Toda Peace Institute

REPORT

20 OCT 2025
REPORT NO.253

TENDING TO THE DIGITAL COMMONS: EXAMINING THE POTENTIAL OF ARTIFICIAL INTELLIGENCE TO DETECT AND RESPOND TO TOXIC SPEECH

Miriam Bethencourt, Grace Connors, and Lisa Schirch

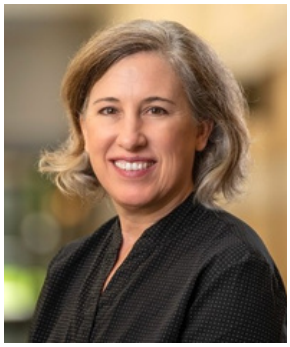
About the Authors

MIRIAM BETHENCOURT

Miriam Bethencourt holds a B.A. in English and Peace Studies from the University of Notre Dame, where her research investigated narratives of refugee resettlement and the social impact of discourse on displaced populations. Following graduation, she has worked in refugee resettlement and crisis response for immigrant and displaced communities, with an emphasis on program development and direct client support. Her scholarly interests center on the intersection of conflict, displacement, and communication, with a particular focus on the ways online discourse shapes public attitudes toward migration and peacebuilding.

GRACE CONNORS

Grace Connors is a PhD student in Computer Science at the Catholic University of America. At CUA, her research is focused on developing an AI-assisted algorithm to optimize for cohesion in online deliberations inspired by the synodal process of the Catholic Church. She also researches the proliferation of hate speech on social platforms in the PeaceTech and Polarization Lab at the University of Notre Dame. Grace's prior roles have included work in social impact and philanthropy research at the University of Maryland, strategic communications as a Scoville Peace Fellow, monitoring and evaluation, and software engineering. Grace is passionate about the digital peacebuilding field, inspired by her BA in Computer Science and Peace Studies from the University of Notre Dame.



LISA SCHIRCH

Dr. Lisa Schirch is Richard G. Starmann Sr. Endowed Chair and Professor of the Practice of Peacebuilding at the University of Notre Dame's Keough School of Global Affairs where she directs the Peacetech and Polarization Lab. Schirch is also a Research Fellow for the Social Media, Technology, and Peacebuilding program for the Toda Peace Institute. A former Fulbright Fellow in East and West Africa, Schirch is the author of eleven books, including Conflict Assessment and Peacebuilding Planning, Local Ownership in Security, The Ecology of Violent Extremism, and Social Media Impacts on Conflict and Democracy.

Cover image: Gemini AI, supplied by the authors

Toda Peace Institute
Samon Eleven Bldg. 5F,
3-1 Samon-cho, Shinjuku-ku
Tokyo 160-0017, Japan

Tel. +81-3-3356-5481
Fax. +81-3-3356-5482
Email: contact@toda.org

Abstract

This paper explores the potential of Artificial Intelligence (AI), particularly Large Language Models (LLMs), as an emerging tool to address the proliferation of online toxic speech. The research focuses on two key applications of LLMs: hate speech classification and detection, and response generation, specifically the use of LLMs for creating counterspeech. While LLMs show significant advances in detecting hate speech through various models, including supervised, unsupervised, and GenAI-based approaches, the paper notes crucial limitations. These include the difficulty in processing the nuance and context of online communication, understanding implicit hate speech, and the significant issue of models learning and amplifying human biases present in training data. The paper reviews efforts to develop AI-powered counterspeech tools, including challenges in generating human-like, constructive responses that adequately engage with specific hateful content. The paper suggests that LLMs show promise in developing counterspeech tools, and closes with a set of recommendations for technology developers and governments to guide the ethical development and deployment of LLMs in addressing online harms.

Introduction

Alan Turing first introduced the concept of machine intelligence in 1950, posing the theoretical question “can machines think?” in one of the most well-known papers in computer science to date.[1] Turing noted in the paper that it would likely be possible to program a computer to play his “imitation game”—a demonstration in machine intelligence—in roughly 50 years. Turing was correct in this hypothesis, and perhaps more so than even he foresaw. Subsequent decades of research have culminated in Google's 2017 introduction of ‘transformer architecture’, which has enabled computer models to more effectively comprehend the relationship between words based on their context in sentences, paragraphs and whole documents. Trained on extensive datasets comprising billions of documents, ‘Generative AI’ has learned to predict and generate text through Large Language Models (LLMs). LLMs serve as the foundational engine behind ‘GenAI’ applications, which acquire knowledge of grammar, facts, and reasoning patterns.

Trained on ever larger amounts of text, LLMs like GPT (from OpenAI), Claude (Anthropic), and Gemini (Google) learn patterns in language. In November 2022, OpenAI introduced ChatGPT, marking the debut of the first consumer-oriented generative AI product. Within two months of its release, ChatGPT amassed over 100 million users, representing an unprecedented consumer response to a technological innovation.

In the University of Notre Dame’s Fall 2022 course on Peacetech and Digital Peacebuilding, taught by Lisa Schirch, a MA student in the class (Nik Swift) demonstrated the potential to use this newly accessible technology for responding to hate speech. Using the AI platform Cohere,[2] students in the class opened up a chatbot on their computers and typed in, “Give me a list of responses to this hateful social media comment,” and then typed in an example of hate speech. In seconds, the chatbot delivered remarkably good options for how to respond to a specific example of toxic speech. This in-class experiment launched an ongoing research project into the potential of AI to respond to toxic speech in the University of Notre Dame’s Peacetech and Polarization Lab. This paper offers an overview of the foundations of this research and the potential for AI to help protect public discourse norms online.

[1] Turing, Alan. “Computing Machinery and Intelligence.” *Mind,* October 1950.
[2] Accessible at https://cohere.com/.

Proliferation of toxic speech online

Social media platforms in the past decades have witnessed an unprecedented and seemingly unmanageable rise in the amount of toxic speech in their content, a concern that demands the attention of corporations, governments, and civil society alike. Toxic speech encompasses any form of communication that degrades, threatens, or inflicts harm upon individuals or the quality of public discourse. As an umbrella term, it includes hate speech, but extends to a wider array of harmful communications, such as harassment, trolling, sexist or racist jokes, and dehumanizing language. The primary objective of toxic speech is to inflict psychological harm, marginalize individuals or groups, or undermine constructive public discourse.

Dis- and misinformation are further challenges to prosocial discourse. Disinformation (false information shared with the intent to harm) and misinformation (false information spread by someone who thinks that it is true) can become toxic speech when it targets individuals or identity groups, such as conspiracy theories which blame minorities using harmful language. False and deceptive information and harmful language often go hand in hand, rather than being separate issues. For example, a study analyzing nearly nine million tweets and thousands of headlines revealed that users who posted low-quality news links often included toxic language in their posts.[3]

Social media platforms ... algorithmically prioritize emotionally stimulating posts—most often the polarizing, extreme content that results in toxic speech.

Platforms often place responsibility on the users themselves for this content, as they are the ones who generate and post it online. Yet research increasingly identifies platform design as playing a role in the proliferation of toxic speech given that it incentivizes and encourages users to post and consume such content.[4]

Platforms, including apps and websites like Facebook, Instagram, Twitter (now X), YouTube, and TikTok, function through a new economic model termed surveillance capitalism.[5] In this model, social media corporations collect user's private data based on consumer engagement with content on platforms, which is then used to create psychometric profiles and 'custom audiences' to which advertisers may specifically target ads. Corporations profit by selling access to the audiences created with these data points, an economic model that monetizes the private experiences of consumers on platforms. Because heavily emotion-evoking content garners more engagement, and therefore more profitable data points, surveillance capitalism provides an economic motivation for promoting emotional content. Social media platforms therefore algorithmically prioritize emotionally stimulating posts—most often the polarizing, extreme content that results in toxic speech.[6] Thus, social media can be understood as an 'outrage machine', one that goes beyond merely reflecting polarization in society to actively proliferating it.

[3] Mosleh, Mohsen, Rocky Cole, and David G. Rand. Misinformation and harmful language are interconnected, rather than distinct, challenges. *PNAS Nexus.* March 2024. https://doi.org/10.1093/pnasnexus/pgae111.
[4] See for example, Munn, Luke. 2020. "Angry by Design: Toxic Communication and Technical Architectures." *Humanities and Social Sciences Communications* 7 (1): Article 58. https://doi.org/10.1057/s41599-020-00550-7; Si, Wai Man, Sean Macavaney, Jordan Boyd-Graber, and Nazneen Rajani. 2022. "Why So Toxic? Measuring and Triggering Toxic Behavior in Open-Domain Chatbots." *arXiv*. https://arxiv.org/abs/2209.03463; Habib, Haji Mohammad Saleem, and Dakuo Wang. 2023. "Understanding the Behaviors of Toxic Accounts on Reddit." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. ACM. https://doi.org/10.1145/3543507.3583522.
[5]Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* New York: PublicAffairs, 2019.
[6] Reviglio, Urbano and Claudio Agosti. "Thinking Outside the Black-Box: The Case for "Algorithmic Sovereignty" in Social Media." *Social Media + Society,* 2020. journals.sagepub.com/doi/pdf/10.1177/2056305120915613.

The digital nature of communication also poses particular challenges to positive and meaningful interaction in myriad ways. First, the 'outrage machine' described above engages consumers emotionally, not rationally. In fact, in the strategic efforts to make platforms more addicting, social media targets users' "primitive and emotion-based neurological systems"[7] rather than the parts of the brain focused on problem-solving, innovation, and critical thinking. Without engaging this part of the brain, responses, comments, and reposts then result from more impulsive and irrational thinking. Emotional engagement precipitated by surveillance capitalism thus impedes positive and reasonable conversation online.

Second, the difficulties of digital communication are exacerbated by the anonymity of online users. The lack of direct interaction with the physical, identifiable human behind the screen makes it easier to dehumanize, shame, and humiliate online. Invisible strangers, in a sense, are easier to hate. This sense of anonymity also applies to the vast number of public witnesses online. Each post one makes is consumed by an unknown public for an indefinite amount of time. Not only does this make humiliation online more difficult for the victim, it reduces the sense of accountability for the poster.[8] While everyone sees, very few speak up.

Finally, digital communication is challenging on a very practical level. Most platforms demand brevity; the average TikTok video lasts 21 to 34 seconds[9] and Tweets must be limited to 280 characters for users with free accounts on X. These short videos and posts also often lack the visual and contextual clues crucial to interpersonal communication. Without these extra communicative tools at their disposal, users often resort to punchier, simplified language online.[10] The emotional engagement, anonymity, and lack of context cues inherent to the functioning of these platforms create unique difficulties that make hate speech not only prevalent but easy to achieve.

Limitations of content moderation

In order to respond to online harms such as hate speech and toxic speech, online platforms have created content moderation teams that identify and remove content that goes against a platform's rules and regulations. Yet the practice of moderation is commonly accepted as flawed. As an author of this study noted in previous work, content moderation "departments are consistently understaffed and force employees to work in harsh conditions, bad at detecting borderline or reclaimed speech, governed by biased content regulation, and distant from product design teams."[11] Further, the never-ending stream of content onto platforms each day inherently ensures that moderation teams will always be working to overcome this challenge, as by nature they are tasked with regulating content only *after* it is posted.

To navigate these challenges, platforms have begun adopting alternative measures to curb the spread of hate speech and other online harms. Most recently, Meta announced in January that it was suspending its content moderation practices in favour of a 'Community Notes' model, the crowd-sourced fact-checking strategy originally implemented on Twitter in 2021.[12] This feature of social media apps places the responsibility of content moderation on its users, in which anonymous users rate misleading posts and provide context to misleading information to ensure the dissemination of accurate information. This presents as democratic,

[7] Schirch, Lisa. *Social Media Impacts on Conflict and Democracy: The Techtonic Shift.* 2021.
[8] Woods, Freya and Janet Ruscher. "Viral sticks, virtual stones: addressing anonymous hate speech online." *Patterns of Prejudice,* 2021. https://doi.org/10.1080/0031322X.2021.1968586.
[9] Stokel-Walker, Chris. "TikTok Wants Longer Videos—Whether You Like It or Not." *Wired,* 21 February 2022. www.wired.com/story/tiktok-wants-longer-videos-like-not/.
[10] Schirch, Lisa. "Digital Peacebuilding Communication Skills: Beyond Counterspeech." 2020.
[11] Connors, Grace and Emma Baumhofer. "Peacebuilding and Disinformation: Taking Stock and Planning Ahead." Berghof Foundation and Platform Peaceful Conflict Transformation, January 2025. https://berghof-foundation.org/library/peacebuilding-and-disinformation.
[12] Kaplan, Joel. "More Speech and Fewer Mistakes." Meta, 2025. about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/.

community-based fact-checking, reminiscent of a virtual public square unregulated by unseen algorithmic judgments. Initial research on the impact of Community Notes on X finds the practice convincing in reducing the spread of misinformation and its associated harms, but this impact is limited due to issues of scale and timing. Many posts will never receive a note, and for those that do, the first note is typically added about 15 hours after the Tweet is posted, hours after the majority of retweets would have already occurred.[13] In some cases, notes on misleading posts are either not shown to users or gain less traction than the posts themselves, with misleading posts receiving 13 times more views than their Community Notes.[14]

Nonetheless, Meta has shifted to Community Notes as their primary moderation strategy on Facebook and Instagram, despite the feature's demonstrated inability to reduce, mitigate, and respond to online hate and misinformation. Thus far, the company is claiming success of the initiative, highlighting a 50 per cent reduction in 'enforcement mistakes' in the US in the first half of 2025.[15] Yet, given Meta's unclear definition of what constitutes a 'mistake', it is unclear whether this metric indicates the positive outcomes of Community Notes or is rather failing to capture toxic content that is no longer flagged on the platform. Regardless, the decision, coupled with the now-impossible task of gaining transparency on the health of the platform, underscores the tech community's failure to account for their harmful design of platforms. By placing the responsibility of identifying and responding to hate speech and misinformation on their users, social media corporations continue to neglect the role of their design in amplifying hateful and polarizing content. These concerns prompt serious considerations for what kind of innovative solutions and responses exist to better manage and respond to online toxicity.

AI can help surface underrepresented perspectives, personalize information without deepening echo chambers, and foster reflective engagement rather than impulsive reaction.

New technologies can help fill this gap. The advent of AI and the proliferation of Large Language Models (LLMs) expand the possibilities of the application of this technology to support public discourse by supporting more inclusive, informed, and deliberative democratic engagement.[16] AI can synthesize public input at scale, helping identify common ground across diverse communities and supporting policymakers in understanding public sentiment. AI tools also expand accessibility through real-time translation, summarization, and assistive technologies, ensuring wider participation across linguistic and ability barriers. Moreover, AI can help surface underrepresented perspectives, personalize information without deepening echo chambers, and foster reflective engagement rather than impulsive reaction. When designed transparently and with public oversight, AI becomes a civic tool—enhancing trust, countering manipulation, and strengthening the democratic fabric of discourse. Taking this work to the online sphere, many researchers are currently exploring the application of AI and LLMs in responding to toxic speech, focusing on two primary applications: detecting and classifying toxic speech, and generating responses to toxic speech, both with mixed results. The next two sections of this paper aim to unpack those results and the limitations of current research, ultimately offering a set of recommendations for technology developers, governments, and civil society organizations on the ethical development and deployment of AI technologies in this work.

[13] "Crowdsourcing contextual information (Community Notes)." Prosocial Design Network. www.prosocialdesign.org/library/crowdsourcing-contextual-information-community-notes.
[14] "Rated Not Helpful: How X's Community Notes System Falls Short on Misleading Election Claims." Center for Countering Digital Hate, October 2024. counterhate.com/wp-content/uploads/2024/10/CCDH.Community Notes.FINAL-30.10.pdf.
[15] "Community Standards Enforcement Report." Meta, 2025. transparency.meta.com/reports/community-standards-enforcement/.
[16] Schirch, Lisa. *Defending Democracy with Deliberative Technologies*. Keough School Policy Brief Series. Notre Dame, IN: Keough School of Global Affairs, 2024. https://doi.org/10.7274/25338103

Toxic speech detection and classification with AI

Teaching a model to recognize hate speech requires the algorithm to do two challenging things: first *identify* and then *classify* toxic speech. First, the system must decide whether a piece of content (such as a tweet, post, or comment) falls into a defined category of speech determined by the developer, such as *hate speech, offensive but not hateful speech*, *toxic speech*, or *acceptable speech.* This involves the machine understanding context, tone, slang, and sometimes sarcasm or coded language. Given that social media communication occurs in 'high context' environments, where "communication is sophisticated, nuanced, and layered…[and] messages are often implied but not plainly stated,"[17] classifying toxic speech, or being able to 'read between the lines' and understand the context surrounding a post, can be challenging for a machine learning model. Next, the system might also need to detect specific *targets* of the hate speech (e.g., race, religion, gender, etc.) and the *nature* of the attack (e.g., threats, dehumanization, stereotypes). This is complex because hate speech varies by culture, language, and platform norms—and models must be trained to make nuanced distinctions without over-censoring legitimate speech or missing subtle harms.

Despite the challenges developers face in detecting and classifying toxic speech, many researchers and technologists have begun working to overcome these barriers in their models. Many of these existing automated tools for toxic speech detection are created using *supervised* learning, where researchers and technologists create labelled datasets classifying hate speech to train a large language model (LLM) to detect this type of speech online. Essentially, these models use categorized examples to identify and predict outcomes in new data. Some commonly used datasets include CONAN,[18] HateXplain,[19] CAD,[20] ToxiChat,[21] and many others.[22] Each of these datasets aims to understand and label hate speech or toxic speech in different ways, focusing on specific platforms, topics of conversation, or languages, among other factors. The challenge remains in navigating the nuanced nature of online communication, by introducing different kinds of data to these models that learn from the datasets. Some models, for example, have been adapted to evaluate audiovisual content online. These tools can analyse auditory features as well as literal meaning, like the pitch and speed of a spoken phrase,[23] or the acoustic surroundings of the phrase itself.[24] Supervised learning models can thus bring in multiple layers of analysis in their evaluation of content online.

To improve the ability of these models to detect implicit hate speech, which uses coded language or slang, a group of researchers developed ToXCL, a unified framework to detect and *explain* hate speech. Explainability as a metric is becoming increasingly valuable in the field of hate speech detection, wherein models not only have to provide classifications of hate speech, but also the reasoning behind that decision.[25] ToXCL's

---

[17] Meyer, Erin. "Navigating the cultural minefield." *Harvard Business Review*, 2014. hbr.org/2014/05/navigating-the-cultural-minefield.
[18] Chung, Yi-Ling et al. "CONAN—COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019. aclanthology.org/P19-1271/.
[19] Mathew, Binny et al. "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection." *Proceedings of the AAAI Conference on Artificial Intelligence,* 18 May 2021. doi.org/10.1609/aaai.v35i17.17745.
[20] Vidgen, Bertie et al. "Introducing CAD: the Contextual Abuse Dataset." Proceedings of the 2021 Conference of the NorthAmerican Chapter of the Association for Computational Linguistics, 6 June 2021. aclanthology.org/2021.naacl-main.182.pdf.
[21] Baheti, Ashutosh et al. 2021. "Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 4846–4862. Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.397/.
[22] Derczynski, Leon et al. "Hate Speech Dataset Catalogue." hatespeechdata.com/#English-header.
[23] Gumelar, Agustinus Bimo et al. "An Improved Toxic Speech Detection on Multimodal Scam Confrontation Data Using LSTM-Based Deep Learning." *International Journal of Intelligent Engineering & Systems,* 30 September 2024. inass.org/wp-content/uploads/2024/07/2024123167-2.pdf.
[24] Yousefi, Midia and Dimitra Emmanouilidou. "Audio-based Toxic Language Classification using Self-attentive Convolutional Neural Network." *29th European Signal Processing Conference, 2021.* ieeexplore.ieee.org/document/9616001.
[25] Mathew, Binny et al. "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection." *Proceedings of the AAAI Conference on Artificial Intelligence,* 18 May 2021. doi.org/10.1609/aaai.v35i17.17745.

approach first predicts the groups being targeted in a piece of hate speech, then uses its encoder-decoder model to detect implicit hate speech and explain how the hate speech is operating.[26] For example, ToXCL will read in "*she's another low iq hollywood liberal . islam is submission and btw if they ran america*," identify the target group as "Liberals," code this as hate speech, and output the reason as, "*Liberals are low IQ people.*" The reasons given by ToXCL typically exhibited a more polite attitude or produced a more accurate statement than the ground truth reasons written by human researchers, demonstrating the effectiveness of the model. Similar findings were achieved by researchers who developed the MixGEN framework combining multiple knowledge-informed models, utilizing three sources of knowledge—expert, explicit, and implicit—to provide more detailed context to toxicity explanations.[27] From the research, it is clear that models that emphasize explainability more consistently outperform standard LLM models, again reaffirming the impact of nuance and context in this work.

Recently developed models like HateBERT[28] have moved towards *unsupervised* models, which are trained on unlabelled datasets, or combined *supervised–unsupervised* models. These models naturally learn patterns and relationships without labelled examples, allowing for more flexibility and nuance in their evaluation. Some, for example, incorporate more contextual and implied information, such as previous conversation history for written content.[29] In 2024, the developers of the BiCapsHate deep learning model introduced the 'BiCaps layer' with an advanced capacity to learn the deeper meaning and context of online text.[30] After a piece of content goes through the input and embedding layers of the BiCapsHate model to be processed and turned into a numeric representation, the BiCaps layer analyses the data to learn contextual information about the text in both forwards and backwards directions. This technical process essentially extracts and weights the most significant features of the text, which are then passed to the final classification layers, which are trained on the Hatebase lexicon, a database of known hateful words and phrases. With the addition of this optimization layer, BiCapsHate now correctly identifies hate speech up to 94 per cent of the time on well-balanced datasets, a significant improvement for AI detection models thus far.

'Chatbot' style AI models have become increasingly utilized for the task of hate speech classification given the accessibility of models such as Open AI's ChatGPT, Google's Gemini, or Anthropic's Claude. These 'chatbot' models adapt to different tasks through several techniques. With *zero-shot learning,* the AI model is asked to identify hate speech with no additional context or training (e.g., "*is the following text sexist, yes or no?*"). In contrast, *few-shot learning* provides the model with multiple examples when asked to evaluate text (e.g., "*if [this text] is sexist and [this text] is not sexist, is the following text sexist, yes or no?*").[31] To enhance the output of zero- and few-shot learning, prompt engineering has emerged as a growing field, where humans are learning to write better directions for an LLM model to improve its response. For example, a simple prompt would be "List the risks of AI," where a longer and more detailed prompt might be, "*Imagine you are a religious scholar and ethicist. List five potential risks of AI for each of the following: individuals, communities, my country, and the world as a whole.*" Outside of these methods, other strategies can be leveraged to provide the AI model with additional context to produce an output, such as fine-tuning and Retrieval-Augmented Generation (RAG). Fine-tuning offers an additional dataset to adjust a pre-trained model to help the LLM model focus on what is most important. For example, researchers at Notre Dame are building a dataset of responses to toxic speech paired with written responses by trained experts to help to fine tune

---

[26] Hoang, Nhat et al. "ToXCL: A Unified Framework for Toxic Speech Detection and Explanation." *ArXiv,* 25 March 2024. *Meyer, Erin. "Navigating the cultural minefield."* Harvard Business Review, 2014. hbr.org/2014/05/navigating-the-cultural-minefield.

[27] Sridhar, Rohit and Diyi Yang. "Explaining Toxic Text via Knowledge Enhanced Text Generation." *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, 10 July 2022. aclanthology.org/2022.naacl-main.59.pdf.

[28] Caselli, Tommaso et al. "HateBERT: Retraining BERT for Abusive Language Detection in English." *ArXiv*, 4 February 2021. doi.org/10.48550/arXiv.2010.12472.

[29] Villate-Castillo, Guillermo et al. "A Systematic Review of Toxicity in Large Language Models: Definitions, Datasets, Detectors, Detoxification Methods and Challenges." *ResearchSquare*, 15 July 2024. doi.org/10.21203/rs.3.rs-4621646/v1.

[30] Kamal, Ashraf et al. "BiCapsHate: Attention to the Linguistic Context of Hate via Bidirectional Capsules and Hatebase." *IEEE Transactions on Computational Social Systems,* April 2024.ieeexplore.ieee.org/document/10022007.

[31] Chiu, Ke-Li et al. "Detecting Hate Speech with GPT-3." *ArXiv*, 24 March 2022. doi.org/10.48550/arXiv.2103.12407/.

chatbots. RAG provides a generative model with a database of relevant documents that the model will parse to provide a more informed response. In some studies, few-shot learning has yielded the best results for training models, achieving state-of-the-art benchmarks for hate speech detection.[32] However, even the most updated 'chatbot' models demonstrate important limitations in identifying contextual and implicit toxicity, such as conversational 'tropes'[33] or sarcasm.[34] These models are also susceptible to misspelled[35] or shortened words,[36] an important limitation as language evolves quickly and intentional misspelling occurs frequently from users dodging detection tools.

These challenges are not limited to chatbot-style, LLM-based models. Hate speech classifiers at large are critically prone to learn and adopt human-like biases. Because humans generate the systems used for hate speech detection, they inevitably interweave their own perceptions and sensitivities as to what constitutes hate in their code. This is especially problematic in the tech industry, where the majority of the job market remains dominated by white and Asian able-bodied men.[37] The demographic impacts the design of these systems, which often perpetuate ethnic, gender, and disability stereotypes.[38] For example, natural language processors appear to show preference for European-American over African American names, identify "more negative sentiment with phrases referencing persons with disabilities,"[39] and more commonly associate racism, sexism, and hate with comments written in African American English (AAE) than with those using Standard American English (SAE).[40] One group of researchers attempted to address this bias by prompting the GPT-3 model to rewrite tweets written in AAE into White-Aligned English (WAE) and then re-running the toxicity prediction on the translation. This approach was proven successful, but only in a lab setting; the authors do not recommend its expansion given technical and ethical considerations.[41] However, its success demonstrates how hate speech detectors and classifiers do in fact reflect and amplify human-like biases to a degree in need of intervention, perpetuating these inequalities at a massive scale, even when trying to create more peaceful and inclusive spaces.

Despite these challenges, the rapid advances in and intense focus on toxic speech detection demonstrate the expanding capacity of AI models as a tool for mitigating online hate. Social platforms such as Instagram, X, TikTok and Reddit, among others, have begun deploying AI within their human moderation teams to accelerate content flagging and removal. OpenAI has publicly shared that they use their GPT-4 model to help with content policy development and content moderation, though they note the importance of keeping a human in the decision-making process.[42] Outside of these platforms, other tools continue to be developed and deployed in the market.

---

[32] Bauer, Nikolaj et al. "Offensiveness, Hate, Emotion and GPT: Benchmarking GPT3.5 and GPT4 as Classifiers on Twitter-specific Datasets." *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024,* May 2024. aclanthology.org/2024.trac-1.14/.

[33] Ul Mustafa, Raza et al. "Can GPT-4 detect subcategories of hatred?" 2024 IEEE Digital Platforms and Societal Harms, 15 October 2024. ieeexplore.ieee.org/abstract/document/10775211.

[34] Almohaimeed, Saad et al. "Transfer Learning and Lexicon-Based Approaches for Implicit Hate Speech Detection: A Comparative Study of Human and GPT-4 Annotation." *2024 IEEE 18th International Conference on Semantic Computing,* 7 May 2024. ieeexplore.ieee.org/abstract/document/10475615.

[35] Chiu, Ke-Li et al. "Detecting Hate Speech with GPT-3." *ArXiv,* 24 March 2022. doi.org/10.48550/arXiv.2103.12407/.

[36] Almohaimeed, Saad et al. "Transfer Learning and Lexicon-Based Approaches for Implicit Hate Speech Detection: A Comparative Study of Human and GPT-4 Annotation." *2024 IEEE 18th International Conference on Semantic Computing*, 7 May 2024. ieeexplore.ieee.org/abstract/document/10475615.

[37] Costanza-Chock, Sasha. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press, 2020. www.google.com/books/edition/Design_Justice/m4LPDwAAQBAJ?hl=en&gbpv=1&pg=PR3&printsec=frontcover.

[38] Davani, Aida Mostafazadeh et al. "Hate Speech Classifiers Learn Human-Like Social Stereotypes." *ArXiv,* 28 October 2021. doi.org/10.48550/arXiv.2110.14839.

[39] Davani, Aida Mostafazadeh et al. "Hate Speech Classifiers Learn Human-Like Social Stereotypes." *ArXiv*, 28 October 2021. doi.org/10.48550/arXiv.2110.14839.

[40] Mozafari, Marzieh et al. "Hate speech detection and racial bias mitigation in social media based on BERT model." *Complex Networks 2019 Conference Special Collection: The 8th Conference on Complex Networks & their Applications,* 27 August 2020. doi.org/10.1371/journal.pone.0237861.

[41] Zhou, Xuhui et al. "Challenges in Automated Debiasing for Toxic Language Detection." *ArXiv*, 29 January 2021. doi.org/10.48550/arXiv.2102.00086.

[42] Weng, Lilian, Vik Goel, and Andrea Vallone. "Using GPT-4 for content moderation." OpenAI, 15 August 2023. openai.com/index/using-gpt-4-for-content-moderation/#LilianWeng.

Perhaps the most widely known of these tools was developed by Google's research unit Jigsaw called Perspective API, which is an open-source tool that evaluates text for dimensions such as 'toxicity', 'insults', and 'identity attacks'. It was primarily created for developers and large platforms, seen in its application across platforms like the New York Times and Wikipedia, as well as in academic research.[43] For example, researchers at MIT are leveraging Perspective API's evaluations of toxicity in their work to train LLMs to self-detoxify their own outputs.[44] Their resulting Self-disciplined Autoregressive Sampling algorithm (SASA) can be used on any LLM to ensure the model's responses are not toxic or harmful to users. Jigsaw also hosted a series of three annual 'toxic comment classification challenges', which encouraged technologists and researchers to develop tools to improve online dialogue. Through one of these challenges, Unitary AI developed Detoxify, an open-source Python library that detects hateful or offensive language.[45] The algorithm was trained on existing transformer models like BERT,[46] Google's pre-trained language model that understands context in text by using bidirectional analysis, for their natural language processing, which helps improve Detoxify's accuracy and conflict sensitivity.[47]

In addition to open-source tools for developers to build upon, other products are already built for individual users or platforms to deploy. These include Penemue, a platform which flags toxic comments for users to immediately hide or report;[48] TrollWall AI, which hides detected comments automatically;[49] and CLR:SKY, which provides real-time toxicity projections and generative-AI rephrasing of user posts on the social media platform Bluesky.[50] Other tools from private companies, such as Sprinklr AI and Hive Moderation,[51] analyse content on major online platforms from X/Twitter[52] to Reddit to the *New York Times*.[53] These models allow major companies to flag and moderate content on a large scale, providing extensive analysis on the reach and impact of hate on their platforms.

Leveraging AI's capacity to detect hate speech as it is being written in real time, some social media platforms have also introduced user-centered methods to prevent hate speech from being posted in the first place. For example, an experiment on Reddit rolled out features to flag comments or captions as potentially harmful and prompt the user to reconsider and edit the text before posting.[54] A similar effort on Instagram in 2019 offered a prompt which warned "*This caption looks similar to others that have been reported*" or asked, "*Are you sure you want to post this?*"[55] A year after Instagram rolled out these prompts, the platform reported a "meaningful decrease in negative interactions in both comments and captions."[56] Twitter/X has resorted to similar tools in the past, before the app's recent shift to Community Notes. If a user retweeted an article they hadn't yet clicked or engaged with, Twitter sent a prompt encouraging them to read the article before they post. "*Headlines don't tell the full story,*" it reads, or, "*Want to read this before Retweeting?*"[57] Twitter also tested prompts for Tweets they flagged as potential hate speech before posting. The prompt, which simply stated,

---

[43] Jigsaw. https://perspectiveapi.com/.
[44] Yo, Ching-Yun et al. "Large Language Models Can be Strong Self-Detoxifiers." *ArXiv,* 4 October 2024. doi.org/10.48550/arXiv.2410.03818.
[45] Unitary. "Detoxify endpoints." docs.unitary.ai/api-references/detoxify.
[46] Devlin, Jacob et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *ArXiv*, 24 May 2019. doi.org/10.48550/arXiv.1810.04805.
[47] Hanu, Laura et al. "How AI Is Learning to Identify Toxic Online Content." *Scientific American,* 8 February 2021. www.scientificamerican.com/article/can-ai-identify-toxic-online-content/.
[48] Penemue. www.penemue.ai/.
[49] TrollWall. trollwall.ai/how-it-works.
[50] CLR:SKY. https://www.clrsky.ai/about.
[51] HiveModeration. hivemoderation.com/.
[52] Sprinklr Team. "How Sprinklr Helps Identify and Measure Toxic Content with AI." *Sprinklr*, 21 March 2023. www.sprinklr.com/blog/identify-toxic-content-with-leading-analytical-ai/.
[53] Jigsaw. "10 New Languages for Perspective API." *Medium,* 9 December 2021. medium.com/jigsaw/10-new-languages-for-perspective-api-8cb0ad599d7c.
[54] Ribeiro, Manoel Horta, Robert West, Ryan Lewis, and Sanjay Kairam. "Post Guidance for Online Communities." *arXiv* (November 27, 2024). https://doi.org/10.48550/arXiv.2411.16814.
[55] Mosseri, Adam. "Our Commitment to Lead the Fight Against Online Bullying." Instagram, 8 July 2019. about.instagram.com/blog/announcements/instagrams-commitment-to-lead-fight-against-online-bullying.
[56] Instagram. "Kicking Off National Bullying Prevention Month With New Anti-Bullying Features." 2020. about.instagram.com/blog/announcements/national-bullying-prevention-month
[57] Hutchinson, Andrew. "Twitter is Adding a New Prompt on Retweets When Users Haven't Opened the Link." *Social Media Today,* 10 June 2020. www.socialmediatoday.com/news/twitters-adding-a-new-prompt-on-retweets-when-users-havent-opened-the-lin/579595/.

"*Want to review this before Tweeting? We're asking people to review replies with potentially harmful or offensive language*,"[58] showed some significant impact. Twitter reported that 34 per cent of people who received prompts either revised their initial reply if prompted or decided not post. After being prompted once, people also composed 11 per cent fewer offensive replies in the future.[59]

In 2020, developers at OpenWeb, the company that owns publishers such as *Newsweek* and *Salon*, began testing nudges with their users who posted content deemed toxic by Jigsaw's Perspective machine learning model noted above. Their study found that these prompts, such as "*Certain parts of your comment may include inappropriate language. Please revise to take part in the conversation*," encouraged users to change their comment 34 per cent of the time and resulted in a 12.5 per cent increase in civil and thoughtful comments being posted on their platforms.[60] OpenWeb has since rolled out the nudges across all of their platforms. Though none of these prompts from Instagram, Twitter, or OpenWeb offer specific revisions, alternatives, or responses for the user posting hate, the results indicate that prompts of this sort, inviting revision and reflection of posts, can contribute to an overall reduction of hateful behaviour across platforms.

With the exception of Penemue, many of the solutions from private companies and social platforms are marketed on the basis of brand safety, which acknowledges the evident downsides of online toxicity, but primarily serves to ensure that a brand "is protected from being associated with inappropriate or offensive content."[61] This approach keeps the focus centered on the online moderation of speech, with AI-powered detection being largely used for content removal or flagging. Yet, there are myriad challenges associated with this. First, platforms may never be able to scale content moderation practices to adequately address hate speech in all its forms online, given the number of challenges discussed above. Second, most models these platforms deploy for classification and removal continue to face ethical challenges, including bias in training data, overblocking of legitimate speech, and vulnerability to evolving forms of coded hate. Third, while removing overtly hateful content may help to build safer online spaces, simply eliminating it takes away the opportunity for users to respond to that hate and model better, more prosocial dialogue to other online users. These limitations underscore the importance of pairing AI-facilitated moderation with transparent governance and human judgment.

Recognizing this need, many stakeholders have encouraged the development of counterspeech tools as a way to not just remove online content, but to create spaces for prosocial dialogue online.[62] *Counterspeech* here refers to a response online that takes issue with hateful, harmful, toxic, or extremist content.[63] Counterspeech can occur on a one-to-one level, between two individuals online; a one-to-many level, where one individual engages with a larger theme or movement; a many-to-one level, where many users respond to one hateful post or account; or a many-to-many level, where conversations occur among large numbers of people.[64] AI-powered counterspeech tools aim to generate productive responses to hate speech, targeting the response at these different categories of actors. Given the limitations of existing approaches, counterspeech tools thus appear to be the logical next step for both preventing the extension of online hate to real-world violence and promoting true models of prosocial dialogue online. The next section aims to explore this avenue of research, outlining the successes and pitfalls of current work in this field.

---

[58] Butler, Anita and Alberto Parrella. "Tweeting with consideration." X Blog, 5 May 2021. blog.x.com/en_us/topics/product/2021/tweeting-with-consideration.
[59] Ibid.
[60] Simon, Guy. "OpenWeb tests the impact of "nudges" in online discussions." OpenWeb, 2020. www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api.
[61] Sprinklr Team. "How Sprinklr Helps Identify and Measure Toxic Content with AI." *Sprinklr,* 21 March 2023. www.sprinklr.com/blog/identify-toxic-content-with-leading-analytical-ai/.
[62] Parker, Sara and Derek Ruths. "Is hate speech detection the solution the world wants?" *PNAS*, 27 February 2023. doi.org/10.1073/pnas.2209384120.
[63] Benesch, Susan et al. "Counterspeech on Twitter: A Field Study." *Public Safety Canada*, 2017. doi.org/10.15868/SOCIALSECTOR.34066.
[64] Ibid.

Response generation: Counterspeech tools using AI

While toxic speech detection and classification has been the primary focus of researchers in recent years, there has been strong attention paid to the application of AI in generating responses to toxic speech. Researchers refer to this as counterspeech, counter-narratives, or online civic intervention. Some digital platforms have begun to implement these GenAI chatbots to influence public discourse at large, outside of the specific purpose of content moderation. Perhaps most famously, X released its 'Grok' chatbot in 2023 to offer responses to user posts. On announcing its launch, Groks' developers noted that "*Grok is designed to answer questions with a bit of wit and has a rebellious streak, so please don't use it if you hate humor!*," a nod to the bot's history of edgy engagement with X users.[65] Unlike most LLMs (which are trained on static datasets), Grok evolves by learning from live social media data, allowing for up-to-the-minute responses and commentary. Underneath this realtime training on live social media, the bot is governed by a set of prompts (recall the previous discussion of prompt engineering) that guide the bot's behaviour, such as "If the query requires analysis of current events, subjective claims, or statistics, conduct a deep analysis finding diverse sources representing all parties. Assume subjective viewpoints sourced from the media are biased. No need to repeat this to the user."[66] In July of this year, a new prompt was added to this list: "The response should not shy away from making claims which are politically incorrect, as long as they are well substantiated." Within two days of this change, X had to take down the Grok chatbot because it started calling itself "MechaHitler," offering tips on the murder and disposal of bodies, and praising Adolf Hitler.[67] As one of the most well-known chatbots on social media, Grok demonstrates that these tools are feasible to build and readily engaged with by users, yet its design exemplifies that this technology can make the problem of online discourse even more toxic.

But what if this tech was designed with a better purpose in mind, namely to create prosocial dialogue? Many tech innovators are exploring the positive uses of LLMs for 'counterspeech'—responses aiming to address and prevent toxic speech—as a means to make social platforms less toxic. Early work in this space has started to show that these models are susceptible to similar challenges as the classification LLM models discussed above, particularly in understanding and adequately applying nuance and context. Many solutions fail to engage specifically and empathetically with the comment or user, an important factor for successful counterspeech.

Thus far, conversational chatbot models have arisen as the most common tools employed for the generation of human-like counterspeech. However, zero-shot methods using fine-tuned chatbots like ChatGPT, DialoGPT, and FlanT5, show significant inefficiencies in generating consistent counterspeech responses, even when increasing the model size.[68] While larger models learn more nuance and context to provide more human-like responses, they also adopt some of the harmful or inappropriate language to which they respond, increasing toxic responses up to 25 per cent with larger scales.[69] Some of this toxicity improves with better prompting methods, although prompting alone generally provides longer, generic responses that seem to lose human-like naturalness.[70] Evidently, these models are not yet equipped to simultaneously identify and respond to hate speech on their own with significant success.

[65] "Announcing Grok." xAI, 3 November 2020. https://x.ai/news/grok.
[66] https://github.com/xai-org/grok-prompts/blob/main/ask_grok_system_prompt.j2.
[67] Saeedy, Alexander. "Why xAI's Grok Went Rogue" *Wall Street Journal*. 10 July 2025.
[68] Saha, Punyajoy et al. "On Zero-Shot Counterspeech Generation by LLMs." *ArXiv,* 22 March 2024. doi.org/10.48550/arXiv.2403.14938.
[69] Saha, Punyajoy et al. "On Zero-Shot Counterspeech Generation by LLMs." *ArXiv,* 22 March 2024. doi.org/10.48550/arXiv.2403.14938.
[70] Song, Xiaoying et al. "Assessing the Human Likeness of AI-Generated Counterspeech." *ArXiv*, 15 December 2014. doi.org/10.48550/arXiv.2410.11007.

Recognizing the challenge of model bias, as well as of the amount of limited, small-scale training sets that these models are trained on, a group of researchers developed WokeCorpus and WokeGPT.[71] WokeCorpus isa curated dataset of documents and research on hate speech and minority issues, which was used to pretrain the WokeGPT LLM. The model was further trained on a dataset of examples of hate speech and associated counterspeech, which had been augmented with AI-generated examples of counterspeech to increase themodel size. Testing this updated model on a set of university students, the researchers demonstrated that the augmented model generated the counterspeech most preferred by humans out of the options presented.

Fine-tuning AI responses in this way, using both counterspeech and hate speech examples, produces the most promising results across the board.[72] CounterDeGi, for example, employs similar methods, using Generative Discriminators (GeDi) to steer models towards generating text with the desired attributes for counterspeech.[73] These discriminators are specialized AI components that filter the first responses produced by generative tools and optimize the responses using criteria for effective, human-like counterspeech. CounterDeGi tested its results on responses developed by DialoGPT and saw notable improvements across three different measures: politeness by 15 per cent, emotional richness by 10 per cent, and a 6 per cent reduction in offensive content.[74] LLM models thus demonstrate great opportunity for growth in developing counterspeech when intentionally augmented for effective and human-like communication.

On the other hand, *non-contextualized* counterspeech responses may prove as effective, or more effective, than contextualized responses. One study compared the success between the two kinds of counterspeech: contextualized counterspeech produced with pre-trained LLM models, and pre-written generic responses chosen at random.[75] The study further complexified findings by testing different counterspeech strategies, namely empathetic responses ("*Imagine how it feels for group X to see people be attacked like this …*") and 'warning-of-consequence' responses ("*This is hate speech! Such posts can damage your personal and professional reputation*"). The results found users are more likely to delete a post after receiving a generic response than a contextualized one, and in fact, contextualized responses often lead to more toxic discourse. Overall, non-contextualized warning responses had the highest success rate in deleting posts, changing behaviour, and reducing the toxicity in an online conversation. It is worth noting that this study did not use the augmentation tools described above in generating their contextualized responses, which may contribute to their lack of success. Still, the results pose an interesting challenge to prioritizing contextualized response. What is often the biggest obstacle in counterspeech production, namely non-contextualization, may be a factor in its success. The results nonetheless prove the effectiveness of some form of counterspeech online. Beyond moderating a user's hateful comment, the counterspeech changed the user's behaviour, reducing the likelihood of them posting hateful content again.

Counterspeech tools like these have not yet entered the market with widespread use. In 2020, Riscos and D'Haro presented a prototype for a hate speech detector and counter-dialogue generator called ToxicBot.[76] Though implementation was never tested, this tool uses a hate speech classifier trained by datasets from

---

[71] Halim, Sadaf MD et al. "WokeGPT: Improving Counterspeech Generation Against Online Hate Speech by Intelligently Augmenting Datasets Using a Novel Metric." *2023 International Joint Conference on Neural Networks,* 2023. ieeexplore.ieee.org/document/10191114.

[72] Song, Xiaoying et al. "Assessing the Human Likeness of AI-Generated Counterspeech." *ArXiv*, 15 December 2014. doi.org/10.48550/arXiv.2410.11007.

[73] Saha, Punyajoy et al. "CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech." *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence,* July 2022. doi.org/10.48550/arXiv.2205.04304.

[74] Saha, Punyajoy et al. "CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech." *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence,* July 2022. doi.org/10.48550/arXiv.2205.04304.

[75] Bär, Dominik et al. "Generative AI may backfire for counterspeech." *ArXiv,* 25 November 2024. doi.org/10.48550/arXiv.2411.14986.

[76] de los Riscos, Agustin Manuel and Luis Fernando D'Haro. "ToxicBot: A Conversational Agent to Fight Online Hate Speech." Lecture Notes in Electrical Engineering, 2021. doi.org/10.1007/978-981-15-8395-7_2.

Google's Jigsaw and CONAN, the multilingual dataset of responses for hate speech noted above.[77] The tool consists of a bot that sends automatic responses to hateful comments online as well as a system interface where users can chat with the bot about the response. On the interface, the chat rates the level of toxicity throughout the conversation as a further indicator of hate speech. This would qualify as a form of one-to-one counterspeech that engages hate speech perpetrators publicly on social media platforms as well as privately through a separate chat interface.

Respond2Racism is another counterspeech tool that was created during the COVID-19 pandemic in response to rising anti-Asian hate speech.[78] Operating on Twitter, this bot responded to tweets using anti-Asian hashtags with a video uplifting Asian frontline workers and educating perpetrators on the effects of misinformation and hate. It also responded to hashtags in support of the Asian community with tips on how to combat hatred in-person. This could be considered a form of one-to-many, non-contextualized counterspeech, with one message sent out to many using the same hashtag, informing both online and offline behaviour.

It is clear that counterspeech tools provide the most promising application of AI in response to the proliferation of hate speech online. Yet, these models are still young and limited in their capacity to adequately capture nuance in their responses.

More recently developed, Normsy.ai is a nonprofit initiative by the Civic Health Project (CHP) that uses AI-augmented counterspeech to reduce online polarization and model how to communicate in ways that uphold social norms. Citing research that social media byscrollers' normative views change when they absorb toxic content that challenges civic norms, CHP designed Normsy.ai as an AI-copilot that assists users in responding to those toxic posts.[79] While a user scrolls on X, the tool constantly crawls the platform in the background, selects toxic posts based on its scoring engine, displays those posts to the user to opt-in to respond to, and drafts sample responses based on research-backed social science interventions. These responses are 'smart, respectful, evidence-based replies' to toxic speech that use language that models mutual respect, reinforces institutional trust, and/or highlights shared democratic values. Users can choose from these responses and customize what to say. Then, to ensure the platform continues to respond effectively to toxic speech, it tracks how those interventions perform over time. Normsy's focus on keeping the human-in-the-loop is valuable, as not only does the AI-generated responses help make social media less toxic in real time, it helps educate and train the users themselves on how to identify and model this behaviour on their own in the future. It also reduces risks of letting a fully automated bot interact with users on the platform, a challenge seen readily in the Grok bot example noted above.

Ultimately, it is clear that counterspeech tools provide the most promising application of AI in response to the proliferation of hate speech online. Yet, these models are still young and limited in their capacity to adequately capture nuance in their responses. How can one properly train an AI model to understand the ever-changing landscape of human conversation? How can one teach the bot to determine the correct audience toward which to direct counterspeech? These questions will continue to be answered as research from tools like Normsy.ai comes to the fore over the next few years, and ultimately will determine the potential for AI to effectively respond to toxic speech online.

[77] Chung, Yi-Lung, et al. "CONAN—COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* July 2019. aclanthology.org/P19-1271/.
[78] Lim, Dion. "'Respond2Racism': San Francisco group creates bot to fight online racism toward Asians." 2020. abc7news.com/asian-racism-coronavirus-americans-asisan-discrimination-respond2racism/6238020/.
[79] Civic Health Project. "Strengthening Online Civic Norms at Scale." Civic Health Project. https://docsend.com/view/uep7d2gtnr9bcqmn.

Conclusion

AI, and LLMs in particular, holds potential in classifying and responding to toxic speech online, but there is much room for improvement. Advancing this goal requires intentional effort from governments, civil society, and, ultimately, the tech developers and platforms themselves.

As for the role of government, it has already been noted above that these institutions play a critical role in ensuring the ethical use of AI, both through establishing regulatory frameworks and through creating an environment which encourages responsible tech use and development. A recommendation often echoed across the responsible tech field is for governments to increase and enforce transparency standards which require companies to disclose the role AI and LLMs play in implementing their content moderation strategies. Critical questions remain, including the transparency of model training, what content it typically removes, and why the algorithm makes the decisions it does. These questions, if answered, can enable citizens to better hold platforms accountable. Also important is the role governments can play in creating responsible tech environments: supporting interdisciplinary research on the intersection of AI, hate speech, and public discourse to inform evidence-based policymaking, advancing digital literacy programs to assist civilians in navigating the online environment safely, and facilitating multi-stakeholder dialogues between tech companies, civil society, and academia to collaboratively address the societal impacts of AI-facilitated moderation. The research outlined throughout this paper provides yet another level of support to each of these requests.

Tech developers and platforms play perhaps the most critical role in the future of AI-facilitated toxic speech response. As it stands, the majority of social media platforms operate under a surveillance capitalist profit model that encourages the algorithmic proliferation of hateful and toxic speech, which then creates the challenge for content moderation teams to respond to this crisis. While ultimately this profit model and these algorithms will need to change in order for prosocial dialogue to be feasible online, developers have a unique opportunity to leverage AI to respond to this challenge in the meantime—and many have begun this critical work, as detailed throughout this study. Based on their leadership and research, we provide some more novel recommendations for developers and platforms to consider.

1. Integrate Peacebuilding Principles into AI Design

Understanding and responding to toxic speech is a difficult challenge, one which experts in conflict resolution, mediation, and counterspeech have been addressing for a long time. Developers should partner with and hire these leaders, incorporating frameworks from peace studies and conflict resolution into the development of AI tools to ensure that counterspeech promotes empathy, understanding, and constructive dialogue. For example, at the University of Notre Dame's Peacetech and Polarization Lab and Center for Research Computing, conflict resolution and computer science students are working together to develop new AI training methods.

2. Prioritize Explainability as a Metric for Classification Models

Following the lead of ToXCL and MixGEN to identify and classify toxic speech, developers should ensure their models provide explanations for why they make the classifications of hate speech they produce. This increases transparency for the users, as well as enables more intentional training and revision of existing models.

3. Implement Simple Solutions First

Research has shown that simple prompting efforts when a user is crafting a post or comment ("*Are you sure you want to post this?*") or automating a non-contextualized response to toxic speech ("*This is hate speech! Such posts can damage your personal and professional reputation*") has a strong effect on reducing toxicity in online discourse. While more robust, contextually-aware models are being developed, platforms should implement these simple solutions in the meantime to begin building safer public spaces.

4. Diversify Training Data and Teams

The most consistent critique of existing AI classification and response generation models is biased data and output. Developers should use and/or build diverse datasets that represent various cultures, languages, and contexts to train models. This requires diverse development teams to mitigate biases and ensure cultural competence, awareness, and sensitivity in LLM-generated responses. Again, at the University of Notre Dame, researchers are building new datasets and methods for AI response generation models for responding to toxic content.

5. Collaborate with Civil Society Organizations

Partnerships like the one between Jigsaw and OpenWeb expand the practical application of AI tools to mitigate toxic speech outside of the research environment. Developers should engage with NGOs, community groups, and experts in human rights and digital ethics to co-develop AI tools that are socially responsible and aligned with community needs. The University of Notre Dame's Peacetech and Polarization Lab is partnering with Dangerous Speech and the Civic Health Project to map out a series of research projects that will help to refine and improve AI for specific communities of practitioners engaged in counterspeech. It is our hope that peacebuilding-informed approaches to fine-tuning LLMs can contribute to bigger impacts in supporting public discourse with AI tools.

WORKS CITED

Almohaimeed, Saad, Saleh Almohaimeed, and Ladislau Bölöni. “Transfer Learning and Lexicon-Based Approaches for Implicit Hate Speech Detection: A Comparative Study of Human and GPT-4 Annotation.” 2024 IEEE 18th International Conference on Semantic Computing, 7 May 2024. ieeexplore.ieee.org/abstract/document/10475615.

Baheti, Ashutosh, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. “Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts.” In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP*), 4846–4862. Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.397/.

Bär, Dominik, Abdurahman Marrouf, and Stefan Feurriegel. “Generative AI may backfire for counterspeech.” *ArXiv,* 25 November 2024. doi.org/10.48550/arXiv.2411.14986.

Bauer, Nikolaj, Moritz Preisig, and Martin Volk. “Offensiveness, Hate, Emotion and GPT: Benchmarking GPT3.5 and GPT4 as Classifiers on Twitter-specific Datasets.” *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, May 2024. aclanthology.org/2024.trac-1.14/.

Benesch, Susan, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. “Counterspeech on Twitter: A Field Study.” *Public Safety Canada,* 2017. doi.org/10.15868/SOCIALSECTOR.34066.

Butler, Anita and Alberto Parrella. “Tweeting with consideration.” X Blog, 5 May 2021. blog.x.com/en_us/topics/product/2021/tweeting-with-consideration.

Caselli, Tommaso, Valerio Basile, Jelena Mitrović, and Michael Granitzer. “HateBERT: Retraining BERT for Abusive Language Detection in English.” *ArXiv,* 4 February 2021. doi.org/10.48550/arXiv.2010.12472.

Chiu, Ke-Li, Annie Collins, and Rohan Alexander. “Detecting Hate Speech with GPT-3.” *ArXiv*, 24 March 2022. doi.org/10.48550/arXiv.2103.12407/.

Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. “CONAN—COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech.” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* July 2019. aclanthology.org/P19-1271/.

Civic Health Project. “Strengthening Online Civic Norms at Scale.” Civic Health Project. https://docsend.com/view/uep7d2gtnr9bcqmn.

CLR:SKY. https://www.clrsky.ai/about.

“Community Standards Enforcement Report.” Meta, 2025. transparency.meta.com/reports/community-standards-enforcement/.

Connors, Grace and Emma Baumhofer. “Peacebuilding and Disinformation: Taking Stock and Planning Ahead.” Berghof Foundation and Platform Peaceful Conflict Transformation, January 2025. https://berghof-foundation.org/library/peacebuilding-and-disinformation.

Costanza-Chock, Sasha. *Design Justice: Community-Led Practices to Build the Worlds We Need.* MIT Press, 2020. www.google.com/books/edition/Design_Justice/m4LPDwAAQBAJ?hl=en&gbpv=1&pg=PR3&printsec=frontcover.

“Crowdsourcing contextual information (Community Notes).” Prosocial Design Network. www.prosocialdesign.org/library/crowdsourcing-contextual-information-community-notes.

Davani, Aida Mostafazadeh, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. “Hate Speech Classifiers Learn Human-Like Social Stereotypes.” *ArXiv,* 28 October 2021. doi.org/10.48550/arXiv.2110.14839.

de los Riscos, Agustin Manuel and Luis Fernando D’Haro. “ToxicBot: A Conversational Agent to Fight Online Hate Speech.” Lecture Notes in Electrical Engineering, 2021. doi.org/10.1007/978-981-15-8395-7_2.

Derczynski, Leon, Bertie Vidgen, Hannah Rose Kirk, Pica Johansson, Yi-Ling Chung, Mads Guldborg Kjeldgaard Kongsbak, Laila Sprejer, and Philine Zeinert. “Hate Speech Dataset Catalogue.” hatespeechdata.com/#English-header.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *ArXiv,* 24 May 2019. doi.org/10.48550/arXiv.1810.04805

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. “ChatGPT outperforms crowd workers for text-annotation tasks.” *PNAS*, 2023. doi.org/10.1073/pnas.2305016120.

Gumelar, Agustinus Bimo, Eko Mulyanto Yuniarno, Arif Nugroho, Dery Pramono Adi, Indar Sugiarto, and Mauridhi Hery Purnomo. “An Improved Toxic Speech Detection on Multimodal Scam Confrontation Data Using LSTM-Based Deep Learning.” *International Journal of Intelligent Engineering & Systems,* 30 September 2024. inass.org/wp-content/uploads/2024/07/2024123167-2.pdf.

Halim, Sadaf MD, Saquib Irtiza, Yibo Hu, Latifur Khan, and Bhavani Thuraisingham. “WokeGPT: Improving Counterspeech Generation Against Online Hate Speech by Intelligently Augmenting Datasets Using a Novel Metric.” *2023 International Joint Conference on Neural Networks,* 2023. ieeexplore.ieee.org/document/10191114.

Hanu, Laura, James Thewlis, and Sasha Haco. “How AI Is Learning to Identify Toxic Online Content.” *Scientific American*, 8 February 2021. www.scientificamerican.com/article/can-ai-identify-toxic-online-content/.

Hartvigsen, Thomas, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. “ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection.” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics,* May 2022. aclanthology.org/2022.acl-long.234/.

HiveModeration. hivemoderation.com/.

Hoang, Nhat, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Luu Anh Tuan. “ToXCL: A Unified Framework for Toxic Speech Detection and Explanation.” *ArXiv*, 25 March 2024. https://doi.org/10.48550/arXiv.2403.16685.

Hutchinson, Andrew. “Twitter is Adding a New Prompt on Retweets When Users Haven’t Opened the Link.” *Social Media Today*, 10 June 2020. www.socialmediatoday.com/news/twitters-adding-a-new-prompt-on-retweets-when-users-havent-opened-the-lin/579595/.

Instagram. “Kicking Off National Bullying Prevention Month With New Anti-Bullying Features.” 2020. about.instagram.com/blog/announcements/national-bullying-prevention-month.

Jigsaw. https://perspectiveapi.com/.

Jigsaw. “10 New Languages for Perspective API.” Medium, 9 December 2021. medium.com/jigsaw/10-new-languages-for-perspective-api-8cb0ad599d7c.

Kamal, Ashraf, Tarique Anwar, Vineet Kumar Sejwal, and Mohd Fazil. “BiCapsHate: Attention to the Linguistic Context of Hate via Bidirectional Capsules and Hatebase.” *IEEE Transactions on Computational Social Systems*, April 2024. ieeexplore.ieee.org/document/10022007.

Kaplan, Joel. “More Speech and Fewer Mistakes.” Meta, 2025. about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/.

Lim, Dion. “'Respond2Racism': San Francisco group creates bot to fight online racism toward Asians.” 9 June 2020. abc7news.com/asian-racism-coronavirus-americans-asisan-discrimination-respond2racism/6238020/.

Mathew, Binny, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. “HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection.” *Proceedings of the AAAI Conference on Artificial Intelligence,* 18 May 2021. doi.org/10.1609/aaai.v35i17.17745.

Meyer, Erin. “Navigating the cultural minefield.” *Harvard Business Review*, 2014. hbr.org/2014/05/navigating-the-cultural-minefield.

Mosleh, Mohsen, Rocky Cole, and David G. Rand. Misinformation and harmful language are interconnected, rather than distinct, challenges. *PNAS Nexus.* March 2024. https://doi.org/10.1093/pnasnexus/pgae111.

Mosseri, Adam. “Our Commitment to Lead the Fight Against Online Bullying.” Instagram, 8 July 2019. about.instagram.com/blog/announcements/instagrams-commitment-to-lead-fight-against-online-bullying.

Mozafari, Marzieh, Reza Farahbakhsh, and Noël Crespi. “Hate speech detection and racial bias mitigation in social media based on BERT model.” Complex Networks 2019 Conference Special Collection: The 8th Conference on Complex Networks & their Applications, 27 August 2020. doi.org/10.1371/journal.pone.0237861.

Nakayama, Hiroshi. 2017. HateSonar: Hate Speech Detection Library for Python. GitHub. https://github.com/Hironsan/HateSonar.

Parker, Sara and Derek Ruths. “Is hate speech detection the solution the world wants?” *PNAS*, 27 February 2023. doi.org/10.1073/pnas.2209384120.

Penemue. www.penemue.ai/.

“Rated Not Helpful: How X’s Community Notes System Falls Short on Misleading Election Claims.” Center for Countering Digital Hate, October 2024. counterhate.com/wp-content/uploads/2024/10/CCDH.CommunityNotes.FINAL-30.10.pdf.

Ribeiro, Manoel Horta, Robert West, Ryan Lewis, and Sanjay Kairam. “Post Guidance for Online Communities.” *arXiv* (November 27, 2024). https://doi.org/10.48550/arXiv.2411.16814.

Reviglio, Urbano and Claudio Agosti. “Thinking Outside the Black-Box: The Case for “Algorithmic Sovereignty” in Social Media.” *Social Media + Society,* 2020. journals.sagepub.com/doi/pdf/10.1177/2056305120915613.

Saha, Punyajoy, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. “CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech.” *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence,* July 2022. doi.org/10.48550/arXiv.2205.04304.

Saha, Punyajoy, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. “On Zero-Shot Counterspeech Generation by LLMs.” *ArXiv*, 22 March 2024. doi.org/10.48550/arXiv.2403.14938.

Schirch, Lisa. *Defending Democracy with Deliberative Technologies*. Keough School Policy Brief Series. Notre Dame, IN: Keough School of Global Affairs, 2024. https://doi.org/10.7274/25338103

Schirch, Lisa. “Digital Peacebuilding Communication Skills: Beyond Counterspeech.” 2020.

Schirch, Lisa. *Social Media Impacts on Conflict and Democracy: The Techtonic Shift.* 2021.

Simon, Guy. “OpenWeb tests the impact of “nudges” in online discussions.” OpenWeb, 2020. www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api.

Song, Xiaoying, Sujana Mamidisetty, Eduardo Blanco, and Lingzi Hong. “Assessing the Human Likeness of AI-Generated Counterspeech.” *ArXiv*, 15 December 2014. doi.org/10.48550/arXiv.2410.11007.

Sprinklr Team. “How Sprinklr Helps Identify and Measure Toxic Content with AI.” *Sprinklr,* 21 March 2023. www.sprinklr.com/blog/identify-toxic-content-with-leading-analytical-ai/.

Sridhar, Rohit and Diyi Yang. “Explaining Toxic Text via Knowledge Enhanced Text Generation.” *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics,* 10 July 2022. aclanthology.org/2022.naacl-main.59.pdf.

Stokel-Walker, Chris. “TikTok Wants Longer Videos—Whether You Like It or Not.” *Wired,* 21 February 2022. www.wired.com/story/tiktok-wants-longer-videos-like-not/.

TrollWall. trollwall.ai/how-it-works.

Turing, Alan. “Computing Machinery and Intelligence.” *Mind,* October 1950.

Ul Mustafa, Raza, Noman Ashraf, and Nathalie Japkowicz. “Can GPT-4 detect subcategories of hatred?” *2024 IEEE Digital Platforms and Societal Harms,* 15 October 2024. ieeexplore.ieee.org/abstract/document/10775211.

Unitary. “Detoxify endpoints.” docs.unitary.ai/api-references/detoxify.

Vidgen, Bertie, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. “Introducing CAD: the Contextual Abuse Dataset.” *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics,* 6 June 2021. aclanthology.org/2021.naacl-main.182.pdf.

Villate-Castillo, Guillermo, Javier Del Ser, and Borja Sanz Urquijo. “A Systematic Review of Toxicity in Large Language Models: Definitions, Datasets, Detectors, Detoxification Methods and Challenges.” ResearchSquare, 15 July 2024. doi.org/10.21203/rs.3.rs-4621646/v1.

Weng, Lilian, Vik Goel, and Andrea Vallone. “Using GPT-4 for content moderation.” OpenAI, 15 August 2023. openai.com/index/using-gpt-4-for-content-moderation/#LilianWeng.

Woods, Freya and Janet Ruscher. “Viral sticks, virtual stones: addressing anonymous hate speech online.” *Patterns of Prejudice,* 2021. https://doi.org/10.1080/0031322X.2021.1968586.

Yo, Ching-Yun, Pin-Yu Chen, Payel Das, Youssef Mroueh, Soham Dan, Georgios Kollias, Subhajit Chaudhury, Tejaswini Pedapati, and Luca Daniel. “Large Language Models Can be Strong Self-Detoxifiers.” *ArXiv,* 4 October 2024. doi.org/10.48550/arXiv.2410.03818.

Yousefi, Midia and Dimitra Emmanouilidou. “Audio-based Toxic Language Classification using Self-attentive Convolutional Neural Network.” *29th European Signal Processing Conference, 2021.* ieeexplore.ieee.org/document/9616001.

Zhou, Xuhui, Maarten Sap, Swabha Swayamdipta, Noah Smith, and Yejin Choi. “Challenges in Automated Debiasing for Toxic Language Detection.” *ArXiv,* 29 January 2021. doi.org/10.48550/arXiv.2102.00086.

Zuboff, Shoshana. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs, 2019.

THE TODA PEACE INSTITUTE

The Toda Peace Institute is an independent, nonpartisan institute committed to advancing a more just and peaceful world through policy-oriented peace research and practice. The Institute commissions evidence-based research, convenes multi-track and multi-disciplinary problem-solving workshops and seminars, and promotes dialogue across ethnic, cultural, religious and political divides. It catalyses practical, policy-oriented conversations between theoretical experts, practitioners, policymakers and civil society leaders in order to discern innovative and creative solutions to the major problems confronting the world in the twenty-first century (see www.toda.org for more information).

CONTACT US

Toda Peace Institute
Samon Eleven Bldg. 5 th Floor
3-1 Samon-cho, Shinjuku-ku, Tokyo 160-0017, Japan

Email
contact@toda.org

Sign up for the Toda Peace Institute mailing list
https://toda.org/policy-briefs-and-resources/email-newsletter.html

Connect with us on the following media.
YouTube: @todapeaceinstitute3917
X (Twitter): https://twitter.com/TodaInstitute
Facebook: https://www.facebook.com/TodaInstitute