-mediated content moderation efforts. Picking up on implications and dogwhistles requires knowledge of a particular group's language use. This would be difficult to model in AI-mediated systems, and it requires a large number of knowledgeable human laborers who stay up to date on politics and language use. In these regards, the adaptation of toxic speech to filtering efforts could well prove too quick for the development of AI and the recruitment of competent staff (Udupa et al. 2022). Figleaves threaten to overwhelm such efforts in a different manner: they provide additional context that is designed to conceal explicitly racist (or otherwise toxic) content. "I am not racist, but…" is designed to present a racist claim as a 'mere observation', and suggesting that a 'mere observation' cannot be the kind of internal attitude that a racist person would have. Figleaves frustrate content moderation efforts insofar as they suggest that there is never enough context to warrant the restriction of online speech. This fits a common strategy of those who defend their own or others' toxic speech: often their first complaint is that the offending speech was 'taken out of context' (regardless of whether there is any context that would actually add new information).

Proposed Solutions

Some of the challenges posed by bias and lack of content likely will or could at least in principle have technical solutions. As corporations invest in translating their AI solutions for other languages, relevant context will be added by those human laborers who develop and train these solutions. Common and predictable uses of dogwhistles and slight alterations of slurs and other forms of toxic speech will be picked up by AI-mediated hate speech detection (this being the part of the semantic arms race that can easily be won). Training these AI systems on larger corpora, and employing more annotators from more diverse backgrounds should also make it possible to detect implied hate speech with more precision. Furthermore, the use of data sets that are less dependent on keywords could help train AI to pick up on implication rather than explicit speech (Yin and Zubiaga 2022).

These technical solutions will raise immediate ethical concerns. Making larger corpora of text available for AI training will amplify privacy and copyright concerns at the data collection stage (Cortiz and Zubiaga 2021). As mentioned above, the data annotation process will require significant investment in human labor to ensure that annotators are knowledgeable, can work in adequate circumstances, and that their respective biases do not have an overdue influence on the AI system. Given how much unpaid labor in training AI is currently done by users (for instance, in "identify the traffic lights" style 'security' checks), and how much of the work in content moderation is still done by human workers in precarious and harmful work environments (Barrett 2020), this is a tall order. It would require the overhaul of the most powerful social media corporations, and the effective, global implementation of labor protection regulations.

While all this is unlikely to happen (at least in the short term), it would be required for the "ethical scaling" envisioned by Udupa et al. (2022). As my concern here is the improvement of AI-mediated hate speech detection in terms of bias and context-awareness, let us assume that such a shift in corporate culture and legal oversight of AI companies is possible. Let us assume further that the expansion of text corpora needed for the improvement of AI will respect copyright laws and privacy considerations. And finally, let us assume that these changes yield the desired effects: AI-mediated hate speech detection that is efficient and precise, needs little human oversight, but is transparent enough for humans to understand its recommendations and decisions.

How would this ethical progress in the development of this technology lead into a dilemma?

The Dilemma

In a nutshell, the dilemma emerges due to the likelihood of dual use. If "ethical scaling" for AI-mediated hate speech detection could be implemented at large scale, then dual use at large scale becomes likely and indeed possible. The main reason to think that it would be likely are current efforts by bad faith actors to frame unwelcome speech as toxic or dangerous speech. We can think here of Elon Musk claiming that the Twitter account detailing the travels of his private airplane was 'doxxing' him (suspending the account, while he let far-right propagandists back on the platform; Spangler 2022). We might also think of resistance networks in autocratic states, such as China, Iran, or Russia, whose members depend on coded language to be able to exchange information. Since the beginning of the Russian invasion of Ukraine, even small expressions of dissent have been suppressed as 'defamation of the military' by Putin's regime (Maynes 2023). Many oppressive regimes frame critical speech as 'support of terrorist activities' and use these allegations to silence the political opposition.

If AI systems succeed at efficient and precise hate speech detection, then political and business actors who are hostile to democratic values and human rights can use the same tools to suppress, censor, and persecute what they consider 'hate speech', that is, all speech they perceive as a threat to their own political ends. The same AI that would flag coded racist messages could and would then also be used to flag the coded messages of citizens trying to evade the censorship efforts of their government (we can think of the protesters in Iran or dissidents in China).

These reframing efforts are not limited to autocratic states, however. We also see them on a smaller scale, for instance, in widespread complaints about 'cancel culture' that aim to silence counterspeech. The rhetoric is universally one of victimhood, the pretense of a 'woke' minority oppressing the speech of a 'silent majority', and it is used by religious fundamentalists, racists, nationalists, sexists, and self-identified 'gender-critical feminists' alike. That is to say, the tactic is known, and it is being employed at the government level (Russia's war propaganda is unfailingly one of victimhood against 'Western aggression'). If we assume that AI-mediated hate speech detection will also become available to these kinds of actors, then it stands to reason that this will be a significant support for their efforts to suppress counterspeech. This would be disastrous especially for countries and geographical regions in which the sharing of resistant ideas relies heavily on social media (Tufekci 2017). And the problem is not limited to states and other political actors; it also applies to corporate actors, who already have immense and effectively unchecked power to act as censors (Langvardt 2018: 1355).

The ethical dilemma, then, is this: there are good reasons to desire responsible online content moderation on social media. Given the sheer amount of content on social media, achieving this aim necessitates the responsible use and improvement of AI for hate speech detection. At the same time, the improvement of AI in this field is likely to lead to the more effective suppression of politically, socially, and ethically valuable counterspeech. Therefore, there are also good reasons to resist the improvement of AI in this field, and to accept its current flaws for the sake of preserving resistant pockets of speech. This is a real moral dilemma (Kvalnes 2019: 11-19), as it pits the values of free speech against each other: the ability to engage in epistemically and socially valuable communication on well-moderated platforms against the need to be free from governmental and corporate efforts to censor critical speech.

References:

Arsht, A. and D. Etcovitch (2018). "The Human Cost of Online Content Moderation," Harvard Journal of Law and Technology, JOLT Digest, March 2, https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation (accessed 1 April 2023).

Barrett, P. M. (2020). Who Moderates the Social Media Giants? A Call to End Outsourcing. NYU Stern Center for Human Rights and Business.

Celi, L. A. et al. (2022). "Sources of Bias in Artificial Intelligence that Perpetuate Healthcare Disparities: A Global Review," in PLOS Digital Health, doi: 10.1371/journal.pdig.0000022.

Cortiz, D. and A. Zubiaga (2021). "Ethical and Technical Challenges of AI in Tackling Hate Speech," in The International Review of Information Ethics 29, https://informationethics.ca/index.php/irie/article/view/416/389 (accessed 16 November 2022).

Das., M. et al. (2022). "HateCheckHIn: Evaluating Hindi Hate Speech Detection Models," Proceedings of the 2022 Language Resources and Evaluation Conference, LREC 2022, 5378-5387.

Dias Oliva, T., D. M. Antonialli, A. Gomes (2020). “Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online,” in Sexuality & Culture 25, 700-732.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderations, and the Hidden Decisions That Shape Social Media. Yale University Press.

Glozer, S., E. J. Godwin, R. Mota (2022). “Twitter and Elon Musk: Why Free Speech Absolutism Threatens Human Rights,” The Conversation, 7 November, https://theconversation.com/twitter-and-elon-musk-why-free-speech-absolutism-threatens-human-rights-193877 (accessed 1 April 2023).

Gorwa. R., R. Binns, C. Katzenbach (2020). “Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance,” in Big Data and Society, doi: 10.1177/2053951719897945.

Grimmelmann, J. (2015). “The Virtues of Moderation,” Yale Journal of Law and Technology 17, 42-108.

Ibrahim, M. A. (2022). “An Explainable AI Model for Hate Speech Detection on Indonesian Twitter,” in CommIT Journal 16, 175-182.

Jiménez-Durán, R. (2022), “The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter,” New Working Paper Series, No. 324, University of Chicago Booth School of Business, Stigler Center for the Study of the Economy and the State.

Knight, W. (2022). “Elon Musk Has Fired Twitter’s ‘Ethical AI’ Team,” on Wired, 4 November, https://www.wired.com/story/twitter-ethical-ai-team/ (accessed 1 April 2023).

Kvalnes, Ø. (2019). Moral Reasoning at Work: Rethinking Ethics in Organizations. Second Edition, Palgrave Macmillan.

Land, M. K., and L. Helfer (2022). “Value Pluralism and Human Rights in Online Content Moderation,” Lawfare Blog, 27 October, https://www.lawfareblog.com/value-pluralism-and-human-rights-content-moderation (accessed 1 April 2023).

Langvardt, K. (2018). “Regulating Online Content Moderation,” The Georgetown Law Journal 106, 1353-1388.

Mac, R. et al. (2022). “A Verifiable Mess: Twitter Users Create Havoc By Impersonating Brands,” The New York Times, 11 November, https://www.nytimes.com/2022/11/11/technology/twitter-blue-fake-accounts.html (accessed 1 April 2023).

Maynes, C. (2023). “After a Russian Girl Drew an Antiwar Poster, Her Dad Faces Defamation Charges,” NPR All Things Considered, 28 March, https://www.npr.org/2023/03/28/1166630360/after-a-russian-girl-drew-an-antiwar-poster-her-dad-faces-defamation-charges (accessed 1 April 2023).

Milmo, D. and A. Hern (2022). “Twitter Bans Comedian Kathy Griffin For Impersonating Elon Musk,” The Guardian, 7 November, https://www.theguardian.com/technology/2022/nov/07/twitter-will-ban-permanently-suspend-impersonator-accounts-elon-musk-says-as-users-take-his-name (accessed 1 April 2023).

Sap, M. et al. (2019). “The Risk of Racial Bias in Hate Speech Detection,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1668–1678.

Saul, J. (2017). “Racial Figleaves, the Shifting Boundaries of the Permissible, and the Rise of Donald Trump,” in Philosophical Topics 45 (2), 97-116.

Saul, J. (2018). “Dogwhistles, Political Manipulation, and Philosophy of Language,” in D. Fogel et al. (eds.): New Work on Speech Acts, Oxford University Press, 360-383.

Sellars, A. F. (2016) “Defining Hate Speech.” Berkman Klein Center Research Publication No. 2016-20. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244 (accessed November 16, 2022).

Spangler, T. (2022). “Elon Musk Bans Twitter Account That Tracked His Private Jet, After Claiming He Wouldn’t,” Variety, 14 December, https://variety.com/2022/digital/news/elon-musk-bans-twitter-elonjet-account-1235461331/ (accessed 1 April 2023).

Steiger, M. et al. (2021). “The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support,” Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Article No. 341, doi: 10.1145/3411764.3445092

Stolton, S. et al. (2022). “Elon Musk Triggers Twitter Chaos With Mass Firings Worldwide,” Politico, 4 November, https://www.politico.eu/article/twitter-fires-employs-worldwide/ (accessed 1 April 2023).

Tufekci, Z. (2017). Twitter and Tear Gas. The Power and Fragility of Networked Protest, Yale University Press.

Wiegand, M., J. Ruppenhofer, E. Eder (2021). “Implicitly Abusive Language. What does it actually look like and why are we not getting there?” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 576–587.

Udupa, S.(2022). “Ethical Scaling for Content Moderation. Extreme Speech and the (In)Significance of Artificial Intelligence.” Publication of the Shorenstein Center for Media, Politics, and Public Policy, Harvard Kennedy School, https://shorensteincenter.org/ethical-scaling-content-moderation-extreme-speech-insignificance-artificial-intelligence/ (accessed November 16, 2022).

Ugwudike, P. (2021). “AI Audits For Assessing Design Logics And Building Ethical Systems: The Case of Predictive Policing Algorithms,” in AI and Ethics 2, 199-208.

Yin, W. and A. Zubiaga (2022). “Hidden behind the Obvious: Misleading Keywords and Implicitly Abusive Language on Social Media”, in Online Social Networks and Media 30, https://doi.org/10.1016/j.osnem.2022.100210.