AI & SOCIETY
https://doi.org/10.1007/s00146-025-02463-y

MAIN PAPER

NLP as language ideology: discursive and algorithmic constructions of 'toxic' language in machine learning research

Gabriella Chronis[1]

Received: 11 July 2024 / Accepted: 23 June 2025
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2025

Abstract
This article considers natural language processing research as a language-ideological practice, looking specifically at the task of toxic language detection, which impacts nearly everybody online through automated content moderation. Industry discourse constructs the category of *toxicity* through a series of oppositions between civil/healthy/referential/rational and unhealthy/toxic/indexical/emotional. Examples from a toxicity correction dataset demonstrate how this ideology can become encoded algorithmically: a focus on preserving referential content in text "detoxification" results in neglect of important poetic, expressive, and social-indexical functions. Overall, the discursive framings of toxicity construct the ideal speaker in terms of what I call a "referentialist" language ideology, which values rational debate in the (regulated) liberal-democratic public sphere. Ultimately, toxicity detection and other metapragmatic tasks do not merely model the existing pragmatic categories but actively construct them. Toxicity in particular potentially reinforces exclusionary norms of white maleness and promotes online subjectivities that are useful (profitable) to the commercial platforms that shaped the task. Since there is no avoiding NLP as language-ideological practice, independent NLP researchers must acknowledge the political potency of their work by continually reflecting on the categories they work with in relation to models of social-political formation.

Keywords Metapragmatics · Toxic language detection · Language ideology · Automated content moderation · Civility

1 Introduction

Much of the user- and AI-generated content published on the internet is mediated by automated filters. It is easy to find obvious cases of this technology gone awry. For example, a gamer replies to a tweet about an online streamer caught in a sex scandal, contextualizing the gaming community's apologism against its transphobia: "None of those people actually care about kids, they just want only cis straight people to exist." "Your tweet may go against our community standards. Do you want to revise it before sending?"[1] A baseball fan taps out a message to her new Tinder match: "Well anyone kicking the Dodgers ass makes me happy." The first response comes not from her match but from the platform: "Slow down—your match may find this language disrespectful. Are you sure you want to send?"[2] An activist working for an NGO notices that their social media posts mentioning Palestine are viewed far less often than their other posts (Abokhodair et al. 2024). These suggestions are made by automated systems for content moderation.

Many such content moderation systems are machine learning models trained to solve a natural language processing (NLP) task called *toxic language detection*. The aim of developing such models is to reduce harassment online by flagging potentially harmful posts for review or removal. Though their developers advise against using them to remove posts or ban users without review by human moderators, they are commonly deployed without human supervision in 'nudge' mechanisms (Simon 2020; OpenWeb 2020)

✉ Gabriella Chronis
gabriellachronis@utexas.edu

[1] Department of Linguistics, The University of Texas at Austin, 110 Inner Campus Dr, Austin, TX 78712, USA

[1] https://x.com/EddyZacianLand/status/1805720916930052611.

[2] https://twitter.com/Rallyears/status/1407215540687441920.

Published online: 16 July 2025

like the Twitter and Tinder pop-ups that offer users a chance to rewrite their message *before* posting it.

The identities of the authors in these examples—queer, Middle Eastern, woman—hint at one of the main claims of this article: that the category of toxicity as it is constructed in NLP research devalues the very voices such research is supposed to protect. But toxicity is more than just exclusionary; it actively constructs the speaking subject. It is a Foucauldian "technology of the self" (1978). Rather than exclude, the nudge *conditions* inclusion on a particular kind of linguistic behavior. Nudge mechanisms cause a significant proportion of users to edit their post before publishing, whether to remove hate speech or seek creative ways around filters (Simon 2020). Jereza (2024) contends that rather than discouraging racism, content moderation encourages people to change *how* it is expressed and circulates a kind of flexible racism that abides by liberal democratic norms of politeness. If racism can slip through the cracks, what is it that racists are changing about their posts? And how might the authors of the above examples revise themselves in response to this nudge?

What can we learn about content moderation by viewing it not as a repressive mechanism of restriction, but as a constructive mechanism of production? In other words, what kind of subjectivities, in Foucault's sense, does toxicity detection produce?[3] NLP research does not merely reflect *toxicity* as an existing social category but actively constructs it according to a referentialist language ideology. This article characterizes the language-ideological construction of toxicity and illustrates its expression in practical technologies.

After a brief genealogy of toxicity detection in the field of natural language processing (NLP) research (Sect. 3), I turn to the *metapragmatics* of toxicity within the field as a lens into language ideology. At a rhetorical level (Sect. 4), toxicity research leverages frames of *civility*, *health*, *inclusivity*, and *engagement* to refract a Habermasian ideal of rationalist discourse through the lens of commercial interest. Civility, used to contrast toxicity, places emphasis on scientific modes of knowing—evidence, logic, and fact. These values evince a referentialist language ideology associated with the bourgeois public sphere. At the same time, the engagement frame promotes values that clash with the public sphere ideology. As an assemblage of sometimes contradictory frames, the discursive construction of toxicity gestures towards a new political subject of the public sphere. An audit of a training dataset (Sect. 5) illustrates how the metapragmatics of toxicity are presented to machine learning models. Through the process of data annotation, toxicity comes to be associated with non-referential and performative aspects of language like identity negotiation, social connection, and poetic expression.

Ultimately, toxic language detection may do more than combat online harassment. It may in fact enforce hegemonic norms associated with educated white male spaces and aid in the production of subjectivities which serve the commercial goals of the regulating bodies. Ethical NLP demands that researchers must acknowledge and interrogate the ideological and political implications of their work, electing categories that promise to expand rather than contract modes of being.

2 Indexicality, language ideology, and metapragmatic practice

Before proceeding to the analysis, it is necessary to define the linguistic-anthropological concepts it uses and to clarify the relationship between them. One of my central claims is that the category of toxicity reflects a referentialist language ideology. To make this argument, I need to outline different ways of thinking about meaning and about how language works.

One might see language primarily as a tool for communicating information about the world. This capacity of language to *refer* is the main concern of formal linguistic semantics and philosophy of language, which have exerted a strong historical influence on AI research. But language does much more than refer. People use language to do things in the world, like make promises and manage relationships. The main concern of linguistic anthropology and semiotics, which tends to view language primarily as a social practice, is tending to these kinds of meaning and function that aren't easily described by semantics.[4]

The theory of Roman Jakobson (1960), influential among semioticians, identifies six functions of language. Each function maps to a different element of the speech event. The referential function relates the utterance to the physical context (a.k.a. 'the world'). Referential meaning concerns

[3] Toxicity detection systems are part of an *apparatus* in the sense described by Foucault and later elaborated by Giorgio Agamben. The apparatus or *dispositif* is, "literally anything that has in some way the capacity to capture, orient, determine, intercept, model, control, or secure the gestures, behaviors, opinions, or discourses of living beings". (Agamben 2009). Foucault (1979) argued that institutions like prisons, hospitals, factories, schools and so forth do not merely govern subjects; they use apparatuses like surveillance and the threat of discipline to *create* subjects. They form and reform subjectivities by capturing and reorienting the desires of the subject. Part of this process involves organizing 'regimes of truth' which construct categories of people and behavior (such as 'criminal', 'hysteric', or 'teenage mother') and influence how power is exercised over them to create ideal political subjects (Rampton and Richardson 2007).

[4] See James Slotta (forthcoming) on the embattled relationship between linguistic semantics and linguistic anthropology its consequences for the theoretical functioning of indexicality and referentiality.

the truth-conditional meaning of a speech event—the propositional content that is the traditional purview of formal semantics. Jakobson identified five other functions, each of which relates the utterance to another aspect of the speech event: the expressive function relates the utterance back to the speaker. Other functions are conative (relating to the addressee), metalinguistic (relating to the code/language), phatic (relating to the channel/medium), and poetic (relating to the form of the message itself).

One non-referential function of language is the construction of social identity and group formations. Speakers *index* or point to social identities, personalities, and ideological stances through non-referential stylistic features such as register and dialect. Without explicitly *saying* that I'm smart (a strategy that could even backfire), I can *index* or point to my own intelligence by using a big word (Eckert 2019). Indexical associations, even conventional ones like the idea that big words make you smart, posit a physical/causal relation between sign (big word) and object (smarts). Indexes are different from the Saussurian model of the sign dominant in linguistics, which posits an arbitrary relationship between sign (word-shape) and object (conceptual meaning).[5]

A related framework for theorizing meaning beyond reference is speech act theory (Austin 1962). This framework holds as a basic tenet that all speech is also action. Assertives (statements that describe the state of the world) are merely one class of speech acts among others, like apologizing or pronouncing a couple to be married.

Studying indexical meaning led Michael Silverstein to conclude that the lion's share of linguistic meaning is indexical. Studying speech acts led J.L. Austin to conclude that most language is performative. Complexes of related ideas and beliefs of how language works, how it *should work*, are called language ideologies. Language ideology refers to the ideas and conceptions speakers have about language, and the study of how those ideas both influence and are influenced by social dynamics and power relations (Irvine and Gal 2000; Woolard 2020). As people navigate the social world, they draw on and develop ideas and practices around language use: how language can and should be used, and to what ends. These ideas can be expressed more or less consciously and overtly, or through implicit and presupposed cultural assumptions.

Language ideology is not *just* about language; it relates ideals about language use to ideals about other structures within which we are situated: personhood, society, aesthetics, morality. At the broadest level, language mediates social structures (and administers power) by 'controlling the conversation,' or determining the limits of the sayable (Foucault 1978). This work is accomplished by the production of specialized knowledge and techniques which govern, guide, and shape linguistic conduct and subjectivities (Urla 2019). By discussing the ideal speaker in terms of particular frames, and by quantifying this image in datasets, research practices around toxicity detection impact not only the subjectivities of end users whose posts are guided by algorithms, but also the direction and efforts of the research community itself.

This paper explores how NLP research constructs and reifies a "referentialist" language ideology, and probes some of its sociopolitical implications. Referentialism values the symbolic and referential functions of language over its indexical and performative functions. Furthermore, it holds that some kinds of speech are best suited for referential work: logical argument, factual reporting, scientific discourse. Naturally, these registers still index social qualities like a high level of formal education and socio-economic status, but referentialism minimizes indexicality by construing these dominant qualities as socially neutral. In an article which serves as a template for my own analysis, Bauman and Briggs (2000) develop the connected notion of a scientific language ideology by analyzing the texts of John Locke. They characterize it as a complex of beliefs that valorizes "knowledge, truth, universality, rationality and science," over "passion, beauty, belief, passivity, particularity, error, deceit, and rhetoric" (2000, 159). In addition to minimal use of idiomatic, poetic, and rhetorical devices, ideal language on this view stands on its own, limiting intertextual ties to other texts.

The work of anthropologists like Silverstein (1976, 2003), Bauman and Briggs (1990), and Lucy (1992) emphasizes the language-ideological nature of *all* linguistic practice. Silverstein viewed language ideology in terms of people's beliefs about language that they use to justify or rationalize particular linguistic practices over others (Silverstein 1976, 193). One example would be the belief that smart people use standard English. Another is the idea that language can be hurtful. To describe how speakers discuss, transmit, and negotiate social views about language on a micro-interactional level, he developed the concept of *metapragmatic speech*. Metapragmatic speech is language about language *use* (Silverstein 1976). The classic example is a parent telling a child to "say please," which carries within it the instruction about how to use language a set of values about politeness—when and to whom it is important to be polite and show deference and so forth. Roughly, ideology and metapragmatics can be thought of as dialectically connected facets of the same processes operating at micro- and macro-social levels: macrosocial language ideologies constitute the grounds for micro-interactional metapragmatic practice, which in turn influences ideology by shaping "processes of

[5] All kinds of signs perform referential functions, and all speech functions, including referentiality, are socially meaningful. However, non-referential indexicality has been a major focus for theorists interested in the social functions of language. Because I am interested primarily in the middle of the Venn diagram of these concepts, in this paper I use them interchangeably.

producing and receiving texts, affecting who is authorized to speak …and in what sorts of institutional spaces" (Bauman and Briggs 2000, 142).

Below, I analyze two kinds of metapragmatic speech about toxicity: industry publications (where researchers talk to each other and other industry professionals), and a datasets used to train toxicity detection models (in which researchers and annotators address the computational agent). These metapragmatic discourses plays a role in the construction of the category of toxic language, which in turn is used to license particular metapragmatic practices of machine moderation. In other words, NLP in action, because it is metapragmatic practice, is language ideology in action.

3 'Toxicity detection' as an NLP task

Natural language processing (NLP) aims to make accurate computational predictions about unstructured, free form text. NLP *tasks* are abstract goals for research that organize different kinds of predictions: sentiment, information about the author, continuations of the text, translations into another language, summarization, etc. Tasks are social categories, shaped by (but not synonymous with) the shared datasets that the research community uses to train and evaluate models. Like all language classification tasks, work on toxicity detection begins with constructing a dataset. Annotators (typically crowd workers) label each comment 'toxic' or 'nontoxic.' These data points are then used to tune a machine learning model to predict the annotated label.

Before 2018, there was no natural language processing task called "toxicity detection." In 2017, around five published NLP articles mentioned the word toxic.[6] Up until that time, related research (of which there was plenty) was formulated in terms of other taxonomies and labels, including hate speech (Kwok and Wang 2013; Burnap and Williams 2015; Djuric et al. 2015; Davidson et al. 2017), insults (Mahmud et al. 2008; Sood et al. 2012a, b; Sax 2016; Sharma et al. 2018), profanity (Sood et al. 2012a, b; Malmasi and Zampieri 2018), bullying (Xu et al. 2012; Dadvar and Jong 2012; Dadvar et al. 2014), and personal attacks (Wulczyn et al. 2017).

The "ratification" of the toxicity detection task (see Latour 1987) was effected in large part by Google Jigsaw, Google's "geopolitical think/do tank" (Carr 2017). Jigsaw accomplished this by funding the construction of large crowdsourced toxicity datasets and sponsoring public programming competitions with cash prizes for the models that best fit those datasets (Wulczyn et al. 2017; Dixon et al. 2018; Borkan et al. 2019; Jigsaw 2017, 2019, 2020). Google Jigsaw released the "Toxic Comment Classification Challenge" dataset on Kaggle, a data science contest platform, in December, 2017. This dataset, based on news comments, defined toxicity to annotators as "anything that is *rude, disrespectful, or unreasonable that would make someone want to leave a conversation*" (Dixon et al. 2018), emphasis in original; see also (Borkan et al. 2019; Vasserman 2019; Pavlopoulos et al. 2019, 2020, 2022). Thousands of research teams submitted systems for predicting the labels in this dataset. By 2018, over 100 more articles on toxicity detection had been published, all citing Jigsaw's dataset. As of 2024, the count of automated toxicity mitigation articles is closer to the thousands. This research largely adopts Jigsaw's definition, its datasets (which effectively define toxicity to the models) and often rely on its product, Perspective API, to operationalize toxicity in experiments (e.g., Gehman et al. 2020; Lee et al. 2024).

Six years later, 'toxicity detection' has fully ascended to the pantheon of standard NLP tasks. Industry giants invariably list *toxicity* among the problems that 'AI safety' efforts aim to address. OpenAI, Meta and Anthropic use Perspective to develop and evaluate their proprietary large language models (LLMs) (Ouyang et al. 2022; Srivastava et al. 2023). Google uses Perspective to filter LLM training data and to mediate exchanges between LLMs and end-users (Welbl et al. 2021; Vasserman 2023). In short, toxicity has become a *social fact* in the NLP community.[7]

Toxicity research has faced plenty of criticism within the field. Analysts have shown that toxicity filters like Perspective API reflect the racial, gender, and religious biases they are supposed to protect people from; posts which merely mention identity labels are more likely to be flagged as toxic (Dixon et al. 2018; Welbl et al. 2021). Sap et al. (2022) demonstrate that Perspective API has a dialectal bias—comments written in African American English are more likely to be rated toxic. Some contend that toxicity is contextually defined, and call for a community-specific approach (Diaz et al. 2022; Saleem et al. 2022). Others propose techniques for 'mitigating bias' in the data or in the trained model (Zhao et al. 2018; Feng et al. 2021). These interventions, and the burgeoning field of toxicity research itself, reflect the strong current of social consciousness among AI researchers, and

[6] All article counts were determined according to Google Scholar search results for "toxic language detection" and "toxic language NLP.".

[7] Of course, 'toxicity' has risen to prominence in broader popular and internet culture as well, as a metaphorical descriptor of persistently negative people, situations, and behaviors. There is not enough room in this article to trace the recent semantic broadening of the word *toxic* and trace its genealogy in machine learning research. A more detailed picture, complete with accusations of relationship and workplace toxicity against important figures in the story (Franceschi-Bicchierai 2019; Kara Corvus 2020; Izento 2021), is undertaken in Chronis (in progress). See also (Rieder and Skop 2021) for an excellent organizational analysis of Jigsaw.

the awareness that their work is politically charged.[8] However, bias mitigation techniques are demonstrably ineffective, tending to mask bias rather than 'remove' it (Gonen and Goldberg 2019).

As metapragmatic agents engaged in the work of language ideology, NLP practitioners must take a reflexive perspective in their research that accounts for relations between power, language and technology (Benjamin 2019; Barabas et al. 2020; Miceli et al. 2020). Bias mitigation aims to correct for ideological bias in the task at hand. That is, it aims for ideological neutrality.

In this paper, I aim to show how toxicity detection, and by extension other metapragmatic tasks, is ideological in *essence*, and to outline the contours of this ideology: the metapragmatics of the task itself reinforce a referentialist language ideology that devalues speech practices associated with marginalized social groups.

4 Explicit ideologies of toxicity in technical and professional discursive practice

What kind of talk is ideal talk, in toxicity research? Drawing on frames of public/mental health and civic participation in the liberal democratic public sphere, technical NLP discourses construct an image of ideal language as a tool for rational deliberation about ideas. They do so by constructing parallel oppositions between the healthy/civil/rational/engaged user on the one hand, and the unhealthy/toxic/irrational/antisocial individual on the other. The health frame additionally naturalizes the idea of behavioral intervention (in the form of the 'nudge'), a characteristic move of neoliberal governmentality. Ambiguity in the definition of toxicity allows for slippage between ideological frames, enabling it to resonate on both ethical and profit-seeking levels.

4.1 Inclusion

Jigsaw frames toxicity as a problem of inclusivity on the internet. A stated goal of the Jigsaw's Conversation AI team is to "raise voices" that are quieted by hateful and aggressive online behavior, especially among marginalized groups (Vasserman 2019). *The Current*, Jigsaw's attractively designed online marketing journal, has a whole issue dedicated to toxicity. It cites studies reporting disproportionately high levels of online harassment and cyber-bullying for women, people of color, and LGBTQ + respondents. The phenomenon of toxicity "reduces diversity of thought," and toxicity detection facilitates the construction of a "vibrant space" for "connectedness on a global scale" (Jigsaw 2021).

Though toxicity encompasses hate speech, it is not considered co-extensive with identity-based attack. Jigsaw also focuses on "subtler forms of toxicity like sarcasm, condescension, or dismissiveness" (Price et al. 2020, 2). At times, toxicity seems to encompass almost any form of negativity. One recent paper from a major conference proposes word-level detoxification, which would translate a sentence like, "An ugly life for an ugly man," into, "An amazing life for an ordinary man" (Lee et al. 2024). Closely related to the inclusion frame is the engagement frame. The difference is one of perspective. Whereas inclusion focuses on the emotional experience of the user, engagement focuses on the measurable aspects of online behavior like time spent online. Inclusion is positive for the user, and engagement is positive for the platform. The difference is reflected in the contexts of use: the customer engagement frame is foregrounded in promotional materials like *The Current* but backgrounded in research papers. Jigsaw's widely adopted, intentionally vague (Vasserman 2019) annotation instructions rely on intuitions about what would make somebody leave a conversation. This definition resonates with both inclusivity and customer engagement frames.

4.2 Civility

Jigsaw publications consistently oppose toxicity with civility. As noted, Jigsaw promoted its toxicity datasets by hosting coding contests with cash prizes, which led to thousands of teams developing and testing models against them. The second such dataset was called *Civil Comments* (Borkan et al. 2019).[9] Another paper from the Jigsaw automatically generates "civil rephrases of toxic texts" (Laugier 2022). The two suggested functions of online discourse put forward in this paper are to "exchange views" and "express … opinions" that are "legitimate, well intentioned, and constructive" (1442). The motifs of debate, scientific rigor, and their emphasis on substance and facts recur throughout Jigsaw's

[8] Bias in NLP is extremely well-studied (Bolukbasi et al. 2016; Bender et al. 2021), including the area of toxic language detection. However, most of this research uses a framework of inclusion. The inclusion framework, which focuses the conversation on discussions of fair allocation, can obscure other forms of injustice (Burrell 2024). For example, Sap et al. (2022) discuss the problem of dialectal bias in terms of "spurious correlations" between toxicity and dialectal features. The idea of a spurious correlation means it is an *accidental* artifact of the data that texts written in African American English score a higher baseline toxicity score. This paper argues that we must consider that the definition of toxicity, which hinges on civility and rationality, is inseparable from linguistic features which are conventionally associated with groups who are considered to be less than rational.

[9] Jigsaw purchased this data from an online news comment platform of the same name, founded with the explicit goal of "solving the problem of civility in online discussions" (Bogdanoff 2017). Their mission indicates how from the very beginning the idea of toxicity detection was bound up with a conception of speakers as citizens who relate to one another through norms of good and orderly behavior.

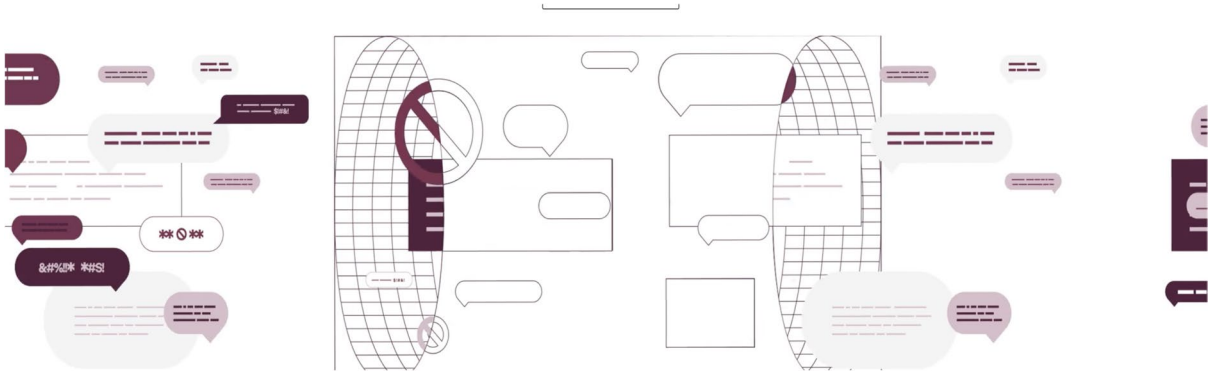

Fig. 1 Screen capture of an animated graphic from the front page of Perspective API's website. Perspective API is a proprietary toxicity detection API service run by Google Jigsaw. It is used by thousands of online platforms to manage user-generated content and has become an industry standard for guiding LLM "safety" research

publications. In a "healthy" conversation, posts are "made in good faith, not overly hostile or destructive, and generally invite engagement. Such a conversation may include *robust engagement and debate*, and is generally (though not always) *focused on substance and ideas*" (Price et al. 2020, emphasis added).

Other Jigsaw-sponsored research develops taxonomies and datasets for detecting *constructive* speech, a category which allows ideal language to be defined in positive terms rather than merely as toxicity's negative. The notion of constructiveness is closely tied to civility: "Constructive comments intend to create a *civil dialog* through remarks that are relevant to the article and *not intended to merely provoke an emotional response*. They are typically targeted to *specific points* and supported by appropriate *evidence*," write the authors (Kolhatkar and Taboada 2017; Kolhatkar et al. 2020, emphasis added). They introduce a taxonomy of constructiveness that ranks opinions with rationale or evidence as more constructive than 'mere' opinions, and toxic but reasoned comments as more constructive than positive but unreasoned comments. Comments like "Thanks Maggie!", a performative speech act which functions phatically to build solidarity and community, are dismissed as non-constructive, "little more than backchannels [that] do not contribute much to the conversation" (2).

The task of finding and ranking constructive comments is gaining traction as a strategy for platform governance. Google recently invested in the annotation of a large constructiveness dataset (Jigsaw 2024). The annotation categories were inspired by characteristics of *New York Times* Top Picks, and the data to be annotated came from the comment section of news articles. Researchers suggest the generalizability of their framework: "While our work focuses on online news comments …we believe that the overall constructiveness labels and the constructiveness sub-characteristics are applicable to many online conversations and commenting platforms." (18). The kinds of speech used to characterize civility and constructiveness—providing evidence, debate, focus on substance and ideas—as well as the researchers' focus on news comments as a primary data source, evince a referentialist language. The standard of polite exchange of views among strangers about current events is established as the ideal form of speech on the internet.

4.3 Public health

In addition to civility, toxicity is contrasted with health, as in the Jigsaw paper "Six Attributes of Unhealthy Conversation" which expands the field of concern "from 'toxic comments' to 'unhealthy conversation'" (Price et al. 2020, 2). The health frame likens toxicity to an unnatural environmental pollutant or a contagious disease. In 2020, Jigsaw partnered with OpenWeb, a platform that manages engagement for scores of major online publishers (Coffee 2022). "There is a crisis of toxicity online," reads the front page of OpenWeb's website, "It's time to save online conversations."[10] This metaphor likens the digital commons to a natural resource that needs to be managed. Indeed, the scrolling animation on Perspective API's landing page shows a stream of comments being filtered through a large cylindrical straw (Fig. 1).[11] The waters of the information stream have been thoughtlessly muddied, and automated content moderation will purify them again by filtering out harmful content that can spread and cause infection in others. The frame of environmental threat blends with that of a threat to public order: the environment that needs saving is civilization itself. An interactive graphic in "The Toxicity Issue" of *The Current* (Fig. 2) shows an 8-bit city skyline labeled 'Comment Town.' As the user scrolls, speech bubbles pop up around the skyscrapers, which deteriorate

[10] https://www.openweb.com

[11] https://perspectiveapi.com

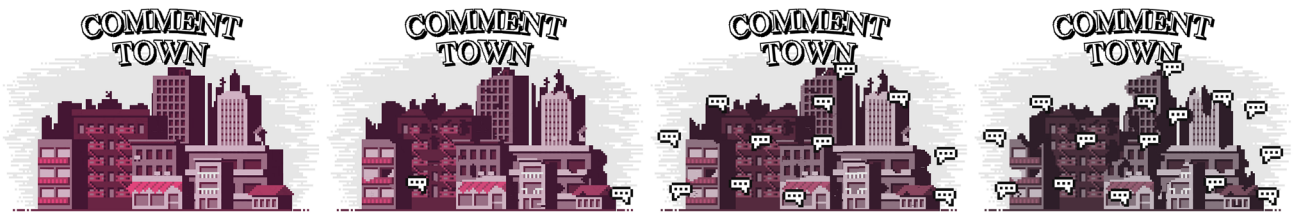

Fig. 2 Interactive graphic from "The Toxicity Issue" of Google Jigsaw's online journal *The Current*. As the reader scrolls, the vibrant skyline of comment town erodes into a toxic ruin

and crumble (G. Jigsaw 2021). The first lines of the journal declaim that, "toxicity online isn't just an issue of civility and healthy discourse. It poses a threat to publishers, platforms, and society."

The health threat that the Perspective promises to neutralize is the toxic individual. According to Jeffey Lin of Riot Games, the behavioral scientist whose toxicity research led directly to Jigsaw's initial toxicity papers, one toxic player is enough to *poison* a whole game (Lin 2013).[12] At the same time, the health frame is used to suggest that toxicity mitigation will save the toxic individual themselves by 'nudging' them towards healthy, constructive, civil language. Jigsaw's constructiveness research speaks of the "positive contagion effect" of "proactive intervention" (Kolhatkar et al. 2020). According to the latest Jigsaw CEO Yasmin Green, internet users want to share accurate (true) information online, but are hindered by the pace of the internet, which stimulates quick and irrational thinking (Green 2022). People *want* to contribute substantively to the discussion of ideas, but they cannot help but say non-constructive things like "Thanks Maggie," at times. Toxicity detection systems are supposed to help people overcome their own biases by automatically capturing their unfiltered responses and redirecting it. It helps them slow their roll, as it were. OpenWeb runs all user comments through Perspective API. If a comment is flagged as toxic, the user is shown a warning encouraging them to rephrase it *before* posting. "We believe encouraging good behavior is equally as important as removing the bad actors to support a thriving community," said CEO and founder Nadav Shoval (OpenWeb 2020).

In framing toxicity as a large-scale public crisis, and toxic individual as both malady and victim, Google positions itself as a public health expert with sufficient resources to tackle the problem, thus justifying the intervention of large platforms in online discourse. Nudge theory, which has roots in behavioral economics (Thaler and Sunstein 2008), exemplifies the model of liberal governance. Jigsaw is engaged almost explicitly in the project of self-formation. It promotes automated toxicity detection and content moderation as Foucauldian "technologies of the self" (1988), transferring the learning task from the algorithm to the user. Like Catholic confession, psychoanalysis and step-counters, toxicity detectors allow users to know themselves and thereby transform themselves into good civil subjects. What do users need to know about themselves to be good civil subjects? Full civic participation requires the ability to anticipate the difference between civil and toxic, healthy and unhealthy, rational and emotional.

4.4 The regulated public sphere

Discourse on toxicity identifies civility with factual, descriptive language, prioritizing "substance" over form, "ideas and opinions" over emotions, and "rational argument" over casual or expressive speech. Referential meaning is valued through contrasts between civil, healthy, referential, and rational on the one hand and toxic, biased, and indexical and emotional on the other. More than just offering a model of ideal communication, this complex of values reflects an ideological model of political organization, one rooted in the Habermasian ideal of the bourgeois public sphere (albeit with significant differences).

Habermas (1989) envisioned the public sphere as a space where free and open rational debate forms public opinion, which lawmakers and politicians then use to inform policy. Thus the ideology of the public sphere is a *metadiscursive* ideology about the role language plays in political organization. The mode of political action in the public sphere also deserves comment. For most of the public, political activity is about speaking rather than acting. The *kind* of speech is also key for Habermas. His "ideal speech scenario" assumes an impartial common ground. Participants arrive at objective truth and intersubjective consensus through reasoned debate, with assertion and factual reference being the primary kinds of speech acts by which people arrive at objective truths (1990). According to Habermas, this kind of discourse naturally leads to optimal solutions. Unfortunately, he opines, it

[12] Google's research on toxic language detection can be traced directly to Lin's efforts to detoxify player behavior in the online game League of Legends. Lin conducted massive-scale psychological priming experiments based on 'nudge theory' mechanisms similar to those promoted by Jigsaw. Again, for an account of Lin's role in the story of toxic language detection, see the genealogy in [redacted] (in progress).

only really existed for a brief time in English coffee houses in the 17th century, after which point commercial interests interrupted its proper functioning.

4.4.1 Exclusion

Jigsaw heralds the internet as a "virtual public square," mirroring Habermas' idealized vision and lamenting its decline in a way reminiscent of his nostalgia for the 17th-century coffee houses. Like Jigsaw, Habermas emphasized the importance of inclusivity for proper functioning of the public sphere. Yet critics of Habermas have rightly pointed out that the public sphere was not inclusive. It excluded women, people of color, criminals, the disabled, workers, immigrants, children—everyone other than educated upper middle class white men (Fraser 1990; Agha 2011). Linguistically, only those with access to educated forms of speech are able to participate in the production of truth.

Habermas himself acknowledged the importance of all kinds of speech acts, but Jigsaw's characterization of 'civil,' 'constructive' speech as rational, dispassionate, and substantive echoes the contemporary imaginary of the public sphere in placing heavy emphasis on referentialist types of speech. This is a classic problem of Standard English hegemony: because civility and rationality are associated with educated white maleness, genres of discourse associated with educated white male spaces are seen as neutral or ideal, while others are marked as deficient. Reformist efforts to mitigate bias in toxicity detection assume that features like civility, rationality, and constructiveness can be separated from identity, when in reality, those qualities are socially constructed and tied to particular groups. Kolhatkar et al. (2020) label "Thanks Maggie" as a nonconstructive comment because it does not contain a reasoned argument. Why place purely phatic communication below reasoned debate in the hierarchy of pragmatic value? The way we define and value civility, rationality, and constructiveness is based on who is seen as possessing those traits. "Thanks Maggie" is constructive of social relationships, a function which not coincidentally is often marked as feminine.

Even if a model could disentangle dialectal features from toxicity, the deeper issue is that toxicity itself is defined in terms of civility and rationality (constructiveness). Historically, rationality and civility have been defined as the negative of Black/Woman/Native and categorically denied to those groups (Wynter 2003). If constructiveness is a necessary prerequisite for participation in the digital public square, and constructiveness is equated to dispassionate, logical debate (a referentialist view), then the concept of constructiveness (and its converse toxicity) may be *constitutively* exclusive.

4.4.2 Useful users

The discursive construction of toxicity hint at a new kind of "ideal subject"—one that differs from the classical public sphere model of political participation. Habermas saw commercial interests as a corrupting influence that ultimately caused the collapse of the bourgeois public sphere. Contemporary critiques of "digital publics" follow a similar line, noting that digital platforms are controlled by a small number of commercial entities with vested interests in shaping online interactions. Yet Jigsaw (a subsidiary of Google) uses public sphere rhetoric to justify its interventions in online discourse. This apparent paradox highlights a key departure from Habermas' vision. The most obvious departure is the use of the public health frame to justify a move away from "absolute" free speech in favor of stricter governance. But Jigsaw's ideal subject differs from Habermas' in their motivation. Jigsaw's vision of the ideal online subject as one who "invites engagement" (Price et al. 2020) resonates with both ethical values of inclusion and corporate values of user engagement. The tech industry defines engagement in terms of clicks, responses, and longer watch times. I have noted that the category of toxicity sometimes extends to encompass all negativity. This is striking because, in the classical public sphere, negativity is essential—for Habermas, criticism and dissent are crucial for arriving at the truth through democratic deliberation. The suppression of negativity suggests a different function for toxicity detection, more in line with the motive of increasing engagement. In addition to combatting harassment, toxicity detection systems may play a material role in technocratic governmentality—its shaping of subjectivities (Rose and Miller 1992; Urla 2019; Rouvroy and Stiegler 2016). Could the blending of frames be working to reconfigure the ideal subject of online discourse—not as a truth-seeker, but as a continuously engaged and engaging user? Without space for deepening the connection to governmentality, my comments can do no more than suggest that toxicity detection is not a neutral technology but a crucial part of a broader transformation in the public sphere.

Others raise the connection between liberal governance and content moderation (Jereza 2024), even linking Jigsaw to the imaginary of the public sphere (Rieder and Skop 2021). In this section, I have built on this foundation by detailing how the idealized model of civic participation that is foundational to the category of toxicity in NLP systematically excludes emotional, expressive, and marginalized forms of speech. The next step is to illustrate that this referentialist ideology can become encoded in the models themselves. How is the ideal speaker represented in a machine-readable format? Sect. 5 illustrates how data annotation practices around toxicity can come to implicitly exclude speech that is emotional, expressive, or socially indexical.

5 Metapragmatics of toxicity in a machine learning dataset

I began this article with the question of how toxicity detection systems might affect the "objects" (actual populations and their language) they model. Yet the above discussion mainly concerns relations between machine learning modelers. Neither internet users nor algorithms attend to researcher discourse nearly as much as other researchers. The nexus between researcher and algorithm is data. Along with other junctures like model architecture and task design, data collection is one of the primary points where the choices and actions of people affect the ultimate behavior of the model (Kockelman 2020).

The annotation of data and its use in training regimes both constitute metapragmatic practice. First, researchers instruct annotators about how to label data for a task, Then, just as a parent teaches a child about politeness, a comment-annotation pair tells a machine learning algorithm to associate an utterance with a moral and ethical value: "This comment *should* be labeled toxic; this one should not." For a given model, its training data effect a semiotic closure over toxicity. Despite the explicitly metapragmatic nature of model training, the important process is implicit: the model must infer features that contribute to the toxic label by honing in on statistical regularities across many examples.

In the spirit of Crawford and Paglen's (2021) excavation of a machine vision dataset, I analyze a 'toxicity correction' dataset to examine how the category of toxicity is encoded. Toxicity *correction* is the task of rewriting a comment to be non-toxic while retaining content (Atwell, et al. 2022; Laugier et al. 2021; Logacheva et al. 2022; Lu et al. 2022). The dataset, called APPDIA (Atwell et al. 2022) is *not* one of the large, influential datasets mentioned above. The data consist of about two thousand offensive Reddit comments, with corresponding style-transferred inoffensive text. The dataset was developed for a "discourse-aware" offensiveness correction model that is better able to understand the influence of context on perceived offensiveness.[13] The research was undertaken by an academic team at the University of Pittsburgh whose members work on ethical issues in NLP including AI safety, inclusion, and democratic conversational AI.

Though APPDIA is relatively marginal, and is technically an *offensiveness* corrector, it was chosen for the unique insight afforded by its parallel structure. Three expert sociolinguists were hired to rewrite Reddit comments, so that each toxic comment is annotated with a detoxified translation. The parallel data are particularly illuminating in that they make it clear what parts of the original comment warranted the label *toxic*. The annotation instructions are not reproduced exactly, but the principle was to (if possible) "remove offense," while preserving "original intent." The paper conflates intent with content, elsewhere framing the task as "eliminating offensiveness from text while preserving [the] original *semantic content*" (1). The word *content* is used 25 times (compared to twice for *intent*), mostly in collocations like *original content*, *semantic content*, and *content preservation*.

Recalling Jakobson, the intent of an utterance can extend beyond its referential (i.e., semantic) content. In each of the examples that follow, the toxic comment is displayed alongside it detoxified counterpart. I consider several of Jakobson's nonreferential speech functions (Sect. 3), showing that when offensive content is removed, priority is given to maintaining reference over other kinds of meaning. By neglecting idiom, poetics, and social meaning, and generally cutting indexical ties, the dataset encodes values reminiscent of the scientific language ideology described by Bauman and Briggs (2000). Through this analysis, we see how moderating language to focus on "substance and ideas" (a core constituent of the toxic/civil binary described in Sect. 4) can devalue other modes of expression.

The main phenomenon I attend to is the identification of *content* with referential content. But of course, translations can and do remove referential content. For the most part the transformations are glaringly obvious: profanity and hate speech are removed, insults are masked. But some changes are subtler: essentializing language is deemed toxic; insulting powerful politicians and institutions is deemed as inappropriate as insulting one's interlocutor. These observations recall and call into question the public sphere ideology that underwrites the task.

5.1 Poetic function

The **poetic function** of language highlights the formal characteristics of the message itself. Online comments often use portmanteau and wordplay, for example in insulting nicknames. In addition to minimizing the insult, detoxification often eliminates the use of portmanteau and wordplay altogether. Example (1a)'s "repugnikkans" becomes "republicans" in detoxified (1b). In Example (2), "Fucker Carlson"

[13] Data available from https://github.com/sabithsn/APPDIA-Discourse-Style-Transfer. The paper calls the task 'offensive language correction.' While at face value this might seem to pose a problem for my argument which centers on *toxicity*, the authors construction of offensiveness in terms of content echoes the constructions of toxicity laid out in Sect. 4. The analysis here, which indicates the influence of the referentialist ideology on the representation of offensiveness, illustrates the general principle that language ideology operates in all kinds of pragmatic classification tasks, from toxic language detection to human preference judgments for creating conversational language models. Since I undertook this analysis, several more parallel datasets have been released or come to my attention which do use the label toxic. Further work is needed to expand the present analysis to these datasets.

becomes "Tucker Carlson" (Tucker Carlson is a notorious right wing pundit and former Fox News anchor in the US). Example (3) contains a portmanteau of *Donald Trump* and fairy tail trickster *Rumpelstiltskin*. Example (4) uses rhyme, and (5) exploits the minimal pair between *dims* and *dems* (short for democrats).

(1)

a. **Repugnikkkans** are the worst fighters… they just run and leave you to fend for yourself
b. **Republicans** are the worst fighters… they just run and leave you to fend for yourself

(2)

a. **Fucker Carlson** is a piece of shit racist
b. **Tucker Carlson** does a lot of racist things.

(3) Suck it up, **trumplstiltskin**. You are on your own.[14]
(4) **Trump Chumps** see themselves when they gaze into the Orange Taint's idiotic eyes. Studies show that the more ignorant you are, the more confidence you have. Describes Trump and his Chumps perfectly. Pathetic.
(5) **Dims** have gotten to be the nastiest trash in our nation.

Each nickname has a kind of poetic argument in it. That argument meant to insult—Republicans are claimed to be repugnant and also racists, with "kkk" referring to the Ku Klux Klan. However, the argument is not a logical, scientific one. It's poetic in the Jakobsonian sense. That is, it relies on the formal qualities of the message itself. The nickname asserts an equivalence of the sign-objects by highlighting the iconic resemblance of the sign-vehicles. The argument goes: "Tucker Carlson *must* be a fucker, because his name has so many of the letters in it already. It's self evident."

5.1.1 Idioms

Idiom is also commonly minimized through detoxification. Example (6a) employs an idiomatic retort to an insult (takes one to know one) to imply that someone, perhaps the addressee, is also a racist.

(6)

a. Takes a racist to know one right
b. Takes racist actions to know racist actions

The speaker was probably accusing someone of racism, whether their interlocutor or a third party. The idiom *takes one to know one* is usually an insult, an *ad hominem* retort. We consider the detoxification of (6a) into (6b) on poetic and referential dimensions.

The poetic dimension of the original insult comes from the idiom. Remember that the poetic dimension relates the message to itself. It's not just what is said but the form of the message that matters. In this case, translation alters the poetic form of the comment: the *takes one to know one* construction usually demands that its 'slots' be filled by potential Experiencers. Replacing the animate noun with an inanimate noun phrase has a secondary effect of *breaking the idiom*.

On a referential level, a statement about essences has been transformed into a (somewhat nonsensical) statement about actions (*a racist* → *racist actions*). The apparent strategy of detoxification in (6b) is to minimize essentializing language. This strategy is strongly indicative of the annotator's ideology of toxicity. For the annotator, polite public discourse bars negative generalization about the *essence* of another. To go past observation of effects (racist actions) to causation (racist essence) transgresses the bounds of what people can justifiably and objectively reason about (an extremely postmodern position!).

Example (7a) uses the variable templatic construction *Just when you think X couldn't get any Yer*.

(7)

a. Just when you think liberals couldn't get any dumber…
b. Just when you think liberals could be smarter…

By reversing the polarity of the sentence, the annotator renders the insult more implicit, cloaked behind a layer of confusing positivity. Again, the idiomatic construction *breaks* through the removal of negation. Example (8) is not so much a fixed expression as it is a creative variation on the phrase *bootlicker*, an English (and I'm told Japanese) idiom for someone who sycophantically worships authority.

(8)

a. Keep licking that boot.
b. Keep doing what you're doing

The use of a fixed form again makes a kind of argument, this time by conventional wisdom. The power of an idiomatic expression—it's truth—derives from being uttered

[14] Examples (3–5) come from the Civil Comments dataset (Jigsaw 2017) In some cases, such as these, the 'detoxification' is not necessary to determine which aspects of the comment the annotator considered toxic.

countless times in countless situations. An idiom possesses a kind of intertextual force that relies on these prior utterances. *Because you can say it in this form, it is true.* Like typing in a password, the truth of an idiom in lies in muscle memory. Only for idioms, the muscle has been trained for generations. Recall that rational argumentation is a defining feature of "constructiveness" (Kolhatkar et al. 2020, see Sect. 4.2). To discount poetic and idiomatic arguments *as* arguments reifies a centuries-long tradition of excluding certain modes of being from the category of rationality (Wynter 2003).

5.2 Expressive function

Speech that relates the utterance or speech event back to the speaker employs the **expressive function** of language. Jakobson cites interjections like "Ouch!" as almost purely expressive. I'll treat nonreferential features that index and construct identity (Silverstein 1976, 2003) in terms of the expressive function. Consider how the author employs creative spelling to index their attitude and social identity in Example (9a).

(9)

a. > tRy AgAiN Typical condescending partisan
b. Stop being so condescending and such a partisan

The nonstandard orthography in "> tRy AgAiN" is conventional: the right carat is used in some online subcultures as a quotative device, and alternating capital letters are a prosodic convention that indicates a mocking tone. In other words, in a move familiar to playground rhetoricians, the speaker repeats their interlocutor's prior comment back to them in a mocking voice. In (9b) these non-referential expressive features are removed (but curiously, the accusations remain, again transformed from essential state to activity).

Most would consider it a stretch to label Example (10a) as toxic. It also uses non-standard English, this time with the term of address *bruh*.

(10)

a. You don't even have a profile bruh
b. You don't have a profile

"Messy" data are considered a necessary evil in NLP—the cost of doing data science on a large scale. Noise or not, such examples influence which patterns models discern during training. The annotator apparently interprets *bruh* as aggressive, or at least the most objectionable thing about the sentence provided, and removes it in detoxified (10b). As an interjection, *bruh* can mean many things. It is most strongly associated with expressing disappointment and frustration, but can also express happiness, sadness, approval, tiredness—it marks an emotional reaction.[15] Use of *bruh* indexes the speaker as someone who is at home in internet and youth subcultures. Stylistic features are important dimensions of online discourse. The direct quote and 'tonal' spelling in (9a) and the slang term *bruh* in (10a) perform non-referential functions of indexing speaker identity, expressing emotional and conversational stance, as well as constructing and negotiating social categories associated with such features.

Identity is also indexed by grammatical structure. Example (11a) elides the subject and the copula *be*.

(11)

a. Really bad stance. What an unbelieveable moron you are
b. That is a really bad stance

Left-edge deletion is common in informal registers, especially in written contexts like diaries, emails, and text messages. In context, the argument of the predicate would be understood to be an earlier comment, or the news article itself. In detoxification, the annotator inserts a pronominal subject and overt copula, thus 'translating' the sentence into Standard English grammar. In translation, the comment also becomes more self-contained—the predicate is given a copula and an argument in the same sentence. The ambiguity in (11) potentially stems from the same kind of left-edge deletion.

(12)

a. Just like Republicans creating themselves narratives to fuel their persecution complex to justify their hatred for everyone that isn't on the Jesus c***
b. Republicans create their own narratives to justify themselves.

Under the first reading, (12a) can be understood as a drawing a comparison between the named group (a derogatory generalization about Republicans that insults Christians) and an unnamed situation apparent from the context (i.e., "*That*'s just like Republicans creating themselves …"). Under the second reading, the elided pronoun is the same expletive *it* we use to describe the weather. Under this reading, the gerund usage is nonstandard (consider the possible

[15] https://www.urbandictionary.com/define.php?term=Bruh. Originating in African American English, as is common with many online linguistic innovations, it's been adopted by white culture and has become an enregistered feature of 'Gen Z' language (read: stereotyped, see (Agha 2003) on enregisterment).

gloss "It's just like Republicans to create …"). The annotator simplifies the comment significantly, eliminating both potential readings. The resulting (12b) stands on its own: any anaphoric reference is cut or rendered implicit. It also standardizes the grammar: the nonstandard construction *create oneself an X* is discarded.

In both (11b) and (12b), the annotators (who could read the parent comment) chose to make the comment more self-contained. This cutting of intertextual ties to other comments echoes the sentiment from constructiveness research that high-quality comments "do not excessively rely on context" (Kolhatkar et al. 2020). It also echoes the nascent ideology of scientific language exemplified by John Locke: "Creating intertextual links, for Locke, is a passive process that deters individuals from following the path to truth and knowledge; the chain of signifiers accordingly must be broken and ideas derived from texts must be 'divested of the false lights and deceitful ornaments of speech' (Locke 1971:93)" (Bauman and Briggs 2000, 153).

5.3 Conative function

The last of Jakobson's functions that I will explore here is the **conative function**, which relates the utterance to the addressee. The conative function is clear in uses of the imperative. The comment in (13a) is a very strong imperative, with an almost incantatory quality.

(13)

a. Fuck off, catch covid and give it to everyone you care about
b. You could get covid and give it to people you care about.

The detoxified (13b) has been transformed into an assertive, a cautionary bit of information about a potential risk. Gone is the profanity as well as the ill wish (though that is perhaps still implicit through thematic clash with the surrounding context), and with it the illocutionary force of the speech act and even the perlocutionary effect of the curse being placed on the addressee. The same thing happens in (14a), which is not even an imperative but itself an idiomatic, Shakespearean curse.

(14)

a. a plague upon you and upon all those who function like you !
b. that you have what you deserve, and all those who are like you

Example (14a) is not *describing* a state of the world in which the addressee is cursed, but *actively bringing it about.* The comment functions both poetically through intertextual reference, and conatively in saying something about (not just saying something about but doing something to!) the addressee. Its detoxified counterpart (14b) still has that archaic orientation, but it's expressed as a cosmic wish rather than a direct consequence of the speaker's words, and it loses the Shakespearean allusion (again cutting intertextual ties).

Given that the overt definition of offensive language refers to speech acts like insults and threats, it is unsurprising when these overt performatives become targets for detoxification, as in the removal of the rude dismissal, "Fuck off" in (13a). However, while the performativity of these insults is perhaps less salient, they remain insults. Detoxified example (13b) is still a threat, albeit a veiled one, and (14b) is still a curse.

5.4 Referentialism in action

Many aspects of the APPDIA rephrasings are easily interpreted in terms of the explicit metapragmatics of toxicity as harassment. Profanity is removed, hate speech redacted, insults blunted. Yet in case after case, features are removed or altered which would seem to have nothing to do with "toxicity" either commonly understood or defined by researchers. Under direction to preserve content, annotators elected to preserve the referential function of the comments, i.e., their denotational content, at the cost of poetic, expressive, and conative functions. Rephrased comments often minimize attention on the concrete particulars of the utterance—the speaker, the addressee, the activity or speech act they are engaged in—and foreground propositional meaning. The annotators' choices reflect the researchers instructions, imbuing the data with valued oppositions characteristic of referentialism: logical argument over poetic argument, standard English over non-standard Englishes, monolog over dialog, fact over feeling, case-building over community building, saying over doing.

5.4.1 Form vs content

In equating symbolic content with referential meaning, the annotators align indexical, non-referential meaning with style. Indeed, the detoxification task is often used as a case study for the task of 'style transfer' (Jing et al. 2020). Thus, symbolic and referential functions give the text its meaningful content, while other indexical and nonreferential functions are merely formal. A comment's potential constructive contribution lies in its content, while toxicity can come from content or form. This connection between form/content and toxic/civil reifies the asymmetrical oppositions between symbolic/indexical, rational/emotional, descriptive/

performative language characteristic of the ideology of the public sphere (Sect. 4).

5.4.2 Politeness

The transformation of essentializing language and the veiling of threats bring harassment into conformity with standards of politeness without removing it. That couching insult in dispassionate terms should make it non-toxic devalues emotion while still permitting mean-spirited interaction in coded language. Internet users commonly evade moderation of racist, sexist, homophobic sentiments by employing dog whistles and scientific language (which often come to absorb over time the negativity of the slurs they replace). The annotators' translation of open hostility into veiled antagonism actually seems to codify this strategy. As Beaver and Stanley (2023) point out, "bureaucratic language is openly antidemocratic. The practices of using bureaucratic language signal membership of an exclusive community of practice, which cannot be freely joined because of systemic societal issues and active gatekeeping" (442).

5.4.3 Saying vs doing

The transformation of performative speech acts into assertives creates a valued opposition between saying and doing that reflects the association of toxicity with forms like rationals debate. This referentialist division between is epistemically limiting. In debate, the primary speech act is assertion. Yet truth seeking through reasoning about statements about the world is just one mode of political discourse among many. Encouraging people through nudges to rephrase themselves in dispassionate, reasoned, assertive speech acts forecloses political subjectivities which rely on emotional force to convey their message. For instance, it discounts the power of pure outcry in building networks of resistance to systems of subjugation and domination.

Examples (1–6) in particular suggest a limit to what counts as valid critique in the regulated public sphere. Even the most powerful political figures are off limits for insults, and essentialization is uncivil even when used to denounce racism. APPDIA and the influential Toxic Comments dataset (Dixon et al. 2018) contain innumerable examples of insults directed at politicians, political parties, and other notable figures, as well as callouts of bigotry and hatred. What are the limits of critique? Should generalizing about socially powerful groups like politicians and CEOs (i.e., "punching up") be treated the same as generalizing about socially marginalized groups (punching across or down)? Habermas and Jigsaw may be divided on this point.

Emotional, poetic, and intertextual language is no less valuable than rational argument to public discourse. However, these features been historically devalued *because* of their association with women and people of color. A toxicity detection system which demands assimilation to referentialist norms, stereotypically associated stereotypically with white, masculine, and scientific discourses, might be labeled *covertly* racist and sexist (Hofmann et al. 2024). An association between toxicity and emotionality is not explicitly racist or sexist, but it does reflect negative societal stereotypes of Black language or women's language as overly emotional. My goal here is more than just to point out possible spurious correlations between toxicity and certain non-referential linguistic features. Rather, in combination with the analysis of overt metapragmatics in Sect. 4, this analysis demonstrates that toxicity research, and indeed NLP in general, is the work of encoding language ideology.

6 Conclusion

This article applies language ideological and metapragmatic analysis to show how the pragmatic category of toxicity is naturalized as an NLP task through professional and technical metapragmatic discourses. It illustrates one of the many ways in which technological distributions—and the socially dominant ideologies of the people who produce them—significantly influence the ideal shape of language and create *useful users*. The task of toxic language detection, designed to improve the online experience of people from marginalized arenas of society, may ultimately reify historical oppositions between referential/non-referential, symbolic/indexical and logical/emotional. These oppositions privilege linguistic forms associated with a referentialist language ideology and the social groups associated with those forms. Analyzing toxicity detection as part of a subjectivity-forming apparatus points to a potential shift in the ideal subject of online discourse away from the classic liberal model of the public sphere, from the truth-seeking to engagement-maximizing. This shift indicates a need to tend to what kinds of speech are privileged by algorithmic governmentality, and what kind of public sphere is being produced. Perhaps the utopia of maximal engagement is one where deliberation never stops—one where users are forever online interacting, reacting, saying much and doing little. Though geared toward the pragmatic end of increasing engagement, toxicity may do more to render the user politically inert than politically potent.

It is risky to critique the category of toxicity because it is designed to address real harms. And in some ways, it succeeds. In the last five years, the reputation of the YouTube comments section has undergone a dramatic transformation. It used to be a "cesspool," a place people were warned away from lest they lose their faith in humanity.[16] Now, it has a

[16] https://x.com/patrickc/status/1487984412092366849

reputation for "wholesomeness." The radical shift is a consequence of content ranking in combination with automated toxicity classification.[17] Some users are angry about the censorship of anything negative. Many give thanks that the comments section can no longer ruin their day. Others warn that genres like dangerous how-to videos necessitate some negativity: "critical comments are key if you want to keep your fingers and your head."[18] The politics of platform governance are complex, and these analytics can't determine a best course of action. But in so far as NLP research actively shapes online discourse in such a significant way, language ideology and metapragmatics are critical tools for investigating power relations and political implications of such research. The status of NLP research as language-ideological work is of particular importance given the importance of AI in the increasingly technocratic sociopolitical landscape.

Much research on toxicity is conducted outside of industry. The APPDIA dataset was developed by in the computer science department at The University of Pittsburgh. Recognizing the *ends* implicit in the *means* of toxicity detection liberates NLP researchers, especially independent academics, to shape these means in service of other ends. For starters, what pragmatic categories might serve the important political functions of playfulness, self-organization, direct action, and expressive force?

Data availability The toxicity detection datasets generated analyzed during the current study are freely available online. The Civil Comments dataseton HuggingFace under a Creative Commons 1.0 License, https://huggingface.co/datasets/google/civil_comments. The APPDIAdataset is available under a Creative Commons Attribution 4.0 International License, https://github.com/sabithsn/APPDIA-Discourse-Style-Transfer.

References

Abokhodair N, Skop Y, Rüller S, Aal K, Elmimouni H (2024) Opaque algorithms, transparent biases: automated content moderation during the Sheikh Jarrah Crisis. First Monday. https://doi.org/10.5210/fm.v29i4.13620

Agamben G (2009) "What Is an Apparatus?" and other essays. Stanford University Press

Agha A (2003) The social life of cultural value. Lang Commun Words beyond Linguis Semiotic Stud Sociocult Order 23(3):231–273. https://doi.org/10.1016/S0271-5309(03)00012-0

Agha A (2011) Large and small scale forms of personhood. Lang Commun Mediatized Commun Compl Soc 31(3):171–180. https://doi.org/10.1016/j.langcom.2011.02.006

Atwell K, Hassan S, Alikhani M (2022) APPDIA: a discourse-aware transformer-based style transfer model for offensive social media conversations. In: Proceedings of the 29th international conference on computational linguistics, 6063–74. Gyeongju, Republic of Korea: International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.530

Austin JL (1962) How to do things with words. Harvard University Press, Cambridge

Barabas C, Doyle C, Rubinovitz JB, Dinakar K (2020) Studying up: reorienting the study of algorithmic fairness around issues of power. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, 167–76. FAT* '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372859

Bauman R, Briggs CL (2000) Language philosophy as language ideology: John Locke and Johann Gottfried Herder. In: Kroskrity P (Ed) Regimes of language: discursive constructions of authority, identity, and power, pp 139–204. https://doi.org/10.1525/ae.2002.29.1.176

Bauman R, Briggs CL (1990) Poetics and performance as critical perspectives on language and social life. Ann Rev Anthropol 19:59–88

Beaver D, Stanley J (2023) The politics of language. Princeton University Press, Princeton

Bender EM, Gebru T, McMillan-Major A, Shmitchell S. 2021. On the Dangers of stochastic parrots: Can language models be too big?" In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 610–23. FAccT '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Benjamin R (2019) Race after technology: abolitionist tools for the new Jim code. Cambridge, UK; Polity

Bogdanoff A (2017) Saying Goodbye to Civil Comments. Medium. https://medium.com/@aja_15265/saying-goodbye-to-civil-comments-41859d3a2b1d

Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: 30th conference on neural information processing systems (NIPS) Barcelona, Spain. https://doi.org/10.5555/3157382.3157584

Borkan D, Dixon L, Sorensen J, Thain N, Vasserman L (2019) Nuanced metrics for measuring unintended bias with real data for text classification

Burnap P, Williams ML (2015) Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet 7(2):223–242. https://doi.org/10.1002/poi3.85

Burrell J (2024) Automated decision-making as domination. First Monday. https://doi.org/10.5210/fm.v29i4.13630

Carr A (2017) "Can alphabet's jigsaw solve google's most vexing problems?" Fast Company. https://web.archive.org/web/20180224203149/https://www.fastcompany.com/40474738/can-alphabets-jigsaw-solve-the-internets-most-dangerous-puzzles

Coffee P (2022) OpenWeb, which helps publishers target readers with ads, raises $170 Million. Wall Street Journal. https://www.wsj.com/articles/openweb-which-helps-publishers-target-readers-with-ads-raises-170-million-11666868402. Accessed 9 July 2024

Crawford K, Paglen T (2021) Excavating AI: the politics of images in machine learning training sets. AI & Soc 36(4):1105–1116. https://doi.org/10.1007/s00146-021-01162-8

Dadvar M, Trieschnigg D, De Jong F (2014) Experts and machines against bullies: a hybrid approach to detect cyberbullies. In: Sokolova M, van Beek P (eds) Advances in artificial intelligence. Springer International Publishing, Cham, pp 275–281

Dadvar M, De Jong F (2012) Cyberbullying detection: a step toward a safer internet yard. In: Proceedings of the 21st international conference on world wide web, WWW '12 Companion, New York, pp 121–26. https://doi.org/10.1145/2187980.2187995

[17] https://x.com/yablak/status/1488182247073091584

[18] https://news.ycombinator.com/item?id=30151021

Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. arXiv. https://doi.org/10.48550/arXiv.1703.04009

Díaz M, Amironesei R, Weidinger L, Gabriel I (2022) Accounting for offensive speech as a practice of resistance. In: Proceedings of the sixth workshop on online abuse and harms (WOAH). Association for Computational Linguistics, Seattle, Washington (Hybrid). pp 192–202. https://doi.org/10.18653/v1/2022.woah-1.18

Dixon L, Li J, Sorensen J, Thain N, Vasserman L (2018) Measuring and mitigating unintended bias in text classification. In: Proceedings of the 8th AAAI/ACM conference on AI, ethics, and society, Madrid, Spain. https://doi.org/10.1145/3278721.3278729

Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, Bhamidipati N (2015) Hate speech detection with comment embeddings. In: Proceedings of the 24th international conference on world wide web, 29–30. WWW '15 Companion. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2740908.2742760

Eckert P (2019) The limits of meaning: social indexicality, variation, and the cline of interiority. Language 95(4):751–776

Gehman S, Gururangan S, Sap M, Choi Y, Smith NA (2020) RealToxicityPrompts: evaluating neural toxic degeneration in language models. In: Findings of the association for computational linguistics: EMNLP 2020, 3356–69. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.301

Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, Hovy E (2021) A survey of data augmentation approaches for NLP. arXiv. http://arxiv.org/abs/2105.03075

Foucault M, Martin LH, Gutman H, Hutton PH (eds) (1988) Technologies of the self: a seminar with michel foucault. University of Massachusetts Press, Amherst

Foucault M (1978) The history of sexuality. Trans. Hurley R. 1st American ed. Pantheon Books, New York

Foucault M (1979) Discipline and punish: the birth of the prison. Trans Sheridan A. Vintage, Oxford, England

Franceschi-Bicchierai L (2019) Google's Jigsaw was supposed to save the internet. Behind the scenes, it became a toxic mess. Vice. https://www.vice.com/en/article/vb98pb/google-jigsaw-became-toxic-mess

Fraser N (1990) Rethinking the public sphere: a contribution to the critique of actually existing democracy. Social Text 25–26(25/26):56–80. https://doi.org/10.2307/466240

Gonen H, Goldberg Y (2019) Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv. https://doi.org/10.48550/arXiv.1903.03862

Green Y (2022) Tech companies are reconsidering an old enemy. Wired, September. https://www.wired.com/story/friction-social-media-moderation/

Habermas J (1990) Moral consciousness and communicative action. MIT Press, Cambridge

Habermas J (1989) The structural transformation of the public sphere: an inquiry into a category of bourgeois society. Translated by Thomas Burger with the Assistance of Frederick Lawrence. Studies in Contemporary German Social Thought. MIT Press, Cambridge, Mass

Hofmann V, Kalluri PR, Jurafsky D, King S (2024) Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv. https://arxiv.org/abs/2403.00742v1

Irvine JT, Gal S (2000) Language ideology and linguistic differentiation. In: Kroskrity P (ed) Regimes of language: ideologies, polities, and identities. School of American Research Press, Santa Fe, pp 35–84

Izento (2021) The toxic psychology of riot lyte. Esportsheaven. https://www.esportsheaven.com/features/the-toxic-psychology-of-riot-lyte/

Jakobson R (1960) Closing statements: linguistics and poetics. In: Sebeok TA (ed) Style in language. MIT Press, Cambridge, pp 350–377

Jereza R (2024) 'I'm Not This Person': racism, content moderators, and protecting and denying voice online. New Media Soc 26(8):4454–4470. https://doi.org/10.1177/14614448221122224

Jigsaw (2021) "Toxicity." The Current: The Toxicity Issue. https://jigsaw.google.com/the-current/toxicity/

Jigsaw (2017) Jigsaw toxic comment classification challenge. Kaggle. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

Jigsaw (2019) Jigsaw unintended bias in toxicity classification. Kaggle. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification

Jigsaw (2020) Jigsaw multilingual toxic comment classification. Kaggle. https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification

Jigsaw (2024) Announcing experimental bridging attributes in perspective API. Medium. Jigsaw. https://medium.com/jigsaw/announcing-experimental-bridging-attributes-in-perspective-api-578a9d59ac37

Jing Y, Yang Y, Feng Z, Ye J, Yu Y, Song M (2020) Neural style transfer: a review. IEEE Trans Visual Comput Graph 26(11):3365–3385. https://doi.org/10.1109/TVCG.2019.2921336

Kara Corvus (2020) "Discussing My Abus3 from Riot Lyte/Jeffrey Lin." https://www.youtube.com/watch?v=Z3FKgCBJflM. Accessed 14 Mar 2024

Kockelman P (2020) The epistemic and performative dynamics of machine learning praxis. Signs Soc 8(2):319–355. https://doi.org/10.1086/708249

Kolhatkar V, Taboada M (2017) Constructive language in news comments. In: Waseem Z, Chung WHK, Hovy D, Tetreault J (Eds) Proceedings of the first workshop on abusive language online, 11–17. Vancouver, BC, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3002

Kolhatkar V, Thain N, Sorensen J, Dixon L, Taboada M (2020) Classifying constructive comments. arXiv. http://arxiv.org/abs/2004.05476

Kwok I, Wang Y (2013) Locate the hate: detecting tweets against blacks. Proc AAAI Conf Artif Intell 27(1):1621–1622. https://doi.org/10.1609/aaai.v27i1.8539

Latour B (1987) Science in action: how to follow scientists and engineers through society. Open University Press, Philadelphia

Laugier L, Pavlopoulos J, Sorensen J, Dixon L (2021) Civil rephrases of toxic texts with self-supervised transformers. In: Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main Volume, 1442–61. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.124

Laugier L (2022) Analysis and control of online interactions through neural natural language processing. PhD thesis, Institut Polytechnique de Paris. https://theses.hal.science/tel-03884481

Lee B, Kim H, Kim K, Choi YS. 2024. "XDetox: text detoxification with token-level toxicity explanations. In: Proceedings of the 2024 conference on empirical methods in natural language processing, 15215–26. Miami, Florida, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.848

Lin J (2013) the science behind shaping player behavior in online games. Game Developers Conference. San Francisco. https://www.gdcvault.com/play/1017940/The-Science-Behind-Shaping-Player

Logacheva V, Dementieva D, Ustyantsev S, Moskovskiy D, Dale D, Krotova I, Semenov N, Panchenko A (2022) ParaDetox: detoxification with parallel data." In Proceedings of the 60th annual meeting of the association for computational linguistics. Association

for Computational Linguistics, Dublin, Ireland, pp 6804–6818. https://doi.org/10.18653/v1/2022.acl-long.469
Lu X, Welleck S, Hessel J, Jiang L, Qin L, West P, Ammanabrolu P, Choi Y (2022) Quark: controllable text generation with reinforced Unlearning. arXiv. https://doi.org/10.48550/arXiv.2205.13636
Lucy JA (1992) Language diversity and thought: a reformulation of the linguistic relativity hypothesis. Cambridge University Press, New York
Mahmud A, Ahmed KZ, Khan M (2008) Detecting flames and insults in text, Brac University, Dhaka, Bangladesh. http://dspace.bracu.ac.bd:8080/xmlui/handle/10361/714
Malmasi S, Zampieri M (2018) Challenges in discriminating profanity from hate speech. J Exp Theor Artif Intell 30(2):187–202. https://doi.org/10.1080/0952813X.2017.1409284
Miceli M, Schuessler M, Yang T (2020) Between subjectivity and imposition: power dynamics in data annotation for computer vision. Proc ACM Hum-Comput Interact. 4(2):1–25. https://doi.org/10.1145/3415186
OpenWeb (2020) Can machines change human behavior? OpenWeb. https://www.openweb.com/blog/can-machines-change-human-behavior-openweb-using-jigsaws-perspective-api-releases-case-study-measuring-the-effects-of-real-time-feedback-and-nudges-in-decreasing-toxicity
Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C et al. (2022) Training language models to follow instructions with human feedback. arXiv. https://doi.org/10.48550/arXiv.2203.02155
Pavlopoulos J, Sorensen J, Dixon L, Thain N, Androutsopoulos I (2019) ConvAI at SemEval-2019 Task 6: offensive language identification and categorization with perspective and BERT. In: Proceedings of the 13th international workshop on semantic evaluation. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp 571–576. https://doi.org/10.18653/v1/S19-2102
Pavlopoulos J, Sorensen J, Dixon L, Thain N, Androutsopoulos I (2020) Toxicity detection: does context really matter? In: Proceedings of the 58th annual meeting of the association for computational linguistics, 4296–4305. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.396
Pavlopoulos J, Laugier L, Xenos A, Sorensen J, Androutsopoulos I (2022) From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In: Proceedings of the 60th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Dublin, Ireland, pp 3721–3734. https://doi.org/10.18653/v1/2022.acl-long.259
Price I, Gifford-Moore J, Fleming J, Musker S, Roichman M, Sylvain G, Thain N, Dixon L, Sorensen J (2020) Six attributes of unhealthy conversation. arXiv. https://doi.org/10.48550/arXiv.2010.07410
Rampton B, Richardson K (2007) Deborah Cameron, Elizabeth Frazer, Penelope Harvey, power/knowledge: the politics of social science. In Discourse Reader, 2nd ed. Routledge, London
Rieder B, Skop Y (2021) The fabrics of machine moderation: studying the technical, normative, and organizational structure of perspective API. Big Data Soc 8(2):205395172110461. https://doi.org/10.1177/20539517211046181
Rose N, Miller P (1992) Political power beyond the state: problematics of government. Br J Sociol 43(2):173–205
Rouvroy A, Stiegler B (2016) The digital regime of truth: from the algorithmic governmentality to a new rule of law, La Deleuziana. 3:6-29
Sap M, Swayamdipta S, Vianna L, Zhou X, Choi Y, Smith NA (2022) Annotators with attitudes: how annotator beliefs and identities bias toxic language detection. arXiv. https://doi.org/10.48550/arXiv.2111.07997
Silverstein M (2003) Indexical order and the dialectics of sociolinguistic life. Lang Commun 23(3):193–229. https://doi.org/10.1016/S0271-5309(03)00013-2
Sax S (2016) Flame wars: automatic insult detection. Department of Computer Science, Stanford University
Sharma HK, Kshitiz K (2018) "NLP and machine learning techniques for detecting insulting comments on social networking platforms. In: 2018 International conference on advances in computing and communication engineering (ICACCE), pp 265–272. https://doi.org/10.1109/ICACCE.2018.8441728
Silverstein M (1976) Shifters, linguistic categories, and cultural description. In: KH Basso, HA Selby (Eds): Meaning in Anthropology. University of New Mexico, Albuquerque, pp 11-55. https://cscs.uchicago.edu/mslv-library/
Simon G (2020) OpenWeb Tests the Impact of 'Nudges' in Online Discussions. OpenWeb. https://www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api
Sood SO, Churchill EF, Antin J (2012b) Automatic identification of personal insults on social news sites. J Am Soc Inform Sci Technol 63(2):270–285. https://doi.org/10.1002/asi.21690
Sood SO, Antin J, Churchill EF (2012) Using crowdsourcing to improve profanity detection: 2012 AAAI spring symposium. In: Wisdom of the Crowd - Papers from the AAAI Spring Symposium, AAAI Spring Symposium, 69–74. http://www.scopus.com/inward/record.url?scp=84865029152&partnerID=8YFLogxK
Srivastava A, Rastogi A, Rao A, Shoeb AA, Abid A, Fisch A, Brown AR et al. (2023) Beyond the imitation game: quantifying and extrapolating the capabilities of language models. arXiv http://arxiv.org/abs/2206.04615
Thaler RH, Sunstein CR (2008) Nudge improving decisions about health, wealth, and happiness. Yale University Press, New Haven
Urla J (2019) Governmentality and Language. Ann Rev Anthropol 48:261–278
Vasserman L (2019) Combating AI Bias at Scale Jigsaw. Presented at Data Council Conference, New York. https://www.youtube.com/watch?v=t_Pm2WrOp5E
Vasserman L (2023) Reducing toxicity in large language models with perspective API. Medium - Jigsaw. https://medium.com/jigsaw/reducing-toxicity-in-large-language-models-with-perspective-api-c31c39b7a4d7
Welbl J, Glaese A, Uesato J, Dathathri S, Mellor J, Hendricks LA, Anderson K, Kohli P, Coppin B, Huang PS (2021) Challenges in detoxifying language models. arXiv. http://arxiv.org/abs/2109.07445
Woolard KA (2020) Language ideology. The international encyclopedia of linguistic anthropology. Wiley, pp 1–21. https://doi.org/10.1002/9781118786093.iela0217
Wulczyn E, Thain N, Dixon L (2017) Ex machina: personal attacks seen at scale. arXiv. https://doi.org/10.48550/arXiv.1610.08914
Wynter S (2003) Unsettling the coloniality of being/power/truth/freedom: towards the human after man, its overrepresentation–an argument. CR New Centennial Rev. 3(3):257–337. https://doi.org/10.1353/ncr.2004.0015
Xu JM, Jun KS, Zhu X, Bellmore A. 2012. Learning from Bullying Traces in Social Media. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics, Human Language Technologies
Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW. 2018. Gender bias in coreference resolution: evaluation and debiasing methods. In: M Walker, H Ji, and A Stent (Eds) Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 15–20. https://doi.org/10.18653/v1/N18-2003

Saleem HM, Kurrek J, and Ruths D (2022) Enriching Abusive Language Detection with Community Context. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). Seattle, Washington, pp 131–142

Chronis, G (2025) Ways of speaking of speaking machines. Manuscript in preparation. Department of Linguistics, The University of Texas at Austin

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.