2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

A Multidisciplinary Lens of Bias in Hate Speech

Paula Reyero Lobo
Knowledge Media Institute
The Open University
Milton Keynes, UK
paula.reyero-lobo@open.ac.uk

Joseph Kwarteng†
Knowledge Media Institute
The Open University
Milton Keynes, UK
joseph.kwarteng@open.ac.uk

Mayra Russo†
L3S Research Center
Leibniz Universität Hannover
Hanover, Germany
mrusso@l3s.de

Miriam Fahimi†
Digital Age Research Center
University of Klagenfurt
Klagenfurt, Austria
miriam.fahimi@aau.at

Kristen Scott†
Leuven.AI
KU Leuven
Leuven, Belgium
kristen.scott@kuleuven.be

Antonio Ferrara†, Indira Sen†
GESIS Leibniz Institute for Social Science
RWTH Aachen University
Cologne, Germany
{Antonio.Ferrara, Indira.sen}@rwth-aachen.de

Miriam Fernandez
Knowledge Media Institute
The Open University
Milton Keynes, UK
miriam.fernandez@open.ac.uk

***Abstract*—Hate speech detection systems may exhibit discriminatory behaviours. Research in this field has focused primarily on issues of discrimination toward the language use of minoritised communities and non-White aligned English. The interrelated issues of bias, model robustness, and disproportionate harms are weakly addressed by recent evaluation approaches, which capture them only implicitly. In this paper, we recruit a multidisciplinary group of experts to bring closer this divide between fairness and trustworthy model evaluation. Specifically, we encourage the experts to discuss not only the technical, but the social, ethical, and legal aspects of this timely issue. The discussion sheds light on critical bias *facets* that require careful considerations when deploying hate speech detection systems in society. Crucially, they bring clarity to different approaches for assessing, becoming aware of bias from a broader perspective, and offer valuable recommendations for future research in this field.**

***Index Terms*—hate speech, bias, multidisciplinary methods**

I. Introduction

With the rapid spread of harmful online content [1], social media platforms have turned to the implementation of automated online content moderation systems. However, Machine Learning (ML) systems exhibit biases and lack robustness, which disproportionately affects minoritized and historically disadvantaged populations [2], [3]. Increasingly, larger models are being used for this task, as seen in the case of prompt-based large language models for hate speech (HS) detection [4]. In doing so, it can become even more challenging to understand and address the risks associated with misclassification, than when employing smaller pre-trained language models fine-tuned for the task. Some reasons for this are the lack of traceability, and the societal biases existing in the growing pre-training corpora needed to feed these models.

While biases in HS detection have to some extent been analysed [5] and reviewed [6], they are often addressed in isolation. Previous work mainly focuses on concerns about the inaccurate detection in the presence of specialized language used by the most frequently targeted communities and non-White aligned English. These interrelated issues, of bias, robustness, and disproportionate harms are only weakly addressed in current model evaluation approaches, particularly in content moderation [7]. That is, bias is only implicitly assessed in a subset of the data perturbations used to stress the reliability of these models [8]. Towards closing this divide, we bring together a cohort of experts on bias in Artificial Intelligence (AI) from multiple disciplines to work on a data challenge specifically designed to optimise hate speech detection while minimizing bias. With guidance from practitioners in the specific downstream application, we re-imagine concerns surrounding fair and trustworthy models in the context of content moderation. We think critically about technical, social, ethical, and legal considerations to provide a more holistic view of existing and underexplored issues. As a result, we identify and highlight key bias *facets* along multidisciplinary axes, particularly focusing on approaches to measure or contextualise these risks more effectively. Our findings and recommendations complement recent work [9], [10] that zooms out on the academic ML research into hate speech detection and critically lays out gaps in the real-world adoption and use of such systems.

Our main contribution is: An interdisciplinary discussion of notions that must be considered when deploying HS detection in society. We seek to promote a nuanced understanding and encourage developing solutions that address not only technical requirements, but also the interrelated ethical, societal and legal aspects of the problems surrounding these systems.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias". Views are those of the authors only.

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11.
https://doi.org/10.1145/3625007.3627491

II. Bias and Hate Speech: "Facets"

In this section, we present the identified bias *facets*. In each *facet*, we discuss implications, real-world examples, and our proposed recommendations. The discussion initiates in the context of a data challenge, the specifications of which are outside the scope of this work. Detailed descriptions of the challenge can be found in our repository: https://zenodo.org/badge/latestdoi/665530454.

A. Technical

1) Intersectionality: Crenshaw introduced the concept of intersectionality to describe how different social identities (e.g., race, ethnicity, gender, sexual orientation, religion) overlap, creating unique systems of discrimination [11]. Thus, hate speech can exhibit substantial variation depending on the intersections of these identities. Understanding and addressing intersections and diverse forms of identities is essential for developing effective detection systems. Modelling intersectional characteristics requires a nuanced understanding of the specific experiences and vulnerabilities faced by different groups [3] (e.g., hate that targets both racial and gender identities [12]). Critically, it requires careful consideration of the unique experiences and discriminations faced by each group [13]. We propose that, to reflect intersectionality, metrics in this context should not only focus on overall accuracy or performance but also account for the nuanced experiences of different marginalised groups. The making of such detection tools should be more *intersectionally*-sensitive. Thus, diverse and representative training data, feature engineering considering intersectional characteristics to capture the specific language and context associated with hate speech targeting multiple marginalized groups, and the evaluation process should include assessments of bias and fairness to ensure that the model does not disproportionately harm any marginalized group.

2) Fairness evaluation: Many HS detection models, or their variants, are deployed in real-life content moderation settings, entailing that these models are used for high-risk decision-making (e.g. Perspective API used by New York Times). Indeed model errors can lead to wrongful stigmatization and de-platforming of minoritized individuals [2]. While a large body of literature has looked into the biases of algorithmic decision-making systems [14], including examinations of explicit, implicit, and unintended bias of HS detection models, Blodgett et al., discuss that many of these investigations are abstract and decontextualized from real-life harms [15]. Bias investigations and debiasing usually address *unintended identity bias*, where non-harmful content containing identity terms is often misclassified as hateful. Other complementary directions focus on testing the robustness of HS detection models through adversarial benchmarks [8] or test suites [7]. While the underlying dataset used to train models and the concentration of identity terms in the hateful class could be driving this bias, the root cause is still unclear. While some debiasing techniques have shown promise [16], [17], their generalizability to unknown identity terms is an open question. Besides, techniques like data augmentation that can improve the robustness of models [18], [19] can *also* introduce bias [20]. Future research avenues should include more nuanced and interpretable fairness metrics that are grounded in real-world harms and deeper investigations into the source of bias.

3) Multimodality: Despite the strong focus on textual data, hate speech is expressed in different data modalities or in their combination. The multimodal nature of the problem poses additional challenges in the parameterization of relevant features and creation of data sources [21]. Notably, when anticipating additional evaluation challenges under this setting, seeking human-centered evaluation is essential. Systems should not be based on the aggregation of metrics from the different modalities, but rather respond to the preferences of a diverse sample of the users interacting with these systems [22]. An open-ended feedback mechanism can help guide complex systems to respond to a more diverse set of human preferences.

4) Data Sources: Algorithmic systems, such as HS detection models, are reliant on datasets that are mainly elaborated from social data; user-generated data obtained from different online platforms, e.g., social media and networking platforms, search platforms, and collaborative sites, among many others [23]. This practice has become widespread due to the elevated data dependency of these systems, as this provides a faster and less costly way to access larger volumes of data [24]. While a practical approach, the resulting datasets can have a series of limitations, e.g., they lack representativeness, are unreliable, inconsistent, or irrelevant, and have bias and quality issues. Moreover, to be used in supervised or semi-supervised ML settings, these datasets need to be annotated. More often than not, data annotation tasks are performed by human annotators under different sets of constraints, e.g., unclear annotation instructions, inadequate domain expertise and targets of such hates, time constrains. These factors can ultimately be reflected in the characteristics of the obtained annotations.

In terms of bias, [23], [25], demonstrates how the whole process of data collection is susceptible to different types of biases creeping in. This implies that great efforts in mitigation at the pre-processing and post-processing stages are required, translating into resource-intensive tasks that still do not guarantee bias-free algorithmic systems. In order to avoid reactionary interventions, dataset production and curation is due for reassessment. Primarily, it is imperative that these processes go beyond easily accessible data and extractive practices. Ideally, careful consideration of the downstream task, the identification of potential biases that could arise at an early stage, and iterative value-sensitive documentation of datasets [26], could already undercut quality issues later down the pipeline. Further, value-based data collection practices, such as data donation initiatives or co-design sessions with experts, e.g., linguists [27], and affected communities [28], could also support the production of higher-quality datasets. Similarly, engaging in ethical data annotation practices that contemplate task formulation that account for downstream tasks as well as annotator expertise and that proactively engage them in annotation tasks without undermining their expertise

or obfuscating their world-view are key considerations [29].

B. Social

1) Annotation disagreement: The process of annotating data often involves subjective judgments, as annotators may interpret and label information differently. Critically, annotators' social identities, such as their cultural background, education, or personal experiences, can influence their perceptions and understanding of the data [13], [30]. It is crucial to consider the subjectivity and impact of annotator demographics when analysing disagreements, as it can lead to variations in labelling decisions. For instance, cultural differences or language proficiency may result in different interpretations of certain concepts. Additionally, individual biases and perspectives can affect how annotators perceive and classify information. Acknowledging these factors is needed to better understand the sources of annotation disagreement, critically in the classification of hate speech content. While it is common to focus on aggregated annotations to derive a consensus, solely relying on the final agreement overlooks the potential value of disagreements [31]. Disagreements can provide valuable insights into the complexity of the annotation task and the underlying data. They highlight the diverse perspectives and interpretations of annotators, which may uncover hidden patterns or nuances in the data that would have otherwise been missed. Such examples and similar approaches [32] go as far as recommending the retention of annotator-level documentation and making this information available when releasing new datasets.

2) Data labour: AI-systems, e.g., HS detection, in the current landscape are highly data-dependent; this fundamental characteristic creates a huge demand for data and, consequently, datasets. Furthermore, deployed AI systems still often fail to function as promised [28], which makes it necessary for the on-demand involvement of humans to step in order to attenuate their deficiencies [33]. In order to produce these use/research-ready datasets and perform intermediary or post deployment interventions, an intensive amount of, and globally distributed, human labour is required [34]. In this work, we use the working definition of data labour given in [24], as "activities that produce digital records useful for capital generation". While the activities alluded to are vast, for our purposes, we highlight the following: data creation, collection, aggregation, labelling, content disambiguation, and content moderation. In doing so, we aim to emphasize how, for the most part, performing these labours implies concerns related to low or inconsistent wages, lack of job security, exploitation, and mental distress, and at the same time, remain under discussed. Data labourer's well-being should be at the forefront when tackling ethics in AI development in general, and HS detection systems in particular. Involvement in this task exposes data labourers to hateful, obscene, offensive and sensitive material, that has been known to cause psychological distress or more permanent mental health issues. While there is still a long way to go to assure better working conditions and due recognition for this labour, it is imperative that researchers and practitioners working on ethical issues pertaining to AI are also accountable and voice their concerns on these matters, especially because as actors, we often find ourselves hiring or performing these tasks (to varying degrees of precariousness). Groundbreaking research in this area [24], [34], [35] advocates for the implementation of frameworks that empower labourers, with recommendations such as: transparent communication channels between contractors and workers; mechanisms to proactively incorporate input and feedback into production workflows; sustained mental counsel; better remuneration; solidarity with labour organization movements, to name a few.

3) Offline world impacts: Both scientific research and journalistic investigations have shown the real-world impact of online hate speech [36], specifically "dangerous speech" which has led to violence spilling over from social media platforms [37]. Research has also shown mobilization and radicalization effects on platforms with highly hateful content [38]. On an individual level, hateful content can affect people's psychological well-being [39] and lead to spirals of silence [40]. Taken together, there is a strong need to moderate such types of content. On the flip side, excessive moderation can lead to the curtailment of people's right to speech and veer into censorship, implying that HS detection and its use for content moderation is neither technically nor socially trivial. It is a rather complex socio-technical issue, akin to a 'wicked problem' that requires careful design choices and thinking about trade-offs.

C. Ethical

1) Tackling hate speech from the source: What is perceived as hate speech in 'the real-world' depends on the concrete situation, the position of the person speaking and the interpretation of the one who is affected by it. Conceptually, hate speech is a *co-constructive process* between a person who speaks hate and the person who receives it, entangled with their positionalities and societal embedding. At the same time, the roots of hate speech go beyond the co-construction between two persons, as hate has structural origins. For instance, hate speech against a Jewish person is linked to systemic anti-semitism in society. The data challenge opened up the possibility to discuss such structural dimensions of hate speech. Insights from disciplines such as gender studies can be of value in order to tackle hate speech from the source. For example, in "Excitable Speech", Judith Butler (1997) [41] argues how hate speech is also always *state speech*. She argues that the state enables structural forms of discrimination, that lead to individuals learning hate. We think that much research needs to be done to dismantle how structural forms of discrimination enter collective and individual speech.

2) Deciding on hate speech: Differences in what is considered hateful or not can vary across cultures, geographic boundaries, and social groups. Additionally, contextual factors such as topic, time of day, community mores and perceptions and who might be 'listening' play a role. Any individual decisions, but also high level rules about what is permitted are going to be disagreed with by some people or groups. Rather

than attempting to frame any hate speech policy as neutral or universal, or to make it acceptable to everyone, there is a need to be explicit about the reasons and philosophy behind the policy. It is not bad, or untenable, for different groups and communities to have disagreements about acceptable speech within their own online community. What is a problem is if the decisions on policies for all people's experience are being made by some small, powerful group, in a top-down matter; this is particularly a risk when online communities are managed by a few massive companies. In this case, users who feel they are sanctioned wrongly, unprotected, or simply want different experiences may have limited powers of influence and paths for recourse. An automated application of a single set of rules is particularly troubling when considering a global context. Extremely varied cultural contexts increase the amount of disagreement about what constitutes hate speech. A uniform approach is unlikely to properly address localized nuances around speech and discrimination, and is also unlikely to work equally at detecting hate against all groups or treating all groups equally.

The Santa Clara Principles [42] are a set of high-level guiding principles which are meant to encompass the agreement of how online hate speech policies should be applied. These principles are in their second iteration and were created by "a broad coalition of organizations, advocates, and academic experts". They take a human rights based approach to the question of how to ensure responsible platform moderation. The five foundational principles are summarized as follows:

- *Human rights and due process* should be considered in all aspects of moderation, and the mechanism by which it is should be made clear to the users.
- Companies should clearly communicate *understandable rules and policies* about what is not permitted and what actions will be taken.
- Moderating must be done only by people with *cultural competence* in the relevant language and socio-political context, and users must have access to all rules and recourse mechanisms in their own language.
- Any *state involvement in content moderation* should be made explicit to users.
- *Integrity and explainability* of moderation systems should be ensured through monitoring and results publicly shared.

In terms of the specific approaches, a given platform should take to moderation in line with these principles; while there is room for developing much more grounded ways of conceptualizing and operationalizing hate speech. This could be done either based on (social science) theory or participatory methods by taking input from people most affected by it (i.e. targets of hate).

D. Legal

1) Proprietary and Non-transparency: The content displayed by private social media platforms is often rendered opaque by the use of algorithmic systems that are created, trained and developed behind "closed doors" of proprietary rules. The same holds true for the non-transparent mechanisms in place to flag or moderate content and hate speech. Usually, social media platforms provide terms of service (TOS), in which they display some of their guidelines. However, the models that are in place to comply with these guidelines, and the specificities of the TOS as such, are decided within the corporation. We discuss how this could introduce a *corporate bias*, where corporate actors (secretly and mundanely) decide which content is hateful and which is not. What is more, users of platforms do not have the possibility of intervening in their decisions. We discussed how we need more spaces for community-driven interventions and guidelines on which forms of content should be moderated, or flagged as hate speech. For instance, the recent protest against censoring bare female breasts (and no male breasts) by Instagram evoked the platform to change some of its flagging practices [43]. Such successful intervention by users may also serve as an inspiration for HS detection.

2) Regulatory gaps: HS detection is made more difficult due to regulatory gaps in relation to discrimination, as what is understood and labelled as 'hate speech' is usually influenced by the prevailing legal understandings of discrimination. Particularly, we identify three of such regulatory gaps. First, legal regulations do not protect against discrimination based on socio-economic status and social class. As a result, *class bias* often remains invisible and receives little attention even in computer science. Models for HS detection may overlook class bias or fail to perceive them as hate speech, it may be unclear which terminologies refer to which social class. We discussed that hate speech that is directed towards precarious groups should definitely be categorized as such – not least because of the intersections with other markers of difference [44]. We identified a second regulatory gap as *speciesist bias*, i.e. directed towards animals or the environment. In times of climate crisis and the increasing conservative backlash and conspiracy discourse, we find it important to act against, and detect such hate speech. Finally, we discussed a *decolonial bias* that may narrow down HS detection to a Eurocentric view. In contrast, a decolonial perspective that centres other forms of knowledge and languages may relate every form of European language with oppression [45]. Further, how hate is expressed in indigenous languages and how it is then translated into online spaces remains an open question. After all, language models may still work best for European languages, and hate is expressed differently across different languages, cultures, and regions.

III. Conclusion

This paper presents a multidisciplinary analysis of bias in hate spech detection, drawing on insights from fairness, online content moderation, gender studies, ethics, and law. Through a collaborative activity, we emphasize key challenges that involve system developers, platforms, regulatory bodies, data workers, and society as a whole. The multi-stakeholder, multidisciplinary perspective provides a more holistic view of biases, the actors related to bias, and their critical role in the design of automatic detection systems.

References

[1] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of hate speech in online social media," in *Proceedings of the 10th ACM Conference on Web Science*, ser. WebSci '19, 2019, p. 173–182.

[2] O. L. Haimson, D. Delmonaco, P. Nie, and A. Wegner, "Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas," *Proc. ACM Hum.-Comput. Interact.*, oct 2021.

[3] J. Kwarteng, S. C. Perfumi, T. Farrell, A. Third, and M. Fernandez, "Misogynoir: challenges in detecting intersectional hate," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–15, 12 2022.

[4] J. Ji, W. Ren, and U. Naseem, "Identifying creative harmful memes via prompt based approach," in *Proceedings of the ACM Web Conference 2023*, ser. WWW '23, 2023, p. 3868–3872.

[5] M. Wiegand, E. Eder, and J. Ruppenhofer, "Identifying implicitly abusive remarks about identity groups using a linguistically informed approach," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jul. 2022.

[6] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling bias in toxic speech detection: A survey," *ACM Comput. Surv.*, jan 2023.

[7] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "HateCheck: Functional tests for hate speech detection models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Aug. 2021, pp. 41–58.

[8] A. Calabrese, M. Bevilacqua, B. Ross, R. Tripodi, and R. Navigli, "AAA: fair evaluation for abuse detection systems wanted," in *WebSci '21: 13th ACM Web Science Conference*. ACM, 2021, pp. 243–252.

[9] W. Yin and A. Zubiaga, "Towards generalisable hate speech detection: a review on obstacles and solutions," *PeerJ Computer Science*, vol. 7, p. e598, 2021.

[10] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," *Language Resources and Evaluation*, vol. 55, pp. 477–523, 2021.

[11] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]," in *Feminist Legal Theory: Readings in Law and Gender*, 2018, vol. 1989, no. 1, pp. 57–80.

[12] M. Bailey and Trudy, "On misogynoir: Citation, erasure, and plagiarism," *Feminist Media Studies*, vol. 18, no. 4, pp. 762–768, 2018.

[13] J. Kwarteng, G. Burel, A. Third, T. Farrell, and M. Fernandez, "Understanding misogynoir: A study of annotators' perspectives," in *Proceedings of the 15th ACM Web Science Conference 2023*, 2023, pp. 271–282.

[14] S. Dev, E. Sheng, J. Zhao, J. Sun, Y. Hou, M. Sanseverino, J. Kim, N. Peng, and K.-W. Chang, "What do bias measures measure?" *arXiv e-prints*, pp. arXiv–2108, 2021.

[15] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of" bias" in nlp," *arXiv preprint arXiv:2005.14050*, 2020.

[16] D. Nozza, C. Volpetti, and E. Fersini, "Unintended bias in misogyny detection," in *Ieee/wic/acm international conference on web intelligence*, 2019, pp. 149–155.

[17] B. Kennedy, X. Jin, A. Mostafazadeh Davani, M. Dehghani, and X. Ren, "Contextualizing hate speech classifiers with post-hoc explanation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 5435–5442.

[18] M. Samory, I. Sen, J. Kohne, F. Flöck, and C. Wagner, ""call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples," in *Proceedings of the international AAAI conference on web and social media*, vol. 15, 2021, pp. 573–584.

[19] I. Sen, M. Samory, F. Flöck, C. Wagner, and I. Augenstein, "How does counterfactually augmented data impact models for social computing constructs?" in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 325–344.

[20] I. Sen, M. Samory, C. Wagner, and I. Augenstein, "Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 4716–4726.

[21] A. Chhabra and D. K. Vishwakarma, "A literature survey on multimodal and multilingual automatic hate speech identification," *Multimedia Syst.*, p. 1203–1230, jan 2023.

[22] C. Ziems, J. Chen, C. Harris, J. Anderson, and D. Yang, "VALUE: Understanding dialect disparity in NLU," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, May 2022, pp. 3701–3720.

[23] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Frontiers in Big Data*.

[24] H. Li, N. Vincent, S. Chancellor, and B. Hecht, "The dimensions of data labor: A road map for researchers, activists, and policymakers to empower data producers," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, p. 1151–1161.

[25] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021.

[26] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 587–604, 2018.

[27] C. Harris, M. Halevy, A. Howard, A. Bruckman, and D. Yang, "Exploring the role of grammar and word choice in bias toward african american english (aae) in hate speech classification," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[28] I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst, "The fallacy of ai functionality," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[29] E. Denton, I. Kivlichan, M. Díaz, R. M. Rosen, and V. Prabhakaran, "Whose ground truth? accounting for individual and collective identities underlying dataset annotation," 2021.

[30] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, and M. Bailey, "Designing toxic content classification for a diversity of perspectives," in *Proceedings of the 17th Symposium on Usable Privacy and Security, SOUPS 2021*, 2021.

[31] V. Prabhakaran, A. Mostafazadeh Davani, and M. Diaz, "On releasing annotator-level labels and information in datasets," in *Proceedings of the Joint 15th LAW and 3rd DMR Workshop*, Nov. 2021.

[32] V. Basile, F. Cabitza, A. Campagner, and M. Fell, "Toward a Perspectivist Turn in Ground Truthing for Predictive Computing," *arXiv e-prints*, p. arXiv:2109.04270, Sep. 2021.

[33] P. Tubaro, A. A. Casilli, and M. Coville, "The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence," *Big Data & Society*, vol. 7, no. 1, p. 2053951720919776, 2020.

[34] M. Miceli, J. Posada, and T. Yang, "Studying up machine learning data: Why talk about bias when we mean power?" *Proc. ACM Hum.-Comput. Interact.*, jan 2022.

[35] M. Miceli, T. Yang, A. A. Garcia, J. Posada, S. M. Wang, M. Pohl, and A. Hanna, "Documenting data production processes: A participatory approach for data work," *Proc. ACM Hum. Comput. Interact.*, vol. 6, no. CSCW2, pp. 1–34, 2022.

[36] K. Müller and C. Schwarz, "Fanning the flames of hate: Social media and hate crime," *Journal of the European Economic Association*, vol. 19, no. 4, pp. 2131–2167, 2021.

[37] S. Benesch, "Dangerous speech: A proposal to prevent group violence," *Voices That Poison: Dangerous Speech Project*, 2012.

[38] M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, and W. Meira Jr, "Auditing radicalization pathways on youtube," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020.

[39] K. Saha, E. Chandrasekharan, and M. De Choudhury, "Prevalence and psychological effects of hateful speech in online college communities," in *Proceedings of the 10th ACM conference on web science*, 2019.

[40] C. S. C. Olson and V. LaPoe, ""feminazis,""libtards,""snowflakes," and "racists": Trolling and the spiral of silence effect in women, lgbtqia communities, and disability populations before and after the 2016 election," *The journal of public interest communications*, 2017.

[41] J. Butler, *Excitable speech: a politics of the performative*, routledge classics edition ed., ser. Routledge classics. London New York: Routledge Classics, 2021.

[42] Santa Clara Principles on Transparency and Accountability in Content Moderation. Santa Clara Principles.

[43] A. Demopoulos, "Free the nipple: Facebook and Instagram told to overhaul ban on bare breasts," *The Guardian*, Jan. 2023.

[44] G. Winker and N. Degele, "Intersectionality as multi-level analysis: Dealing with social inequality," *European Journal of Women's Studies*, Feb. 2011.

[45] W. Mignolo, *The politics of decolonial investigations*, ser. On Decoloniality. Durham: Duke University Press, 2021.